JUPYTER, LLC
INTELLIGENT SYSTEMS

Enabled Financial Services Functions with Generative AI / LFMs

Services that enable Financial Services Functions with the power of Generative AI

SELECT KEY USE-CASES

- **Automated Customer Support**
  Leverage LLMs to provide 24/7 customer service via chatbots, enhancing customer experience with instant response to inquiries.
- **Fraud Detection and Prevention**
  Deploy LLMs to analyze transaction patterns and identify anomalies, helping to flag suspicious behavior and prevent fraud.
- **Investment Advisory**
  Deploy LLMs that aid investment advice decision-making processes by analyzing market trends, historical data, and customer profiles.
- **Risk Management**
  Implement LLMs to assess and predict risks by analyzing vast amounts of data, enabling financial institutions to develop more effective risk mitigation strategies and compliance measures.
- **Credit Scoring and Underwriting**
  Use LLMs to enhance credit scoring models by incorporating unstructured data, such as social media activity and customer reviews, alongside traditional credit worthiness inputs.
- **Regulatory Compliance Monitoring**
  Apply LLMs to monitor and interpret regulatory changes and compliance requirements, ensuring that financial institutions adhere to legal standards and reduce risks of non-compliance.
- **Document Processing & Analysis**
  Use LLMs to streamline processing of financial documents, such as loan applications, insurance claims, and legal agreements, enhancing efficiency and accuracy in document management and data extraction.

ADOPTION LIFECYCLE

Owing to their **generative power** as well as **reasoning prowess**, LFMs have become the brains behind intelligent applications. Operationalizing LFMs are involved undertakings and the following sections summarize how our **Aware Business Engineers** engage to enable Financial Service Processes.

MODEL SELECTION

Selecting the right model for a given use-case involves key considerations including**:** Consumption & Hosting Cost Strategy • Open Source vs. Proprietary • Base Expertise such as Text, Imaging, Audio, Video, Coding, or Reasoning • Parametric Size • Compute Footprint • Customizability • and more.

MODEL CUSTOMIZATION

Selected models must be customized to become **subject experts** in the domains they will serve. This is a collaborative, iterative effort between our engineers and business domain experts. Stages include:

- Selecting & deploying **Evaluation Frameworks** to systematically and objectively capture LFM performance.
- **Engineered Prompt Testing** to baseline LFM subject-domain expertise, informing on additional business data needed to improve it.
- Implementing **Retrieval Augment Generation** (**RAG**) vector databases store to store knowledgebase documents for non-parametric, in-context LFM learning.
- **Fine Tuning** to parametrically improve an LLMs ability to output particular styles and structures, to boost existing knowledge, and to teach it very complex instructions.
- Implement LLM **Function Calling** to enable web searches, calculations, IDE coding and use of external tools.

LLM-OPS PIPELINING

Best-practice architectures, techniques and tools for the operationalized management of LFMs in production environments.

CASE STUDY

Clients of a F500 **investment advisory firm** have access to tens of thousands of **equities analyst articles** published every month. To help clients narrow this vast library to only articles aligned with their investment objectives, we implemented a **recommender engine** suggesting targeted articles in real-time.

ORIGINAL SULUTION

Implemented circa ~2013, the original recommender was deployed using a traditional **big data pipeline** and **machine learning** algorithms including **collaborative filtering** (i.e., item and user **similarity**) as well as **clustering** (e.g., k-Means). This solution worked well, but lacked what modern application families now enjoy; chiefly, the peculiarity of allowing natural language interaction between systems and users.

MODERNIZED SULUTION

By replacing the original pipeline with one featuring the **Llama-3 LLM** connected to a **RAG database** storing the equities article knowledgebase, the original recommender capability was preserved, while the ability for clients to **actively interact** with the knowledgebase was added. Clients can submit 1-click suggestions, or enter rich search criteria at the input prompt.