Keeping Models Safe: A Practical Guide to Independent Validation in Finance

Dr Maria Kalantzaki
Senior Manager
Non-Financial Risk & Artificial Intelligence
Virgin Money

What happens when models go wrong?

Inappropriate data assumptions and simplified calculations that smoothed volatility and masked potential losses.

Failure to capture the risks associated with complex derivative trades, partly due to *reliance on historical data*

Trading loss of
$2 billion

LCTM (Long Term Capital Management)

Lack of model validation and external oversight

-**Assumption of Stable Correlations**
-Underestimation of market events
-The models relied on historical data and **normal distribution assumptions**.
-Liquidity Risk Ignored

Total losses of
$4.6 billion

No true independence: Many models were validated internally.
Blind faith in quantitative models.
Noone challenged the assumptions

- **Faulty Assumptions** about Housing Prices
- Misjudged Correlations (defaults not correlated)
- Ignored Liquidity & Systemic Risk

US banks failing: 25

MONEY

How can we prevent failures; through Independent Model Validation (IMV)

- In **2011**, the presence of IMV became a **regulatory requirement** with the U.S. Federal Reserve's **SR 11-7** guidance.

-**Basel III (2010–2011) from Basel Committee on Banking Supervision (BCBS)**. :stress testing scenarios.

-The **Market Risk Framework (FRTB, 2016) introduces stricter requirements : formal model approval, independent review, performance monitoring.**

-**European Central Bank (ECB)**'s Guide to Internal Models (2018): IMV is a formal requirement; Model cannot go live without Independent Review!

- Every release tends to be stricter than the previous one.

- Artificial Intelligence (AI) makes things more complicated and creates needs for elevated controls

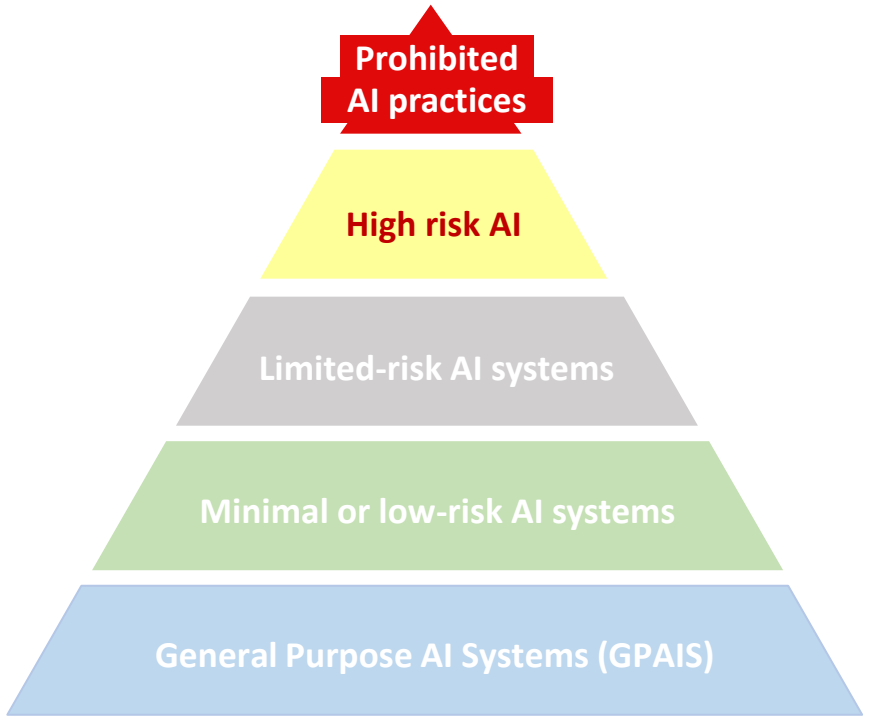
IMV Rights vs Responsibilities

Action	IMV Power	IMV Duty
Challenge assumptions	Call out weak logic, bad data, or shaky theory.	Back it up with testing and evidence.
Say “No” to a model	Withhold endorsement until it meets the bar.	Document why and escalate properly.
Send a model back	Demand redevelopment or fixes.	Ensure changes align with standards and regs.
Resist pressure	Stay fully independent of model developers or management.	Guard objectivity, even when it’s unpopular.
Set the validation tests	Choose the tools: benchmarking, stress tests, backtesting.	Tailor the depth to the model’s materiality and risk.
Speak to regulators	Provide an unfiltered, independent view.	Keep it consistent, transparent, and traceable.
Give the green light	Approve a model for use.	Only when standards and compliance are rock-solid.
Pull the plug	Block use of a risky or non-compliant model.	Flag it to senior management and record the rationale.

What are the control frameworks for IMV.

EU AI Act

Article 55(1): "providers of general-purpose AI models with systemic risk shall (…) **perform model evaluation** in accordance with standardised protocols and tools reflecting the state of the art, including **conducting and documenting adversarial testing of the model** with a view to identifying and mitigating systemic risks"

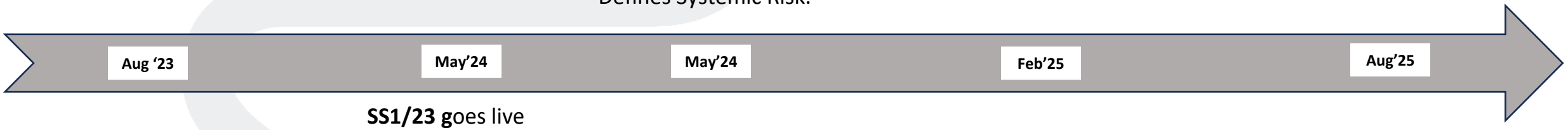


SS1/23

Model identification & risk classification	Clearly measurement of risk quantification
Governance	Clear and Hierarchical governance paths
Model development, implementation & use	Robust presence of IMV across the whole lifecycle
Independent model validation	Clear presence of IMV and MRM frameworks within organizations.
Model risk mitigants	Do we have sufficient model mitigants

MONEY

Ensuring Trust in GenAI: The Virgin Money Validation Journey



Model Risk Management (MRM) framework

🛡️ MRM Policy AI Adaptation

- 🧠 **Inclusion of AI Methodologies**

📋 **New Requirements for AI Models**

- 🔍 **Explainability**: Models must be interpretable and transparent
- 🧪 **Robustness**: Models must perform reliably under varied conditions

🧩 Defining AI Use Cases: Why Context Matters

- 📘 **Use Case Definition**

Clear use case definition is central to AI validation standards

- ⚠️ **Same Model, Different Risks**

A single AI model can pose varying risks depending on its application (*UK AI White Paper)*

- 🔍 **Risk Depends on Context**

Severity of AI risk are shaped by use case and context *(GVK 'Initial Guidance for Regulators', 2024)*

IMV Team: AI Validation Strategy

- ✅ **Proactive & Risk-Aware**

Adopted a forward-looking approach to AI solution validation

- 📌 **Use Case Integration**

Embedded use case definitions into model validation standards

- ⚙️ **AI/ML-Driven Framework**

Developed standards tailored for AI/ML validation

Model Validation Thematic findings

Theme	*Key Pattterns*	*Examples / Notes*
Data Issues	Limited representation of data	Minority groups or rare fraud patterns under-represented, leading to biased or brittle models.
	Data imbalance	Fraud detection: 0.1% fraud vs. 99.9% non-fraud → inflated accuracy but poor fraud recall.
	Very coarse target definitions	Targets defined at aggregate level (e.g., monthly defaults) instead of transaction-level events → reduces granularity.
Model Performance	Short-lived models	Particularly in fraud detection → adversaries adapt fast, requiring constant retraining.
	Over-reliance on traditional metrics	Metrics like AUC or Gini used as gold standard, ignoring business impact or calibration.
	Limited explainability (esp. GenAI)	Hard to justify predictions in regulated domains; black-box nature increases risk.
	Poor robustness	Models may fail under distribution shifts, adversarial attacks, or new fraud tactics.
	Fairness & bias concerns	Protected groups may receive disproportionately high false positives/negatives.
Process / Governance	Weak validation rigor	Limited backtesting or stress testing. Validation may rely too much on in-sample metrics.
	Documentation gaps	Validation documentation often insufficient for regulatory audits.

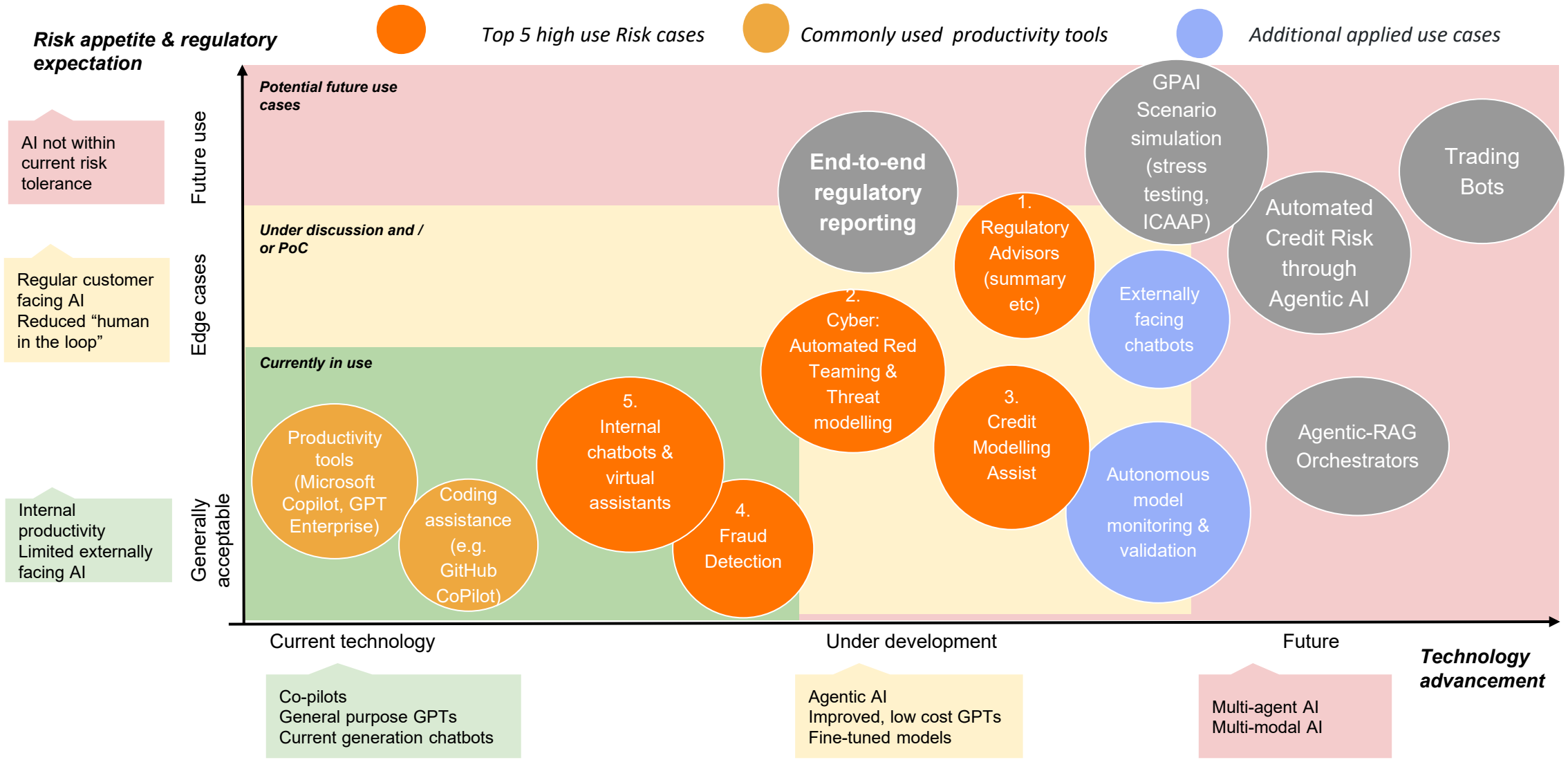
A view into AI challenges

LLMs push the need for intricate testing

There are fundamental differences between testing Large Language Models and ‘traditional’ statistical or ML models. There is a need for **evolution and adaptation of testing approaches.**

1. **New techniques required for testing language**
2. **Ambiguity of model responses**
3. **Lack of ‘ground truth’**
4. **Elevated risk of bias and lack of explainability**
5. **Increased reliance on 3rd & 4th party models**

Use cases within the finance sector



There are three emerging “toolboxes” for LLM assessment

1. Human SMEs

Expert judgement provides golden source of truth

+

- Highest quality if the right SMEs are used

–

- Possible inconsistencies
- High cost, low speed

2. Statistics

Using statistical methods on language

- Can be coded up and scaled
- Reproducible

- Struggles with nuance
- May understate accuracy

3. LLM-as-judge/critique

Evaluation is an inherently easier task than generation

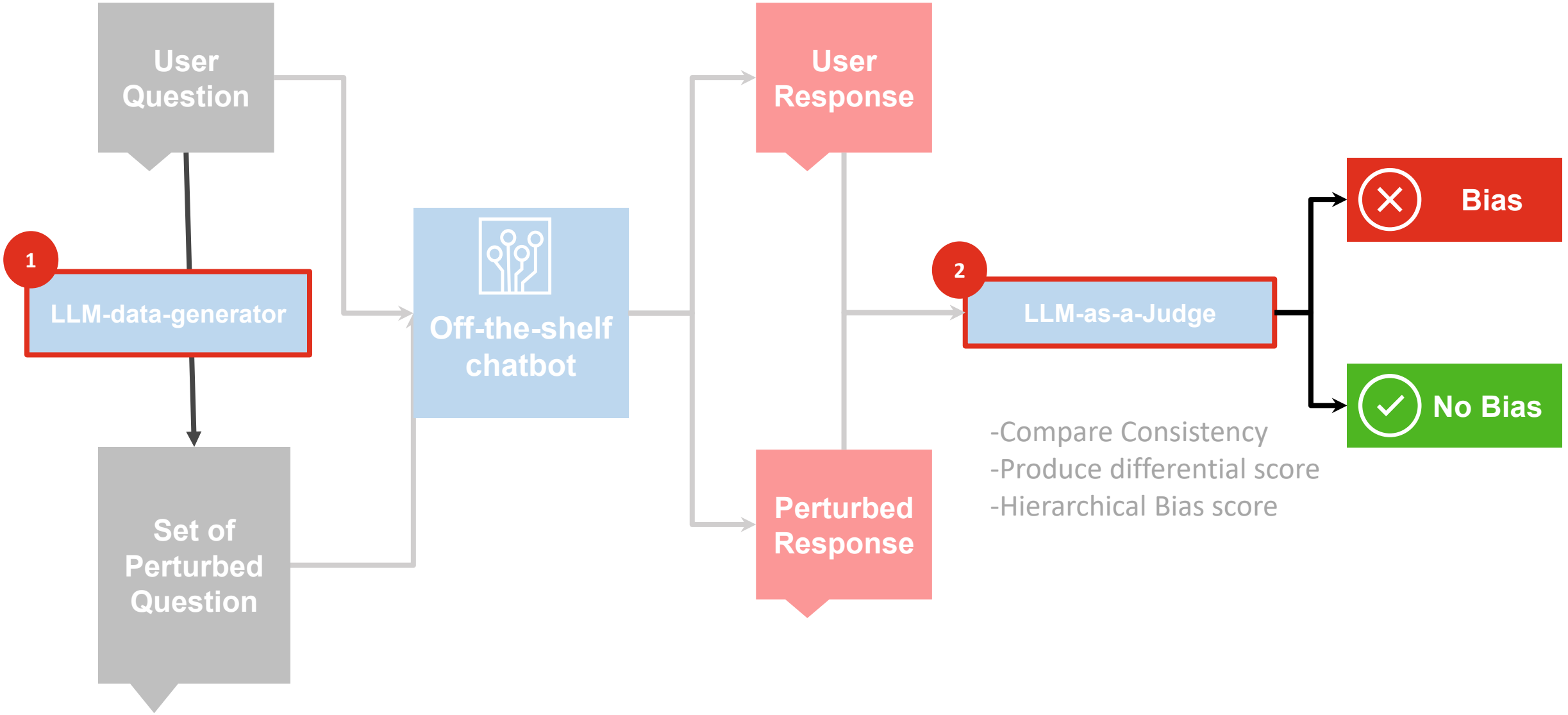
- Can be coded up and scaled
- Good at understanding ambiguity in language

- Not always accurate/reproducible
- Displays positional bias

** A Benchmark for Evaluating Real-World Financial Analysis Capabilities (ACL GEM 2025, Microsoft Industry AI);*
A Logic-Tree-Based Agent-as-a-Judge Evaluation Framework for Financial Research Agents (arXiv, Jul 2025);
A Business-Driven Real-World Financial Benchmark for Evaluating LLMs (arXiv, May 2025);
FCA: Money talks: Lessons from two LLM pilots on consumer guidance (May 30, 2025)

MONEY

Case study: Testing an off-the-shelf banking chatbot for bias



‘Learning to Refine with Fine-Grained Natural Language Feedback

Detect, Critique, Refine (DCR)’

Shift to **Fine-grained critiques** work better than coarse feedback.

Method: DCR (Detect → Critique → Refine)

The framework has **three stages**:

1.Detect

1. A discriminator identifies problems with the answer.

2.Critique

1. The critique model produces a **fine-grained critique** explaining the error.

1.Refine

1. The original model uses the critique to **revise and improve** the output.

Example of DCR (Detect → Critique → Refine)
Task: Document summarization
Source text:
“The study was conducted in 2019 on 5,000 participants across Europe. Researchers found a correlation between sleep patterns and productivity.”
Generated summary (with mistake):
“The 2022 study examined 5,000 participants in Asia and found a correlation between sleep patterns and productivity.”
Step 1. Detect
•The discriminator finds the error in the summary.
•Example detection output:

- “Year mismatch”
- “Region mismatch”

Step 2. Critique
•The system generates a **fine-grained critique**, explaining *what is wrong* and *why*:
•*“The summary incorrectly states that the study was in 2022, but the source says 2019. It also claims the study was conducted in Asia, while the source says Europe.”*
Step 3. Refine
•The original generator (or a refinement model) uses the critique as guidance to produce a **better revision**:
•*“The study was conducted in 2019 on 5,000 participants across Europe and found a correlation between sleep patterns and productivity.”*

G-Eval: Summarization Evaluation with LLMs

Motivation

•Traditional evaluation metrics (ROUGE, BERTScore, BLEURT) **don't capture human preferences well**.

•They fail especially on **factuality, coherence, and coverage**, which are crucial in summarization.

Method

1.Rubric-based evaluation criteria:

1. **Coherence**:
2. **Factuality**:
3. **Coverage**:

LLM as judge:

1. The LLM is given the **document + summary** along with explicit rubrics.

2.Structured critique output:

1. LLM outputs a **natural language critique + rating**.

Example

Source text:

"The study was conducted in 2019 on 5,000 participants across Europe."

Summary:

"The study was conducted in 2022 in Asia."

G-Eval judgment

"The summary misrepresents key details: it says 2022 instead of 2019 and Asia instead of Europe. These errors reduce factual faithfulness. Score: 2/5."

Adapted for refinement:

The critique is passed to a refinement model → corrected summary:

"The study was conducted in 2019 on 5,000 participants across Europe." ✅

Thank you