FINOS

Fintech
Open Source
Foundation

FINOS AI Strategic Initiative

Building a Taxonomy for AI Evals in Finance

From use cases to benchmarks: a shared framework

THE LINUX FOUNDATION

The Challenge & Our Approach

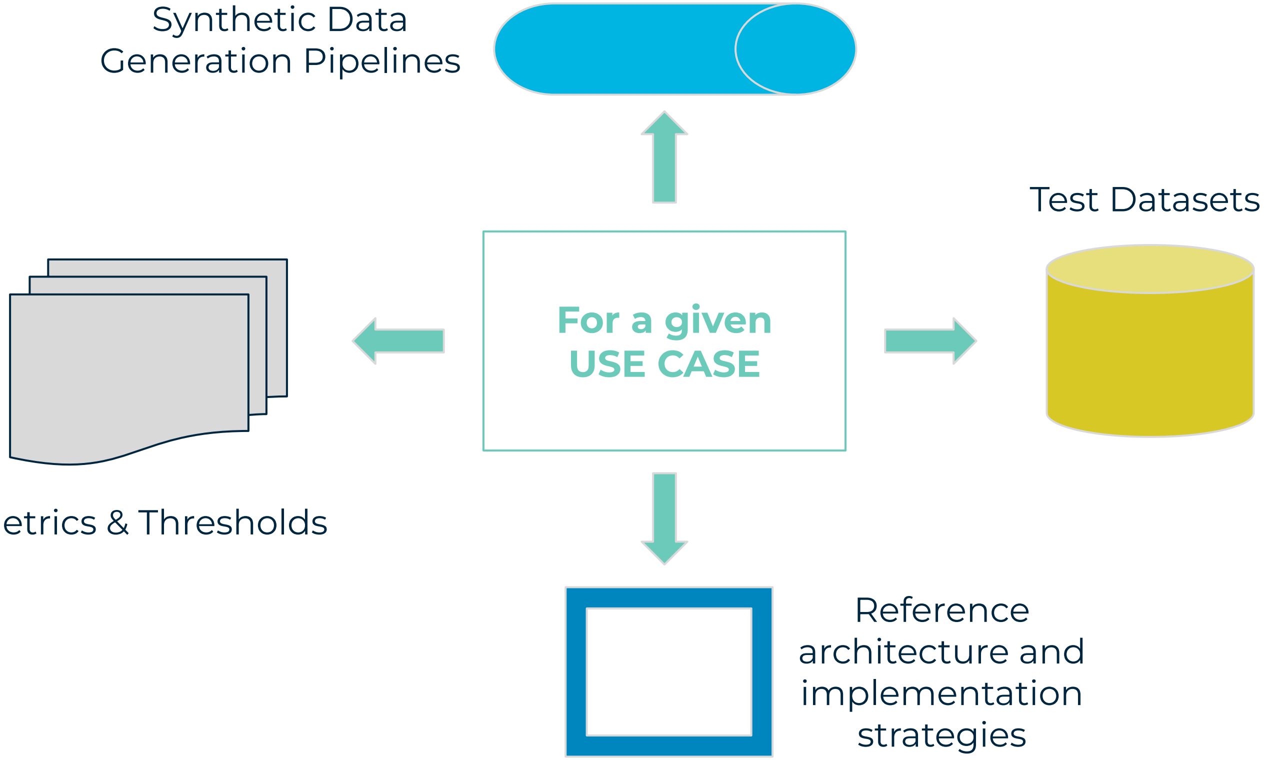
⚠️ Challenges

- AI systems are non-deterministic → no guarantees that it works
- Financial tasks rarely have one “correct” answer

💡 Approach

- Anchor evaluations in **financial use cases taxonomy**
- **Taxonomy Dimensions**: Use case → Risks → Metrics
- **Demo and Example repos** in FINOS Labs

Deliverables



Synthetic Data Generation Pipelines

Test Datasets

For a given USE CASE

Metrics & Thresholds

Reference architecture and implementation strategies

Next Steps

Gather workshop and techsprint artefacts	September	DONE
Identify maintainers	September	WIP
Literature review	October	
Present the project to FINOS TOC	October	
Set up FINOS infrastructure (periodic meetings, repo)	October	
Publish template repos & examples	November	
Pilot with financial institutions	Q1 2026	
Expand shared taxonomy across industry	Q2 2026	

Artefact 1: Defining Initiative Identity

IS

- Forum to define relevant use cases and challenges for FSIs
- Use Cases
- Customer Management Use Cases
- A taxonomy of use cases and eval strategy
- Financial Services use cases catalog <-> eval mapping
- Detecting Hallucinations and inaccurate output
- Hallucination controls
- Reducing bias & discrimination
- Bias / Fairness methodologies
- Bias controls
- Produce test datasets
- Synthetic test data for use cases
- Set of common performance guardrails
- Task benchmarks
- Definition of limits of the framework
- An assessment of open source evals framework in FSI
- Suggest oundational eval stack
- Think tank on eval strategies relevant to FSI
- Reference architecture for evals of different types of AI
- Measuring environmental impact
- Tool performance reporting framework
- Guidance material
- Repeatability

DOES

- Produce Use Cases
- Define and maintain taxonomy
- Map USE CASE -> RISK -> METRICS
- Use Cases specific min guardrails testing spec
- Eval pattern and architecture (testing, monitoring)
- Implementation strategies
- AIGF Mapping and AIGF based eval implementation
- Sample repo with code / tools
- Guidance / Easy to follow
- Techsprint for use case & risk definition
- Use Case agnostic risk mitigation
- Task benchmarks
- Model transparency
- LLM-as-judge tuning and validation
- Trustworthiness scoring methodologies
- Produce test datasets
- Fund eval research (methodology and data)
- Synthetic test data for use cases taxonomy
- Way to measure environmental impact
- Efficiency for evals (LLM-as-judge)

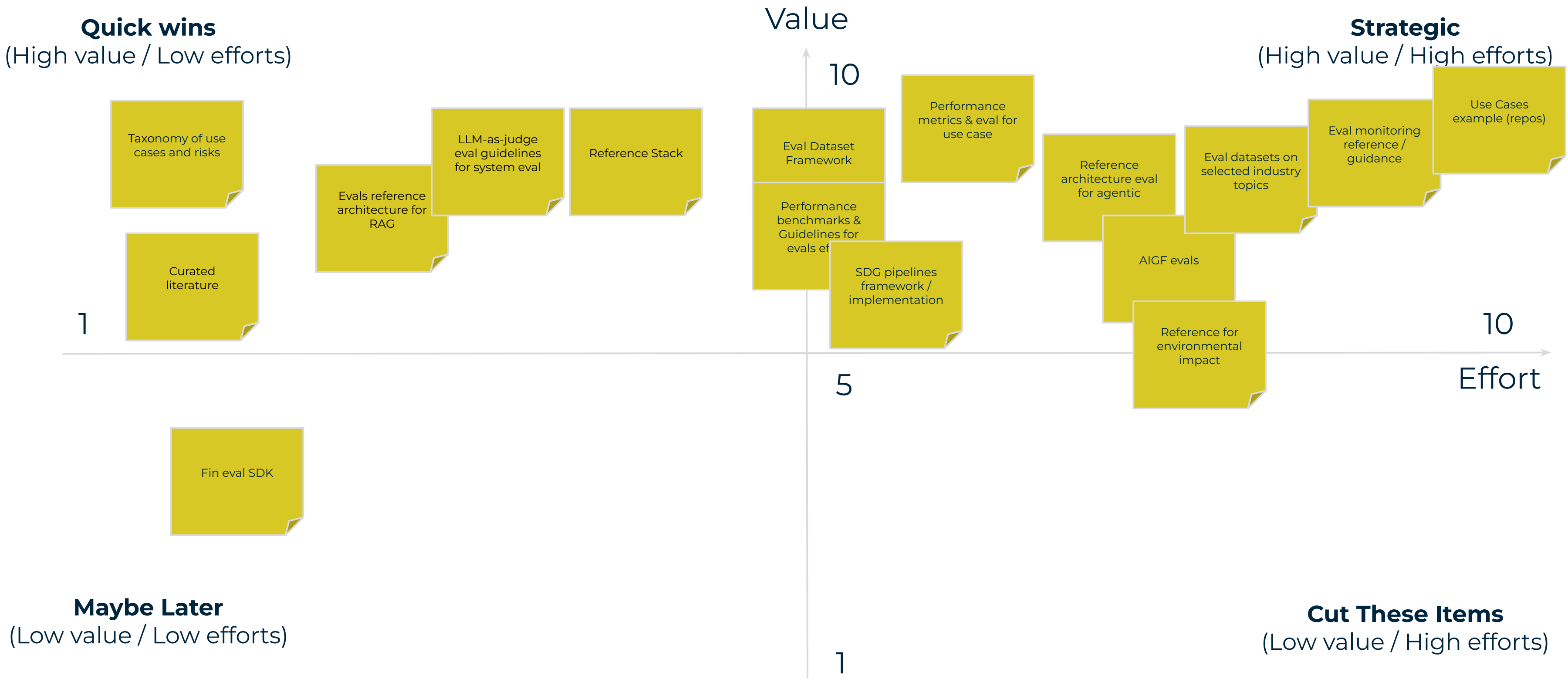
IS NOT

- Tech implementation for testing & verification
- Internal compliance rules
- Definition of business value
- LLM / model factory
- Orchestration platform / ecosystem for accessing evals
- Technical working group building AI solutions

DOES NOT

- Eval leaderboard
- Build software
- Test / eval specific LLM

Artefact 2: Prioritizing Deliverables



Quick wins
(High value / Low efforts)

Strategic
(High value / High efforts)

Maybe Later
(Low value / Low efforts)

Cut These Items
(Low value / High efforts)

Value

10

5

1

Effort

1

10

Taxonomy of use cases and risks

Curated literature

Evals reference architecture for RAG

LLM-as-judge eval guidelines for system eval

Reference Stack

Fin eval SDK

Eval Dataset Framework

Performance benchmarks & Guidelines for evals ef

SDG pipelines framework / implementation

Performance metrics & eval for use case

Reference architecture eval for agentic

AIGF evals

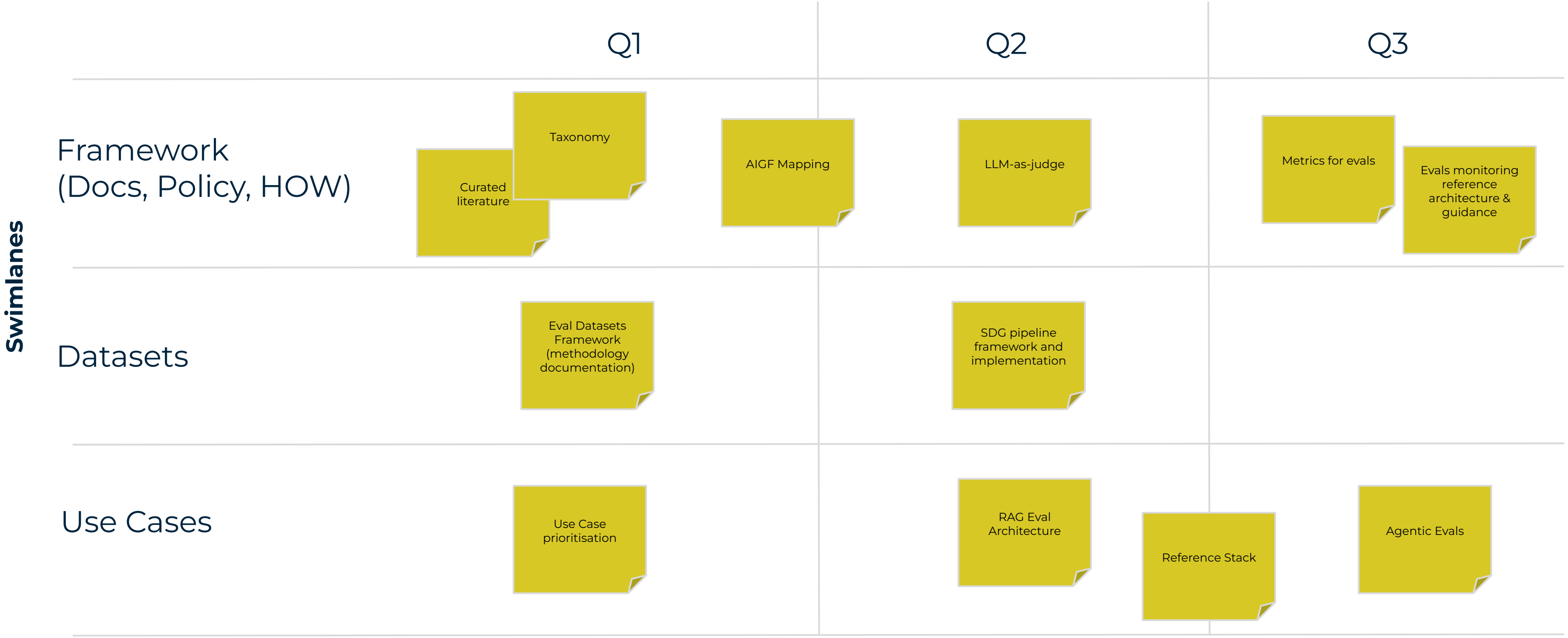
Reference for environmental impact

Eval datasets on selected industry topics

Eval monitoring reference / guidance

Use Cases example (repos)

Artefact 3: FINOS AI Eval Framework Roadmap



Swimlanes

	Q1	Q2	Q3
Framework (Docs, Policy, HOW)	Curated literature Taxonomy AIGF Mapping	LLM-as-judge	Metrics for evals Evals monitoring reference architecture & guidance
Datasets	Eval Datasets Framework (methodology documentation)	SDG pipeline framework and implementation	
Use Cases	Use Case prioritisation	RAG Eval Architecture Reference Stack	Agentic Evals

Artefact 4: FINOS Labs and Hugging Face

- **Agent Quickstart** (with pre-built connectivity to inference endpoints): https://github.com/finos-labs/Agent-QuickStart/blob/main/README.md

- **Synthetic data**: https://huggingface.co/finosfoundation/dataset

- **Literature**: review this recent paper, which surveys emerging approaches to evaluation in the generative AI era and frames key challenges FINOS members are solving for.

- **Use Cases:** Check out this FS use cases list. Ideally we would like to see some of those become "sample agents connected to the techsprint prototype", so that we can show how AI-based, FS-specific workflows could be evaluated and tested at scale.

FINOS

Fintech
Open Source
Foundation

FINOS AI Strategic Initiative

Evaluation and Benchmarking Suite Scoping Workshop (19/09/2025)

THE LINUX FOUNDATION

Evaluation and Benchmarking Suite

General comment: is this the right order? Maybe swap "Vendors" with "Use cases"?

Focus on "System-level" not model-level

Define and maintain taxonomy (multi-layers e.g. tasks etc) (see BIAN for example)

Open Datasets

LLM-as-judge tuning and validation

Common tests

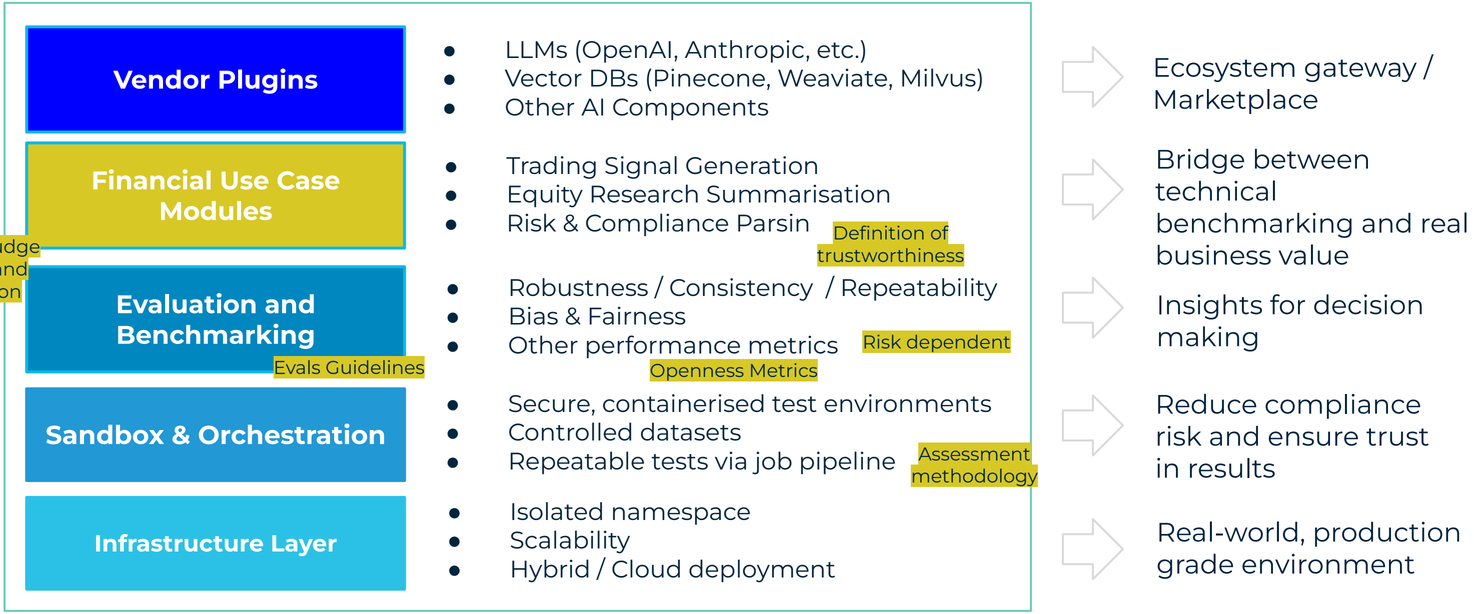
Reference Architectures for the evals

Run time vs. Design

Implementation strategies + code (repo) for the arch. Patterns

Energy spent vs. return

Environmental metrics



Vendor Plugins

- LLMs (OpenAI, Anthropic, etc.)
- Vector DBs (Pinecone, Weaviate, Milvus)
- Other AI Components

Ecosystem gateway / Marketplace

Financial Use Case Modules

- Trading Signal Generation
- Equity Research Summarisation
- Risk & Compliance Parsin

Definition of trustworthiness

Bridge between technical benchmarking and real business value

Evaluation and Benchmarking

Evals Guidelines

- Robustness / Consistency / Repeatability
- Bias & Fairness
- Other performance metrics

Risk dependent

Openness Metrics

Insights for decision making

Sandbox & Orchestration

- Secure, containerised test environments
- Controlled datasets
- Repeatable tests via job pipeline

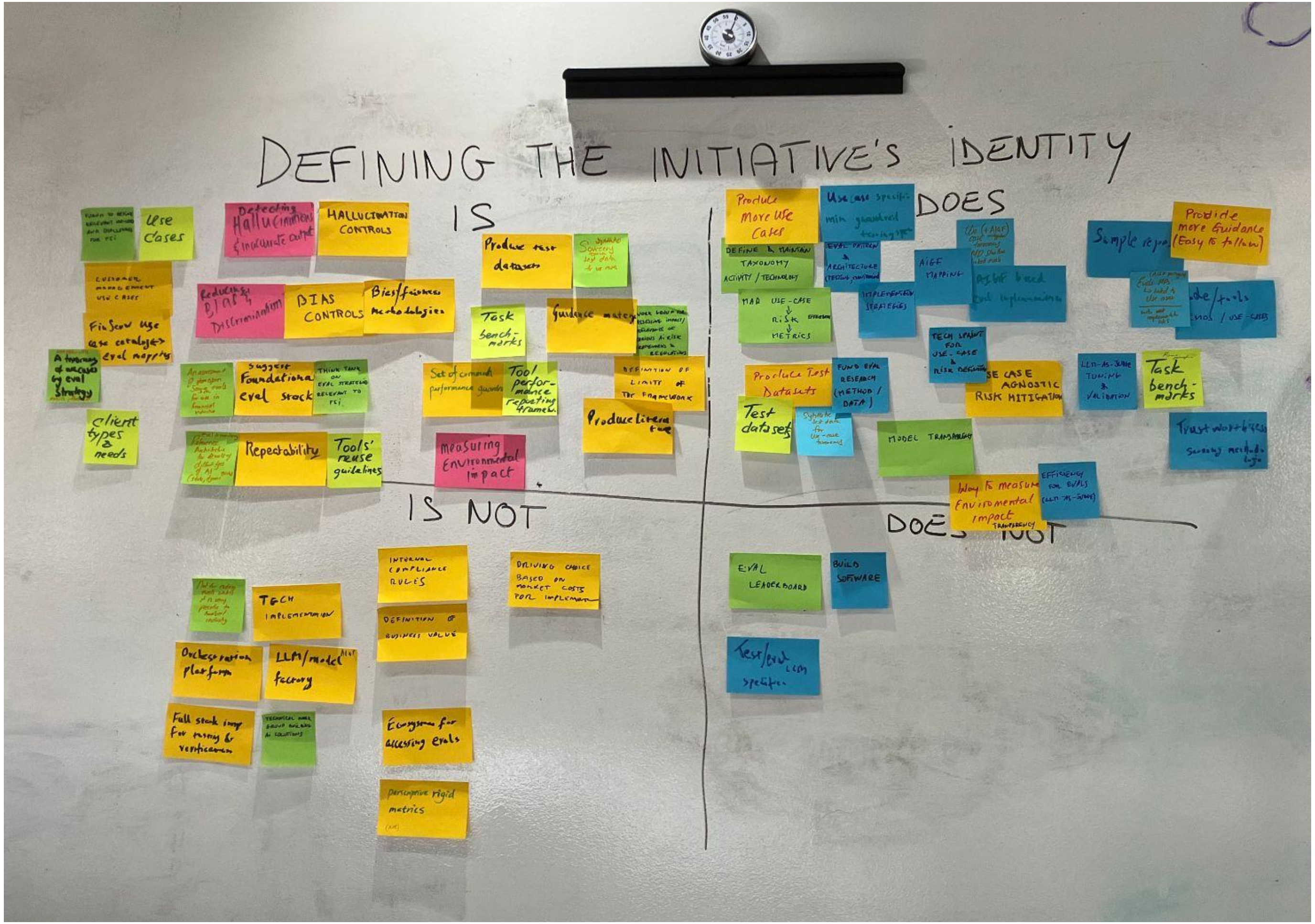
Assessment methodology

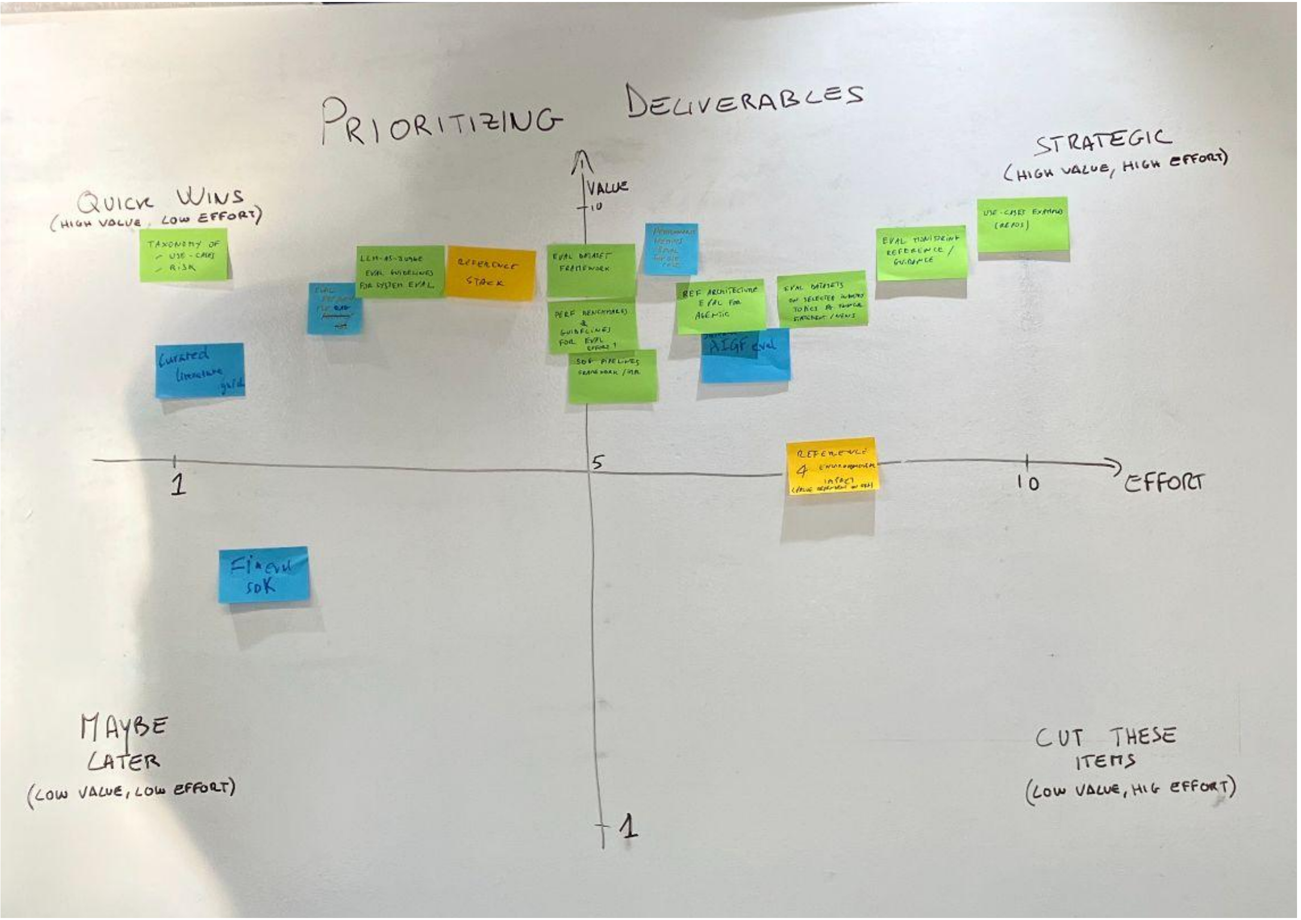
Reduce compliance risk and ensure trust in results

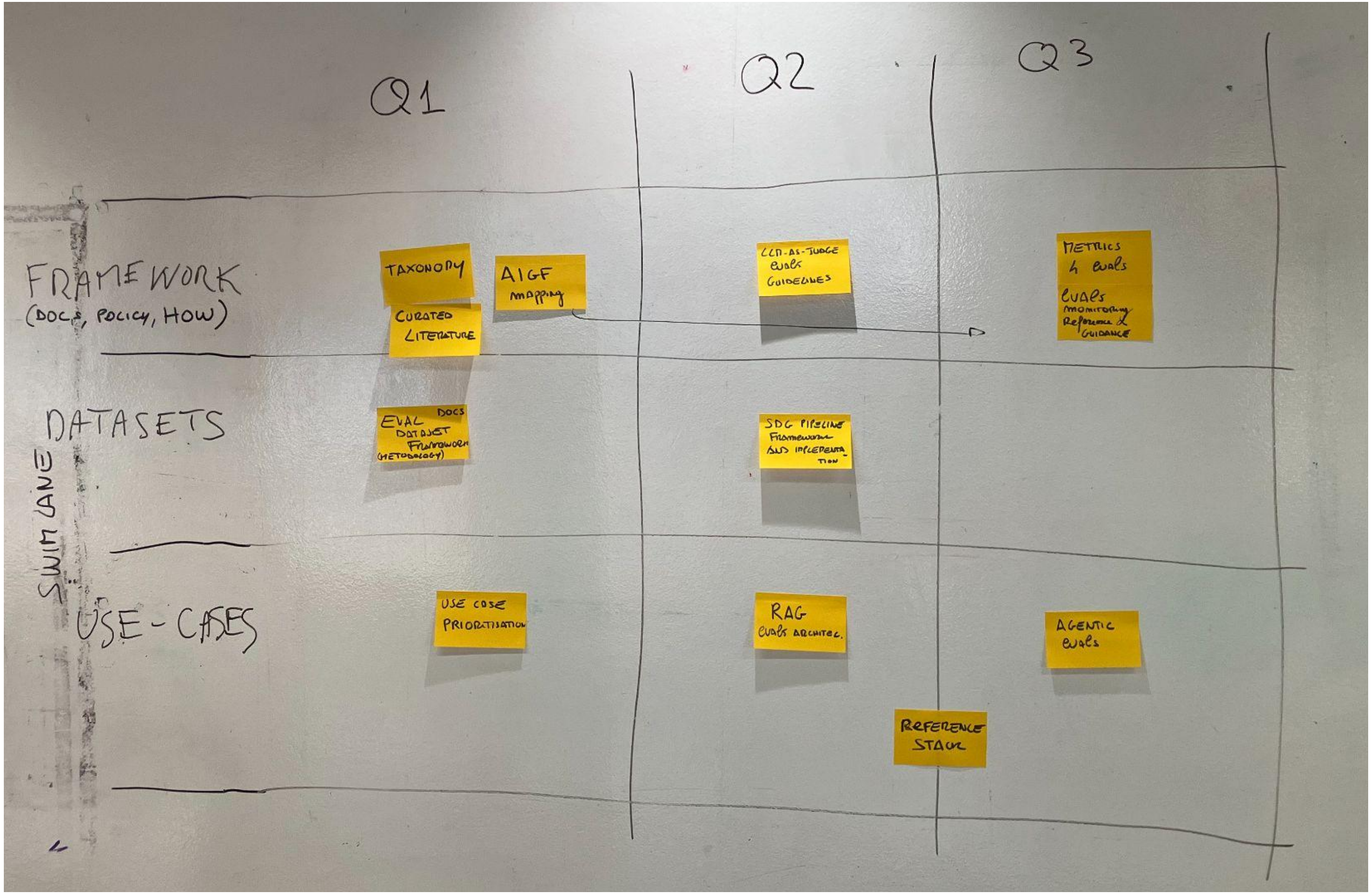
Infrastructure Layer

- Isolated namespace
- Scalability
- Hybrid / Cloud deployment

Real-world, production grade environment


DEFINING THE INITIATIVE'S IDENTITY
IS
Use Cases
CUSTOMER MANAGEMENT USE CASES
FinServ use case catalogues eval mapping
A taxonomy of usecases by eval strategy
client types & needs
Detecting Hallucinations & inaccurate output
HALLUCINATION CONTROLS
Reducing Bias & Discrimination
BIAS CONTROLS
Bias/fairness methodologies
Suggest Foundational eval stack
THINK TANK ON EVAL STRATEGIES RELEVANT TO FSI
Repeatability
Tools' reuse guidelines
Produce test datasets
Task bench-marks
Guidance material
Set of common performance guardrails
Tool performance reporting framew.
DEFINITION OF LIMITS OF THE FRAMEWORK
Produce literature
measuring Environmental impact
DOES
Produce more Use Cases
Use case specific min guardrail testing
DEFINE & MAINTAIN TAXONOMY ACTIVITY / TECHNOLOGY
EVAL PATTERN & ARCHITECTURE
MAP USE-CASE → RISK → METRICS
IMPLEMENTATION STRATEGIES
AIGF MAPPING
AIGF based eval implementation
Simple repo
Provide more Guidance (Easy to follow)
/tools
/ USE-CASES
TECH SPRINT FOR USE-CASE & RISK DEFINITION
USE CASE AGNOSTIC RISK MITIGATION
LLM-AS-JUDGE TUNING & VALIDATION
Task bench-marks
Produce Test Datasets
FUND EVAL RESEARCH (METHOD / DATA)
Test datasets
Synthetic test data for use-case
MODEL TRANSPARENCY
Trustworthiness scoring methodology
Way to measure Environmental impact
TRANSPARENCY
EFFICIENCY FOR EVALS (LLM-AS-JUDGE)
IS NOT
TECH IMPLEMENTATION
Orchestration platform
LLM/model factory
Full stack imp for testing & verification
TECHNICAL WORK GROUP BUILDING AI SOLUTIONS
INTERNAL COMPLIANCE RULES
DEFINITION OF BUSINESS VALUE
DRIVING CHOICE BASED ON MARKET COSTS FOR IMPLEMENTATION
Ecosystem for accessing evals
perscriptive rigid metrics
DOES NOT
EVAL LEADERBOARD
BUILD SOFTWARE
Test/eval LLM specific


PRIORITIZING DELIVERABLES
QUICK WINS
(HIGH VALUE, LOW EFFORT)
STRATEGIC
(HIGH VALUE, HIGH EFFORT)
MAYBE LATER
(LOW VALUE, LOW EFFORT)
CUT THESE ITEMS
(LOW VALUE, HIG EFFORT)
VALUE
10
5
1
1
10
EFFORT
TAXONOMY OF
- USE-CASES
- RISK
Curated literature guide
LLM-AS-JUDGE EVAL GUIDELINES FOR SYSTEM EVAL
REFERENCE STACK
EVAL DATASET FRAMEWORK
PERF BENCHMARKS & GUIDELINES FOR EVAL EFFORT ?
SDF PIPELINES FRAMEWORK / MR
REF ARCHITECTURE EVAL FOR AGENTIC
AIGF eval
EVAL DATASETS ON SELECTED TOPICS
EVAL MONITORING REFERENCE / GUIDANCE
USE-CASES EXAMPLE (REPOS)
REFERENCE 4 ENVIRONMENTAL IMPACT
FinEval SDK



SWIM LANE	Q1	Q2	Q3
FRAMEWORK (DOCS, POLICY, HOW)	TAXONOMY CURATED LITERATURE AIGF mapping	LLM-AS-JUDGE evals GUIDELINES	METRICS 4 evals evals monitoring Reference & GUIDANCE
DATASETS	EVAL DATASET Framework (METODOLOGY) DOCS	SDG PIPELINE Framework AND IMPLEMENTATION	
USE-CASES	USE CASE PRIORITISATION	RAG evals ARCHITEC. REFERENCE STACK	AGENTIC evals

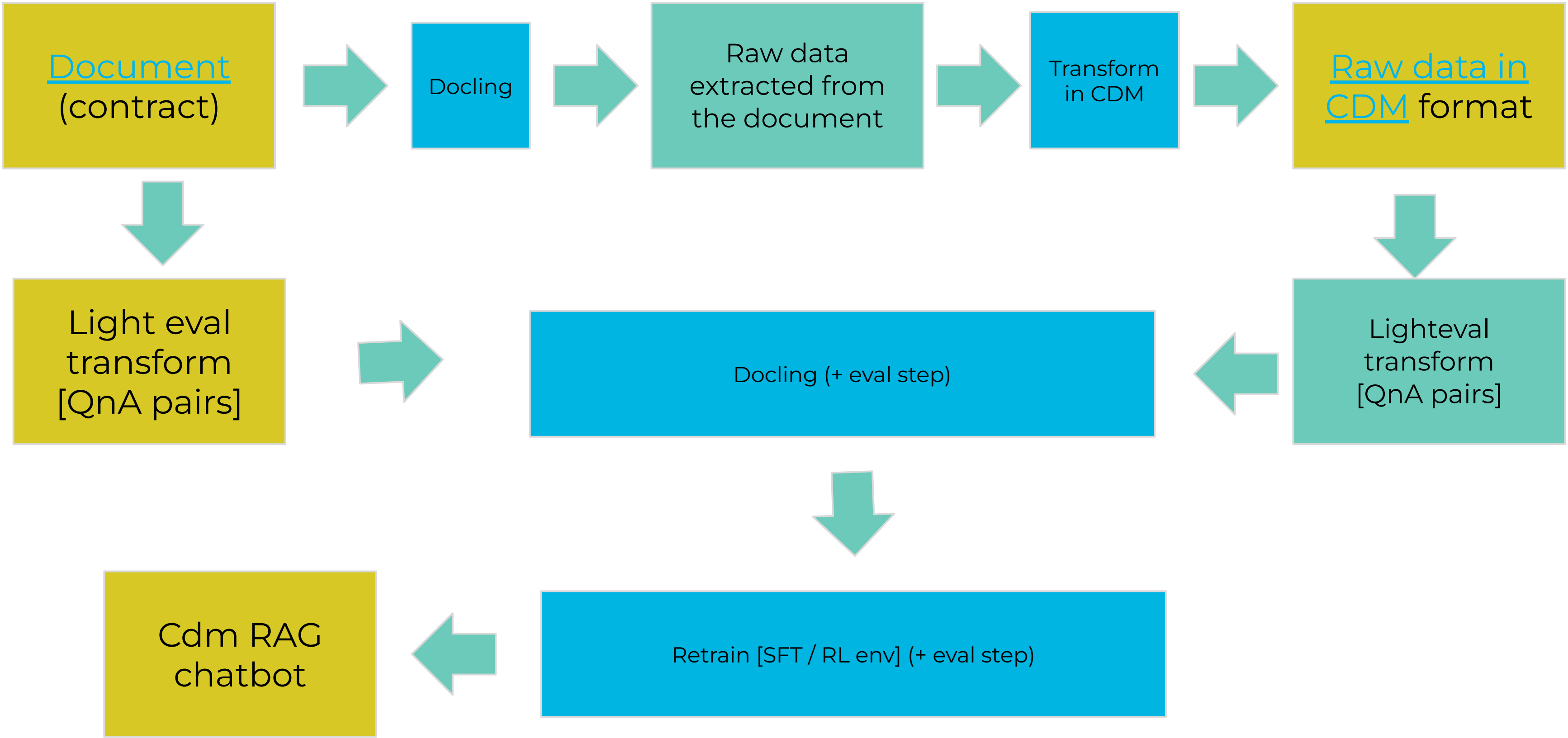
FINOS

Fintech
Open Source
Foundation

FINOS AI Strategic Initiative

Evaluation and Benchmarking Suite
Scoping Techsprint (20/09/2025)

THE
LINUX
FOUNDATION



Document (contract)

Docling

Raw data extracted from the document

Transform in CDM

Raw data in CDM format

Light eval transform [QnA pairs]

Docling (+ eval step)

Lighteval transform [QnA pairs]

Retrain [SFT / RL env] (+ eval step)

Cdm RAG chatbot

Joseph:

- Transforming regulation into CDM schema for derivative reporting

  1. Dataset natural language queries generating CDM code snippet using RAG -> output Response
  2. Embed CDM reference documentation and retrieve based on generated code snippet -> output Citation
  3. Output -> Response with Citation

- Transform CDM (code) dataset to lighteval format

- Transform DORA to lighteval format

- Compress + finetune docling for spiky performance

FINOS Fintech Open Source Foundation

Appendix

THE LINUX FOUNDATION

Evals & Benchmarking (1/5)

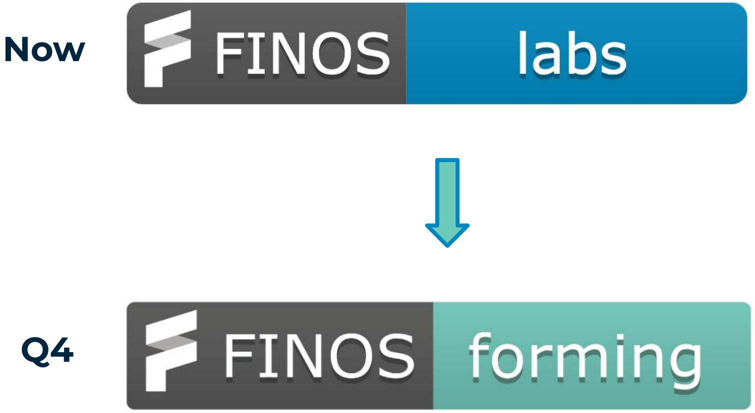
GOAL: Establish a common, open, and transparent evaluation and benchmarking suite for Generative AI (GenAI) applications in financial services.

“I believe that we all recognize that general-purpose AI benchmarks often fall short for the unique requirements of the finance sector. Therefore, a key direction for this initiative is to ***enhance existing evaluation methods.*** *But there are likely other challenges we need to consider, including and not limited to:*

- *The fact that* ***current model-level evaluations often do not fully capture the complexity of AI systems****, which combine AI and non-AI components,* ***nor do they sufficiently address their behavior in specialized domains****.*
- *The* ***challenge*** *of evaluating LLM-based agents presents specific challenges, particularly in assessing their* ***cost-efficiency, safety, and robustness, and in developing fine-grained and scalable evaluation methods.***
- *The pitfall of "****safetywashing****," where general capability improvements in evaluation might be misinterpreted as advancements in safety.“*

Vincent Caldeira, Red Hat CTO for APAC, FINOS TOC

Evals & Benchmarking (2/5)



Now

FINOS labs

Q4

FINOS forming

1. Workshop

- 19 Sept London
- Project Scope Statement and “Must-Have” feature backlog

→ Register here

2. Techsprint

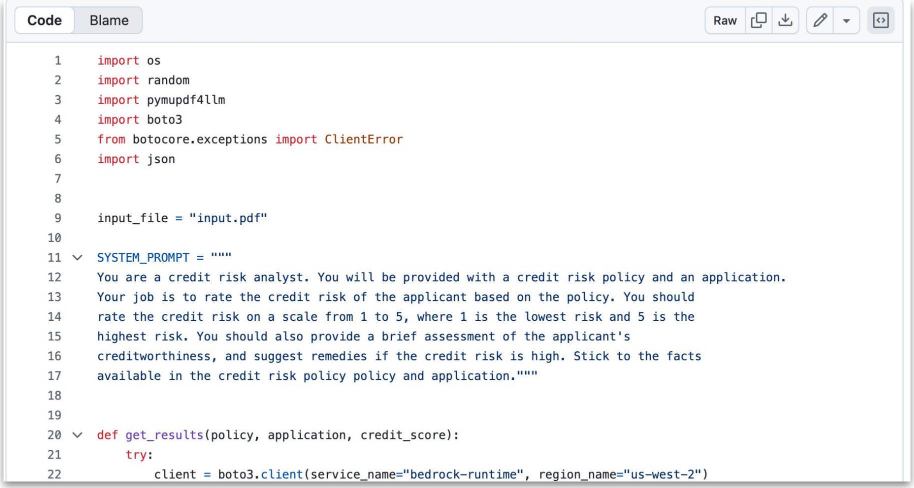
- 20 Sept London and Virtual Red Hat NVIDIA NayaOne
- Hacking a common, open, and transparent evaluation & benchmarking suite for GenAI in financial services

→ Register here

Evals & Benchmarking (3/5)

GOAL: FINOS Labs hosts plenty of use cases that can be used as "**full working example of an agent with dependencies**"

Credit Risk Decisioning: GitHub repo file



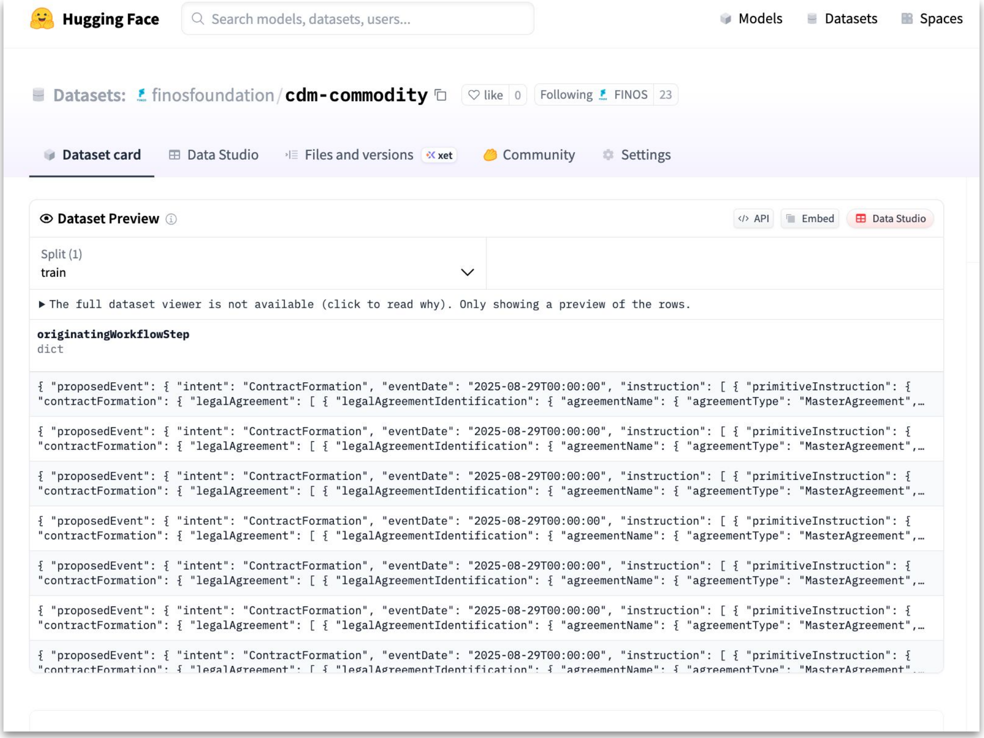
```python
import os
import random
import pymupdf4llm
import boto3
from botocore.exceptions import ClientError
import json


input_file = "input.pdf"

SYSTEM_PROMPT = """
You are a credit risk analyst. You will be provided with a credit risk policy and an application.
Your job is to rate the credit risk of the applicant based on the policy. You should
rate the credit risk on a scale from 1 to 5, where 1 is the lowest risk and 5 is the
highest risk. You should also provide a brief assessment of the applicant's
creditworthiness, and suggest remedies if the credit risk is high. Stick to the facts
available in the credit risk policy policy and application."""


def get_results(policy, application, credit_score):
    try:
        client = boto3.client(service_name="bedrock-runtime", region_name="us-west-2")
```

Evals & Benchmarking (4/5)

Our partner **NayaOne** has contributed a large sample of **CDM synthetic trade records** to power experimentation:

- 10 product types + 21 sub-categories (swaps, options, forwards, exotics)
- Strong emphasis on options, energy, and metals — reflecting real market activity
- 895 unique market identifiers to ensure diversity + realism

The commodity subset is now openly available on [Hugging Face]()



Datasets: finosfoundation/**cdm-commodity**

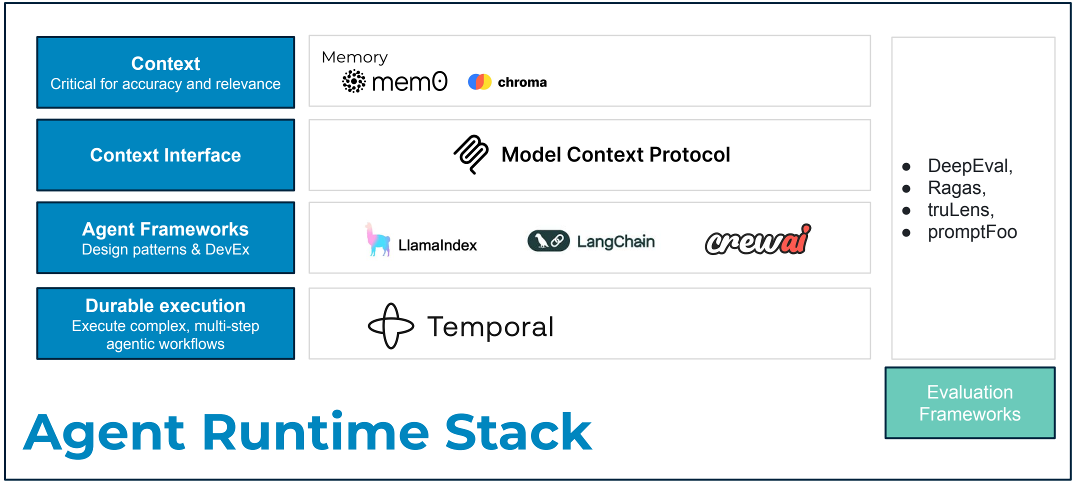
▸ The full dataset viewer is not available (click to read why). Only showing a preview of the rows.

originatingWorkflowStep
dict

```
{ "proposedEvent": { "intent": "ContractFormation", "eventDate": "2025-08-29T00:00:00", "instruction": [ { "primitiveInstruction": {
"contractFormation": { "legalAgreement": [ { "legalAgreementIdentification": { "agreementName": { "agreementType": "MasterAgreement",…

{ "proposedEvent": { "intent": "ContractFormation", "eventDate": "2025-08-29T00:00:00", "instruction": [ { "primitiveInstruction": {
"contractFormation": { "legalAgreement": [ { "legalAgreementIdentification": { "agreementName": { "agreementType": "MasterAgreement",…

{ "proposedEvent": { "intent": "ContractFormation", "eventDate": "2025-08-29T00:00:00", "instruction": [ { "primitiveInstruction": {
"contractFormation": { "legalAgreement": [ { "legalAgreementIdentification": { "agreementName": { "agreementType": "MasterAgreement",…

{ "proposedEvent": { "intent": "ContractFormation", "eventDate": "2025-08-29T00:00:00", "instruction": [ { "primitiveInstruction": {
"contractFormation": { "legalAgreement": [ { "legalAgreementIdentification": { "agreementName": { "agreementType": "MasterAgreement",…

{ "proposedEvent": { "intent": "ContractFormation", "eventDate": "2025-08-29T00:00:00", "instruction": [ { "primitiveInstruction": {
"contractFormation": { "legalAgreement": [ { "legalAgreementIdentification": { "agreementName": { "agreementType": "MasterAgreement",…

{ "proposedEvent": { "intent": "ContractFormation", "eventDate": "2025-08-29T00:00:00", "instruction": [ { "primitiveInstruction": {
"contractFormation": { "legalAgreement": [ { "legalAgreementIdentification": { "agreementName": { "agreementType": "MasterAgreement",…

{ "proposedEvent": { "intent": "ContractFormation", "eventDate": "2025-08-29T00:00:00", "instruction": [ { "primitiveInstruction": {
"contractFormation": { "legalAgreement": [ { "legalAgreementIdentification": { "agreementName": { "agreementType": "MasterAgreement",…
```

Evals & Benchmarking (5/5)

We are preparing tooling for the techsprint participants



Agent Runtime Stack

Context
Critical for accuracy and relevance

Memory
mem0
chroma

Context Interface

Model Context Protocol

Agent Frameworks
Design patterns & DevEx

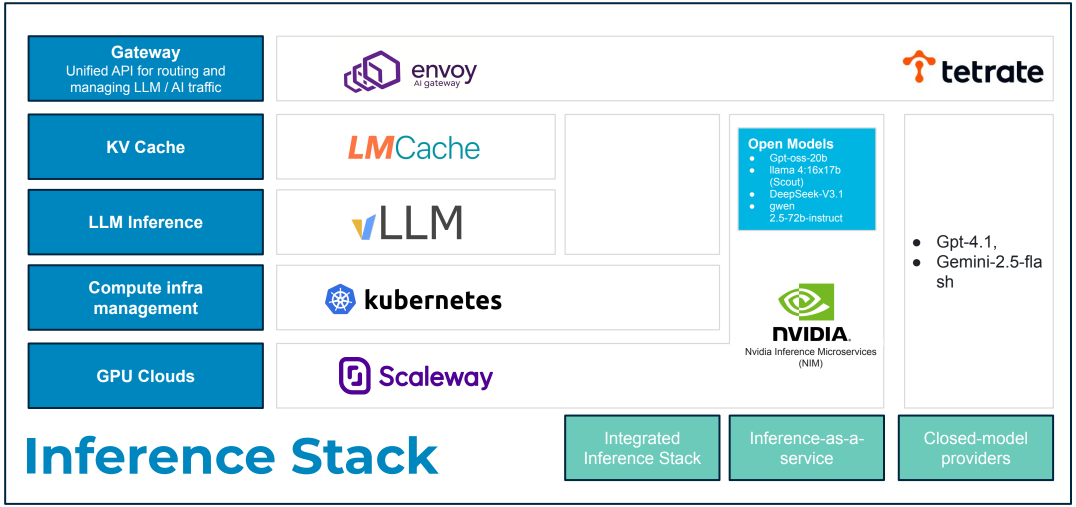
LlamaIndex
LangChain
crewai

Durable execution
Execute complex, multi-step agentic workflows

Temporal

Evaluation Frameworks

- DeepEval,
- Ragas,
- truLens,
- promptFoo



Inference Stack

Gateway
Unified API for routing and managing LLM / AI traffic

envoy
AI gateway
tetrate

KV Cache

LMCache

LLM Inference

vLLM

Compute infra management

kubernetes

GPU Clouds

Scaleway

Integrated Inference Stack

Open Models

- Gpt-oss-20b
- llama 4:16x17b (Scout)
- DeepSeek-V3.1
- gwen 2.5-72b-instruct

NVIDIA
Nvidia Inference Microservices (NIM)

Inference-as-a-service

- Gpt-4.1,
- Gemini-2.5-flash

Closed-model providers