

A photograph of a white lighthouse with a red top on a rocky island, with a sailboat in the water in the foreground, under a sunset sky.

AI-Native Issue-Risk Intelligence: Connecting Issues to Known Risks and Identifying Emerging Risks

Lawrence Dennison-Hall, Data Scientist

Authors: Lawrence Dennison-Hall, Alok Singh, Chris Wells, Ivy Xiong

Affiliations: Roche, Genentech

November 17, 2025 | PHUSE EU Connect 2025

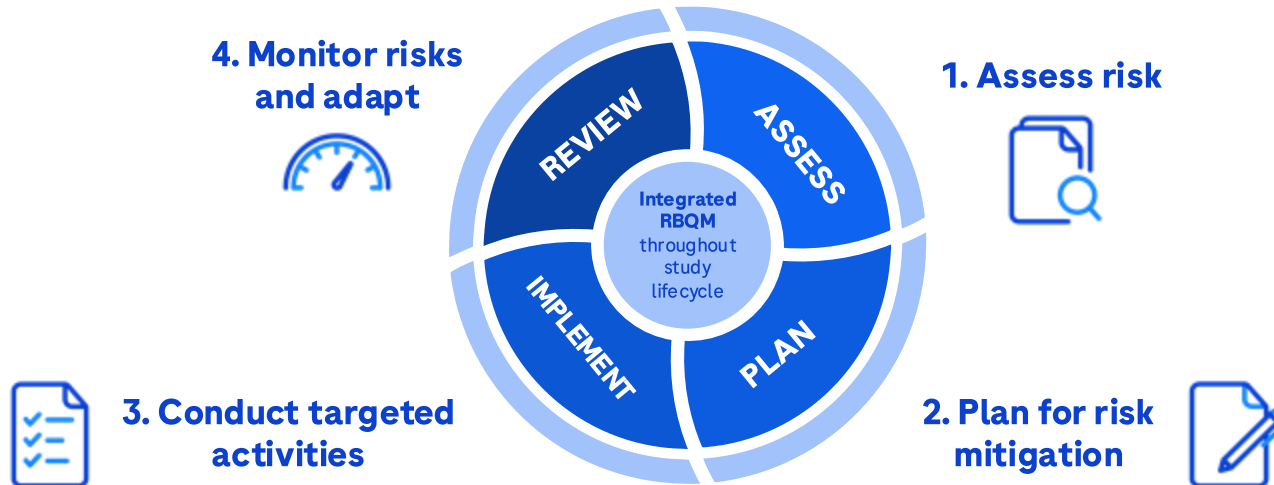
Agenda

- The RBQM Cycle - The Context
- The Problem - Disconnect in the Quality Loop
- Development Approach - How We Worked
- Development Iterations with LLMs - What We Learnt
- SME Review - Results
- Conclusion & Future Work
- Q&A

The RBQM Framework

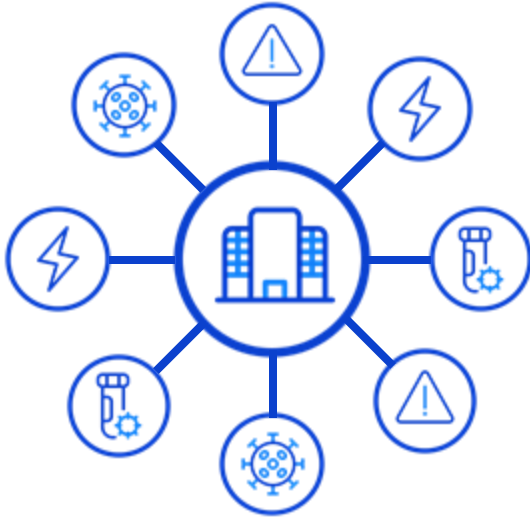
Risk-Based Quality Management (RBQM) is a systematic process for managing risks to ensure the protection of trial participants and the reliability of trial results.

It's a shift from a reactive "*one-size-fits-all*" approach to a **proactive** and **proportionate** framework where oversight and resources are focused on **what truly matters**.



A Disconnect in the Quality Loop

Sponsor risk assessment during study design



Operational issues surfacing at sites during study conduct



**AI Risk
Intelligence**

**Emerging
Risks**

Clinical Trial

**Known
Risks**



Research Objectives

Primary Objective:

To design and build an AI system that programmatically bridges the gap between unstructured issue logs and the formal risk assessment.

Specific Aims:

- To autonomously map operational issues to predefined risks.
- To accurately identify and flag potential emerging risks.

The Value:

- This can enable the next generation of risk review by re-evaluating risks based on the "new knowledge and experience" gained during the trial.
- This new knowledge enables the creation of new risks as well as the improvement of existing risks to ultimately improve quality and patient safety (which can be measured through issue and query reduction).

Risks & Issues



Risks

- A risk is a **potential future event** that **could negatively impact the study**. Risks are identified during the risk assessment phase.
- It's a hypothetical *"if [event] due to [cause] this could lead to [impact]"* statement.
- Example Statement: *"If ineligible patients are randomised due to a misinterpretation of complex eligibility criteria, this could lead to compromised data integrity"*

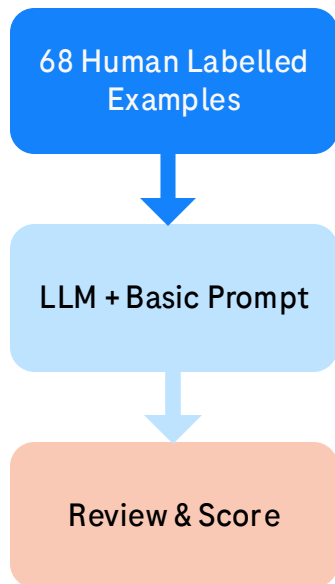


Issues

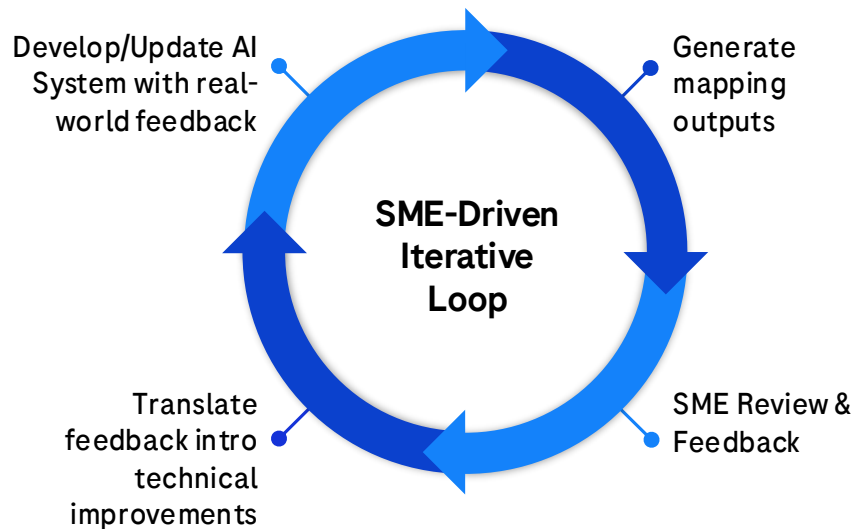
- An issue is an **actual, materialised situation** within trial conduct that **may have negatively impacted the study**. Issues are identified retrospectively. E.g. protocol deviations, AEs, data anomalies, audit findings etc.
- It's a statement describing a specific occurrence
- Examples Statement: *"Participant 205 was randomized on August 29, 2025, despite records showing they are taking a prohibited medication listed in the protocol's exclusion criteria. This was caused by inadequate source review"*

Approach

Phase 1: Concept Validation



Phase 2: Main Development Cycle

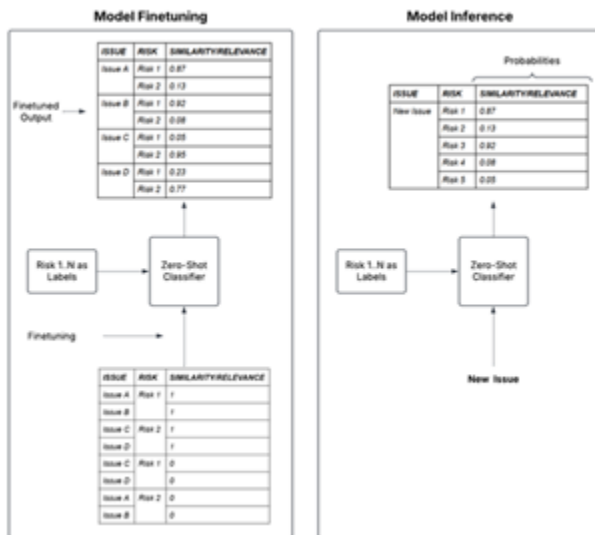


Iteration 1: Problem Framing and Initial Experimentation

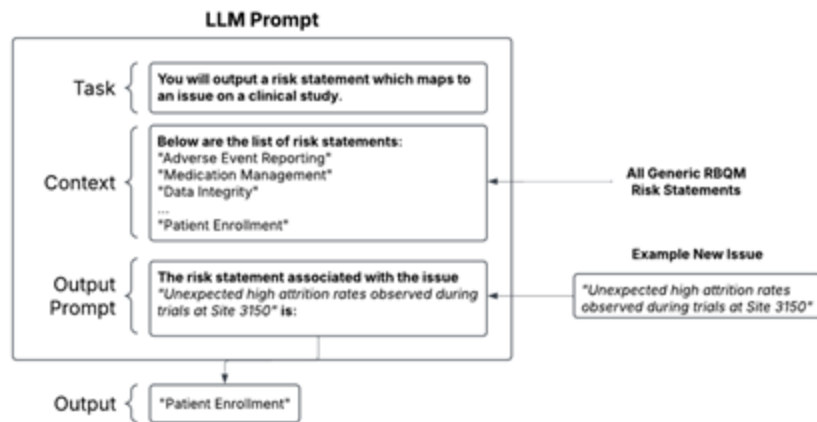
Objective:

- Establish a technical baseline and validate the core concept.
- Determine appropriate the technology

Zero-Shot Small Language (SLM) Approach



Zero-Shot (Prompted) Large Language (LLM) Approach



Iteration 1: SLM vs LLM Results

Finding:

The LLM (GPT-4o mini) demonstrated vastly superior performance (TPR=79.4%) compared to the SLM (BART-Large, TPR=26.5%).

Decision:

An LLM-based architecture was adopted for all subsequent development.

Model	TPR
LLM Baseline (GPT-4o Mini zero-shot prompted)	79.4%
SLM Baseline (BART-Large-MNLI zero-shot)	26.5%

Iteration 1: Zero-Shot LLM Comparison Results

With the LLM approach validated, our next step was to evaluate several state-of-the-art models to select the optimal LLM(s) to enable further development.

Finding:

DeepSeek R1 demonstrated superior performance (TPR=92.6%).

Decision:

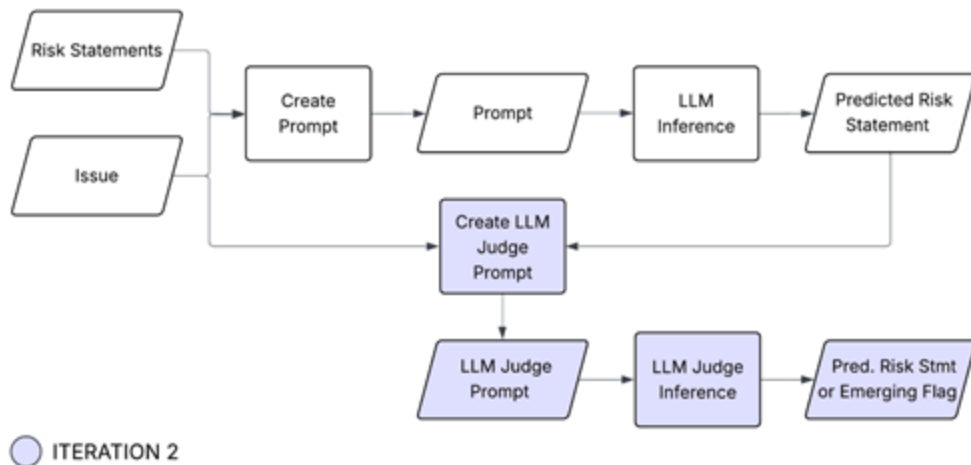
Leverage DeepSeek R1 and/or Claude 3.7 Sonnet for subsequent iterations.

Model	TPR	Correct risk in 3 attempts
GPT-4o mini (OpenAI, 2024)	54 (79.4%)	66 (97.1%)
Claude 3.7 Sonnet (Anthropic, 2025)	58 (85.3%)	68 (100%)
DeepSeek R1 (Deepseek-AI, 2025)	63 (92.6%)	68 (100%)
Llama 3 (Meta, 2024)	53 (77.9%)	66 (97.1%)
Mistral Large (Mistral AI, 2024)	56 (82.4%)	68 (100%)

Iteration 2: Emerging Risk Identification with LLM-as-a-Judge

Objective:

Move from labeled to unlabeled data and begin identifying emerging risks.



Outcome:

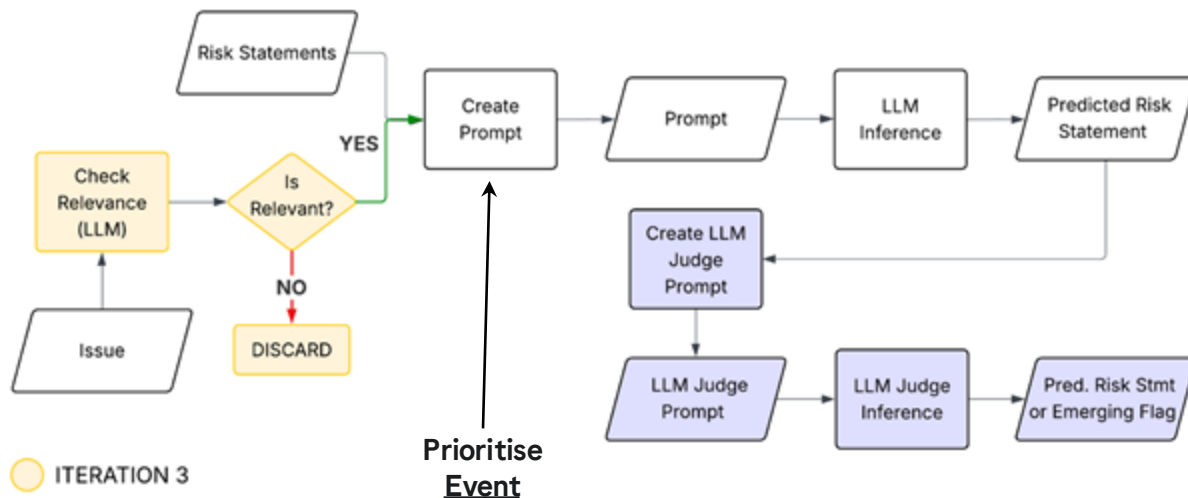
SME review revealed two critical flaws: ❌

- The mapper incorrectly prioritized the 'cause' of an issue over its core 'event'.
- Many entries were not genuine issues ("noise").

Iteration 3: Introducing an Issue Relevance Filter

Objective:

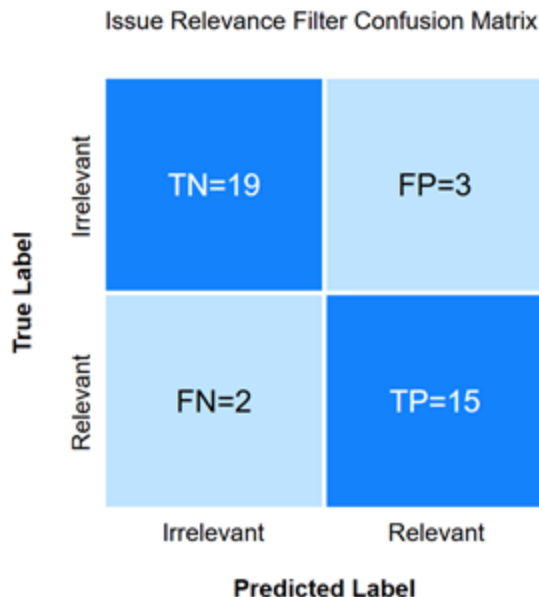
Address the critical flaws of data quality and incorrect mapping logic.



Iteration 3: Issue Relevance Filter Performance

Outcome:

- The filter was highly effective at removing noise 
- the core 'cause' vs. 'event' logic problem persisted 



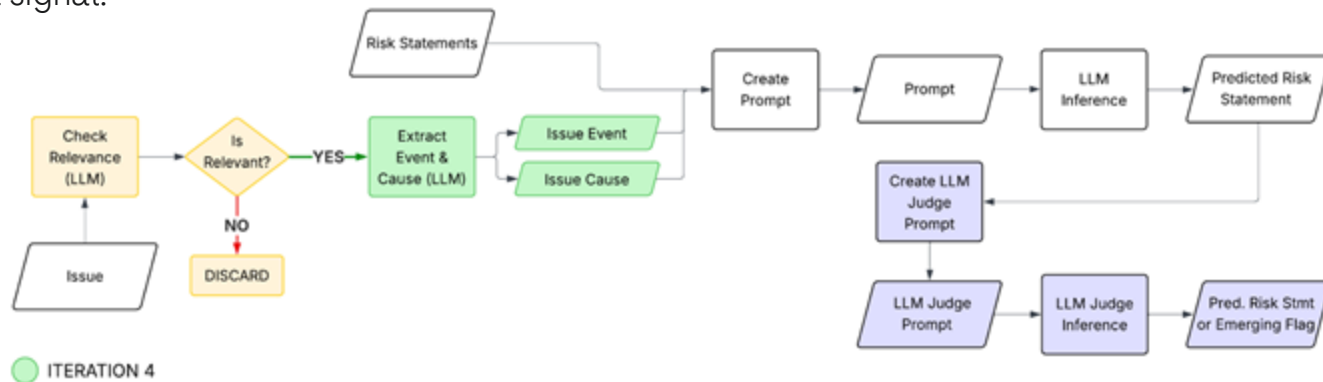
Metric	Score
Accuracy	87.2%
Precision	83.3%
Recall	88.2%
F1 Score	85.7%
TPR	88.2%
TNR	86.4%

Classification performance metrics of the
issue relevance filter (GPT-4.1)

Iteration 4: Input Refinement with Structured Data Extraction

Hypothesis:

The raw, verbose issue text was causing "contextual overload," preventing the model from focusing on the correct signal.



Architectural Change:

Added a pre-processing step to parse the raw issue text and extract two key fields: the core 'event' and its 'cause'.

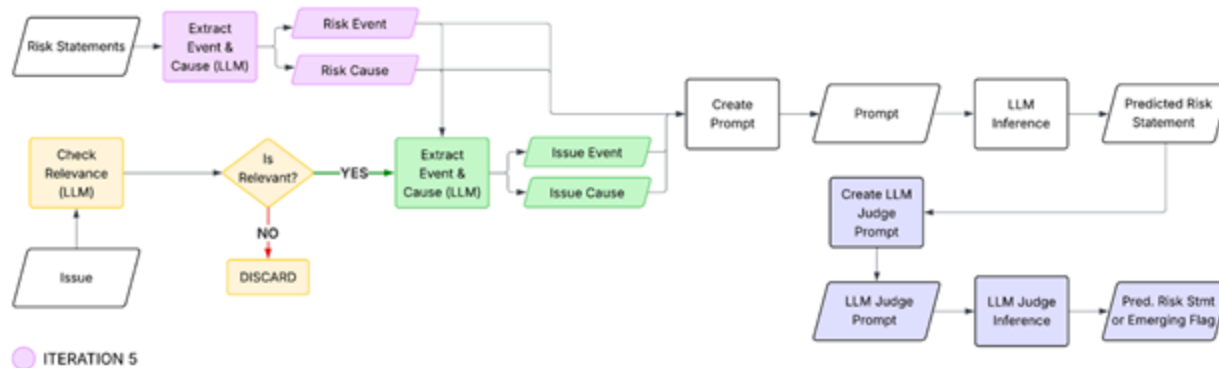
Outcome:

This improved mapping capabilities but surfaced a more profound logical flaw: the system could not distinguish a "materialized instance" of a risk from a mere "trigger" for it.

Iteration 5: Aligning Issue and Risk Data Structures

Hypothesis:

The "trigger vs. instance" problem stemmed from a structural mismatch between the issue and risk data as well as the LLM not fully understanding our definition of issue-risk mapping.



Architectural Change:

Implemented a process to parse the unstructured risk statements into the same event and cause structure as the issues. Improved the risk prediction prompt using an analogy.

Outcome:

A better logical basis was established, but this created two new technical challenges: Context Overload and Task Overload.

Event & Cause Extraction / Data Structuring Example

EXAMPLE RISK STATEMENT

If ineligible patients are randomised due to a misinterpretation of complex eligibility criteria, this could lead to compromised data integrity



**Risk event, cause & impact
extraction LLM**



EVENT: *ineligible patients are randomised*
CAUSE: *misinterpretation of complex eligibility criteria*

**We perform this for every risk in the study
before any mapping takes place**

EXAMPLE ISSUE

Participant 205 was randomized on August 29, 2025, despite records showing they are taking a prohibited medication listed in the protocol's exclusion criteria. This was caused by a misunderstanding of the exclusion criteria.



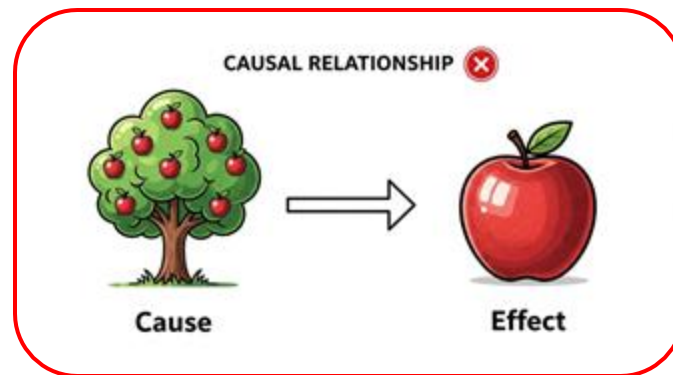
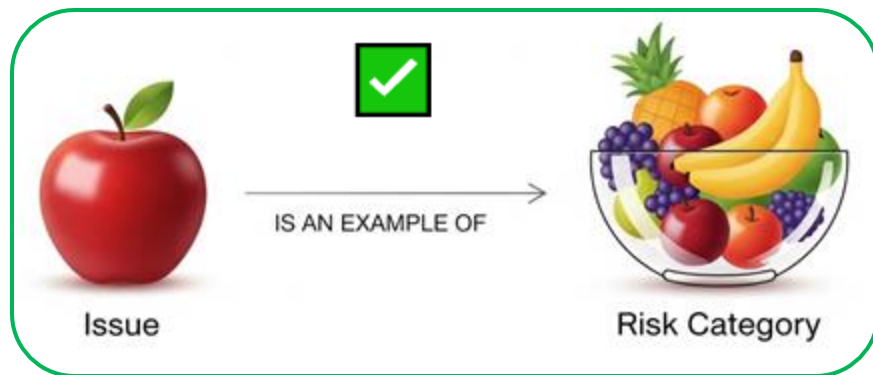
**Issue event & cause
extraction LLM**



EVENT: *Participant was randomized while taking a prohibited medication*
CAUSE: *misunderstanding of exclusion criteria*

**We perform this for every relevant issue
before a mapping can occur**

Issue-Risk Mapping Analogy



✓ Correct Examples:

- Risk Event: "Site Staffing Issues" → Issue Event: "Site coordinator on unexpected long-term leave"
- Risk Event: "Data Entry Errors" → Issue Event: "Incorrect patient weight entered in eCRF"
- Risk Event: "Protocol Deviation" → Issue Event: "Patient visit conducted outside protocol window"

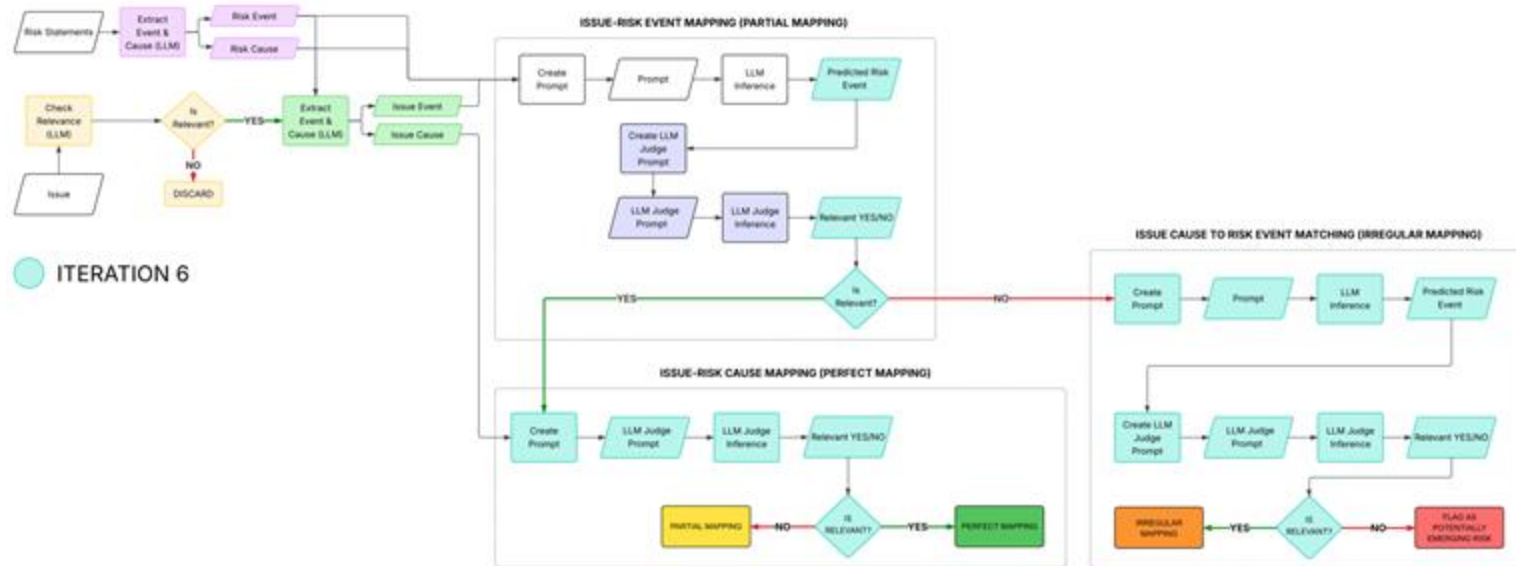
Outcome:

Massively improved the LLMs ability to understand the task; providing LLMs with simple analogies of complex tasks can help to enable their understanding.

Iteration 6: A Hierarchical Classification Approach

Objective:

Deconstruct the overly complex single-step mapping task.



Architectural Change:

Re-architected the analytical core into a multi-step, hierarchical decision process and defined four granular mapping categories (Perfect, Partial, Irregular, Emerging).

Iteration 6: A Hierarchical Classification Approach

Mapping Categories:

- **Perfect Mapping:** The issue's event and its root cause directly align with a risk's event and cause.
- **Partial Mapping:** The issue's event matches a risk event, but the underlying causes differ, suggesting a known risk manifested in an unexpected way.
- **Irregular Mapping:** The issue's root cause maps directly to a known risk event, highlighting an unexpected causal pathway.
- **Emerging Risk:** The issue shows no correlation to any known risk, flagging it for review as a potentially new, uncatalogued risk.

Outcome:

This simplified the task at each stage, but SME review identified a new problem: the "LLM-as-a-judge" explanations were occasionally hallucinating context. It also identified that the LLM was being *too generous* with its mappings.

Iteration 7: Implementing RAG & Chain-of-Thought

Objective:

Ground the system's logic to eliminate hallucinations and improve mapping capability.

Architectural Changes:

- Integrated the study protocol into the workflow with RAG, forcing the model to cite evidence for its reasoning.
- "Panel-of-Experts" Prompting: Implemented a Chain-of-Thought (CoT) technique simulating a debate among AI agents to increase the strictness of the mapping logic.

Outcome:

- Each is LLM "context-aware" of the protocol and is less likely to hallucinate
- The CoT prompting forces the LLMs to reason and therefore be stricter when it comes to predicting a risk.

SYNTHETIC EXAMPLE

AGENT A: Let's begin. Our task is to map this Issue Event to the single best Risk Event based strictly on the observable event and the protocol. AGENT B, please state the facts.

AGENT B: The Issue Event is the randomization of Participant 205. The participant was taking a prohibited concomitant medication at the time of randomization.

AGENT C: On the surface, this maps directly to the risk, 'If ineligible patients are randomised.... this could lead to compromised data integrity'. The events appear to be a direct categorical match.

AGENT B: I agree, but let's verify the specifics. As I recall, the protocol's Appendix 4, Section 8.2, confirms that the specific medication is on the prohibited list, which is a critical exclusion criterion. This validates the patient's ineligibility.

AGENT C: Correct. Therefore, the issue event is a verifiable instance of the general risk category "ineligible patients are randomised." The classification holds.

AGENT A: The risk also mentions a link to "compromised data integrity." Is that connection defined in the protocol, or is it an inferential leap?

AGENT C: The connection is defined. Section 5.4.1 of the protocol discusses data quality and explicitly states that failure to adhere to eligibility requirements is considered a factor that undermines the scientific validity and integrity of the collected data.

AGENT A: Then the consensus is clear. The Issue Event is a direct categorical instance of the Risk Event. The match is fully supported by the definitions in protocol sections 8.2 and 5.4.1. The mapping is confirmed.

A synthetic example of the Chain-of-Thought "panel of experts" dialogue, where AI agents debate the mapping of an issue to a risk by referencing the study protocol.

Quantitative SME Review of the AI System

Evaluation Framework: The SME Review Process

Goal:

To generate a high-quality, consensus-adjudicated "gold standard" dataset

Method:

A structured review of 120 AI-classified items was conducted by two independent SMEs using a custom Streamlit dashboard.

Process:

SMEs could view the full AI workflow and reasoning before adjudicating the final output as correct or incorrect.



Ground Truth Generation

Analysis:

We measured the Initial Agreement Rate between the two SMEs to assess the objectivity of the classification task.

Finding:

The "Irregular Match" category showed a very low initial agreement of only 40%, indicating it was conceptually ambiguous to human experts.

Decision:

The category was methodologically excluded from the final system. Post-discussion agreement on all other categories was >99%.

Agreement Rate Between the SME Reviewers

Section	Initial Agreement Rate	Post-Discussion Agreement Rate
Issue Relevance	92.5% (agreement=37, n=40)	100% (agreement=40, n=40)
Perfect Match	95% (agreement=19, n=20)	100% (agreement=20, n=20)
Partial Match	85% (agreement=17, n=20)	95% (agreement=19, n=20)
Irregular Match	40% (agreement=8, n=20)	Excluded
Emerging Risk Flag	90% (agreement=18, n=20)	100% (agreement=20, n=20)

Issue-to-Risk Mapping Results

Mapping Category	TPR	
Perfect	95.0% (correct=19, n=20)	
Partial	68.4% (correct=13, n=19*)	
Irregular	SME A: 35.0% (correct=7, n=20)	SME B: 90.0% (correct=18, n=20)
Emerging	90.0% (correct=18, n=20)	

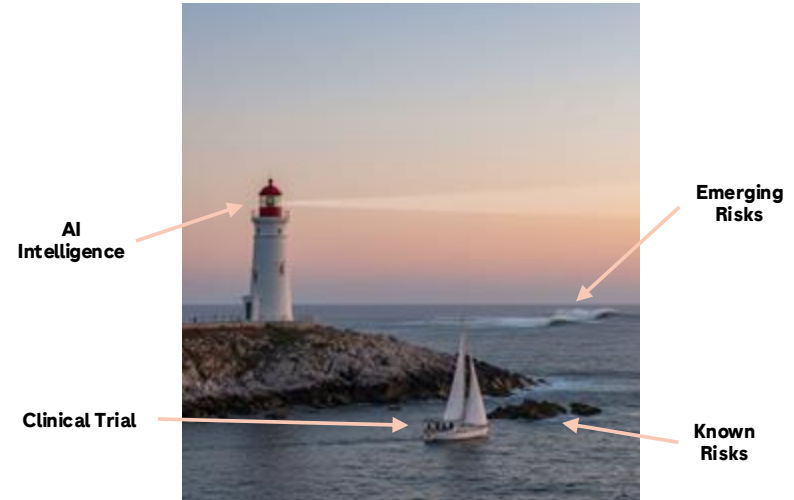
*Excludes example where no consensus was met.

Analysis:

- High Performance (Perfect, Emerging): The system excels at identifying clear, direct semantic/logical connections and clear instances of novelty.
- Moderate-Lower Performance (Partial): The lower TPR reflects the category's inherent cognitive complexity

Conclusion

- Presented a detailed case study of an iterative design process for a complex AI system, demonstrating the necessity of SME co-development.
- Developed a blueprint for transforming RBQM from a static planning exercise into a dynamic learning system.
- Developed a high-performing data quality filter for operational issues.
- Achieved moderate to high performance with our current architecture based on a robust human review of the AI outputs.



Future Work / Next Iteration

Technical Enhancement:

Focus on improving performance on the "Partial Match" category.

System Expansion:

Develop a downstream module to perform automated clustering on emerging risks to identify trends to enable the potential creation of new risks.

GxP Validation & Pilot:

Focus on robust GxP validation to enable a pilot to support the end-to-end RBQM workflow.



Thank you!

Q&A

Your comments and questions are valued and encouraged:
lawrence.dennison-hall@roche.com
chris.wells@roche.com

Doing now what patients need next