FIDELITY CENTER for APPLIED TECHNOLOGY®

Call for Research Proposals

Theme Two: Explainability, Observability, and Monitoring of Generative AI in Financial Services

OVERVIEW

Generative AI (GenAI) and Large Language Models (LLMs) are poised to revolutionize the financial services sector, offering transformative potential in personalized wealth management, automated report generation, synthetic data creation, and intelligent customer assistants. However, the chaining of thousands of probabilistic outputs in GenAI introduces a new layer of risk distinct from traditional discriminative AI where the risk-reward tradeoff can be established before release.

We invite proposals from academic researchers, AI ethicists, and applied ML engineers to develop novel frameworks that enhance the explainability, observability, and monitoring of Generative AI systems within the high-stakes financial environment.

CHALLENGE

Traditional monitoring and Explainable AI (XAI) techniques designed for predictive models (e.g., intent scoring, behavioral modeling) are insufficient for Generative AI due to the complexity and creativity of the outputs. The core challenges include:

Hallucination & Factuality: Detecting when a model convincingly generates incorrect financial advice, cites non-existent regulations, or fabricates market data.

Reasoning Opacity: Unlike regression models, LLMs rely on complex, multi-step reasoning chains (Chain of Thought) that are difficult to interpret and validate for correctness.

Prompt & Output Complexity: The variability in human inputs (prompts) and the vast output space (unstructured text and code) make it difficult to define static boundaries for monitoring.

Safety & Alignment Risks: Ensuring models do not generate biased investment recommendations, violate privacy regulations, or succumb to adversarial prompt injections (jailbreaking) in a production environment.

OBJECTIVES & SCOPE

The primary objective of this program is to fund research that creates tangible mechanisms to trust, trace, and validate Generative AI systems. Proposals should move beyond standard performance metrics (like BLEU or ROUGE scores) and focus on semantic correctness, safety, and auditability.

The scope of this call includes, but is not limited to:

- **Mechanistic Interpretability:** Research into the internal representations of LLMs to understand how specific financial concepts and reasoning patterns are encoded.
- **Semantic Monitoring:** Developing systems capable of evaluating the factual accuracy and logical consistency of generated financial text in real-time.
- **Flow Observability:** Creating tools to trace the lineage of a response—from the initial prompt through various retrieval-augmented generation (RAG) steps to the final output—to identify where failures occur.
- **Evaluation Frameworks:** Defining robust benchmarks for assessing the "safe" operation of GenAI in regulated financial contexts.

FOCUS AREAS

Proposals must address one or more of the following focus areas:

1. Explainability and Reasoning Verification

- Methods for visualizing and validating the "Chain of Thought" in financial models used for advisory or analysis.
- Techniques for attribution: Identifying and showing, which specific segments of training data or retrieved documents (in RAG systems) influenced a particular piece of financial advice.
- Tools that can self-verify mathematical calculations or logic within generated content (e.g., verifying a generated portfolio performance over some historical period).

2. Observability of Unstructured Outputs

- Real-time detection of hallucinations, confabulations, or factual errors in generated financial reports.
- Monitoring for tone drift, policy violations, or emerging bias in customer-facing chatbots.
- Systems that track the semantic intent of user prompts to detect malicious queries or attempted prompt injections.

3. RAG and Context Monitoring

- Frameworks for monitoring the *retrieval* component of RAG systems—ensuring the model is accessing the correct, up-to-date regulatory documents or market data.
- Detecting "context contamination" where irrelevant or noisy data degrades the quality of the generated response.
- Measuring the "groundedness" of a response—quantifying how much the model is sticking to verified facts versus inventing new information.

4. Safety, Guardrails, and Red Teaming as Monitoring

- Automated red-teaming methodologies that continuously probe GenAI systems for financial compliance violations (e.g., making guaranteed returns promises).
- Research on dynamic guardrails that can intercept and correct unsafe generations in real-time without disrupting the user experience.
- Developing feedback loops where human expert corrections are immediately assimilated into the model's monitoring logic to prevent recurrence.

Note that for all research themes, we welcome novel research ideas and approaches to the problem that may not be outlined in the brief.

Fidelity Center for Applied Technology LLC (FCAT®) provides innovative products, services, content and tools, as a service to its affiliates and as a subsidiary of FMR LLC. Based on user reaction and input, FCAT is better able to engage in technology research and planning for the Fidelity family of companies. Unless otherwise indicated, the information and items presented are provided by FCAT and are not intended to provide tax, legal, insurance or investment advice and should not be construed as an offer to sell, a solicitation of an offer to buy, or a recommendation for any security by any Fidelity entity or any third-party. Third-party trademarks and service marks are the property of their respective owners. All other trademarks and service marks are the property of FMR LLC or its affiliated companies. All rights reserved. © 2026.