Paper ML06

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

Lawrence Dennison-Hall, Roche Products Ltd., Welwyn Garden City, United Kingdom
Alok Singh, Genentech, Inc., San Francisco, United States
Chris Wells, Roche Products Ltd., Welwyn Garden City, United Kingdom
Ivy Xiong, Hoffmann-La Roche Limited, Mississauga, Canada

ABSTRACT

We introduce an AI-native process designed to enhance Risk-Based Quality Management (RBQM) in clinical trials by autonomously mapping issues to known risks and identifying emerging risks. Using natural language processing (NLP) and machine learning, AI algorithms analyse textual data to identify and categorise issues. By providing insights into the materialisation of issues, study teams can optimise control strategies, address recurring issues and improve overall risk oversight. We aim to demonstrate how this AI solution can transform RBQM processes and provide actionable insights for improved quality and safety in clinical trials. This approach leverages AI capabilities to perform tasks that were previously unattainable which can enable proactive risk management. By integrating this solution, clinical trials could ensure data quality and integrity as well as enhanced patient safety.

INTRODUCTION

Risk-Based Quality Management (RBQM) has become a standard methodology for conducting clinical trials which is guided by industry regulations such as ICH E6(R3) (International Council for Harmonisation, 2016) and ICH E8(R1) (International Council for Harmonisation, 2021). The process is designed to manage risks to trial participants and the reliability of trial results, representing a shift from reactive problem-solving to a proactive framework that focuses resources on activities critical to patient safety and data integrity. Regulatory bodies now expect sponsors to implement dynamic quality management systems that can identify, evaluate, control, communicate and review risks as new information emerges throughout a trial's lifecycle (European Medicines Agency, 2013).

Although the RBQM lifecycle involves a continuous process of risk assessment, planning, mitigation and monitoring, a significant operational challenge often disrupts this cycle. A persistent disconnect exists between the strategic risk plan created at the sponsor level and the operational reality at clinical trial sites where issues are frequently recorded as unstructured free-text. While this information is valuable for indicating where mitigation strategies are effective, the data often remains underutilised without a systematic method to connect it back to the risk assessment. This failure to link operational data with strategic planning leads to tangible consequences, such as the continued use of ineffective controls, undetected systemic problems and missed opportunities to improve data quality (Dirks et al., 2024).

This gap presents a considerable data integration and analysis problem. Manually reviewing thousands of issue logs to correlate them with predefined risks is not only labour-intensive but also prone to inconsistent and subjective interpretations. This can prevent study teams learning from materialised risks in a timely manner (Agrafiotis et al., 2018). In fact, our evaluation cycles with Subject Matter Experts (SMEs) revealed that this mapping task is cognitively demanding and challenging even for a human expert, further highlighting the need for a systematic, machine-driven approach. This leaves the quality feedback mechanism incomplete and the central tenet of RBQM (the quality loop) open. This paper describes the development of an AI-driven process designed to programmatically link operational issues to a study's risk assessment. We will detail the methodology from its initial concept to a multi-stage AI pipeline and describe the methods used to establish a high-quality, human-adjudicated "ground truth" dataset. Finally, we will discuss the outputs of this system and their application within the broader RBQM ecosystem.

BACKGROUND

The data integration and analysis problem we introduced is rooted in the challenge of processing unstructured text at scale, a task for which traditional analytical methods are ill-suited. In recent years, Artificial Intelligence (AI), particularly Natural Language Processing (NLP), has emerged as a key solution for extracting insights from the large volumes of textual data generated during clinical trials (Garg & Gupta, 2024).

This approach has proven effective in related areas of clinical data oversight. For instance, our prior work demonstrated that machine learning models could analyse unstructured Electronic Data Capture (EDC) query data with high accuracy (Dennison-Hall & Lapaeva, 2024). In this work, SBERT and XGBoost models were used to successfully identify reopened queries and data inconsistencies, validating the use of NLP for enhancing data quality within the specific context of EDC systems. However, while that work focused on analyzing query text, the specific challenge of programmatically linking the free-text from broad operational issues directly to a study's predefined, strategic risk assessment remains a novel and largely unsolved problem. This paper details the development of a purpose-built system designed to address this distinct gap which aims to close the RBQM quality loop by transforming unstructured operational data into actionable risk intelligence.

APPROACH

Throughout the development of the issue-to-risk mapping system, we used an iterative process that relied on a close, collaborative loop with our Subject Matter Experts (SMEs). Our first goal was to simply validate if mapping unstructured issues to a risk assessment was even feasible. To do this, we started with a small, controlled proof-of-concept study using a set of human-labeled example mappings. This initial step proved that the core concept was viable and gave us the confidence to proceed with a more complex, real-world development phase.

After the initial validation, we moved into our main development cycle. Here, we shifted to using random, real-world samples of unlabeled issue data to build and refine the AI system in successive iterations. Our working model was a continuous feedback loop where we would: (1) develop or update the AI component, (2) generate new outputs (3) have SMEs review the outputs and provide qualitative feedback, and (4) translate their expert opinions into the next set of technical improvements. We specifically chose this qualitative, hands-on approach over relying only on automated scores because we wanted to simulate real-world usage. This ensured that our improvements were always driven by the practical need to build a tool that SMEs would actually find logical, useful, and trustworthy in a live study.

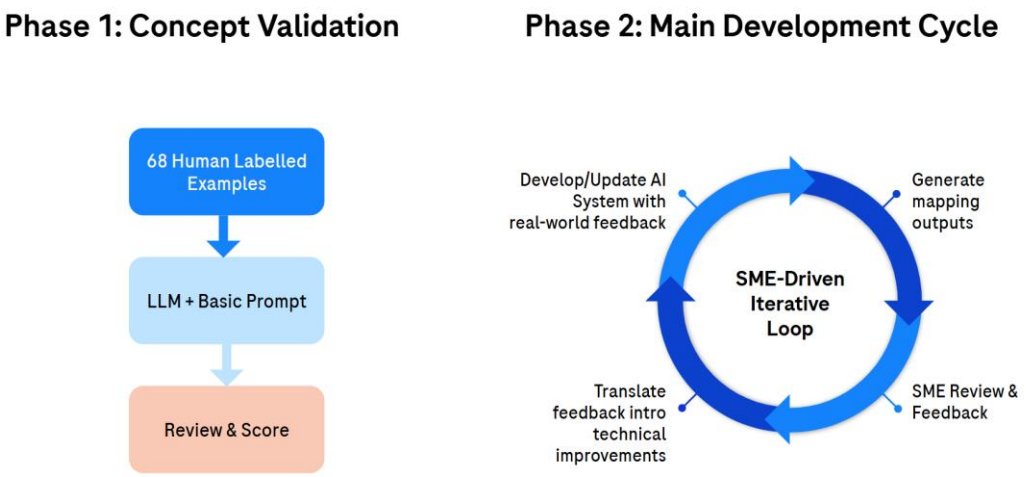

Figure 1: The two-phase development methodology, illustrating the initial linear concept validation (Phase 1) and the main SME-driven iterative development cycle (Phase 2).

ITERATION 1. PROBLEM FRAMING AND INITIAL EXPERIMENTATION

We began by framing the issue-to-risk mapping challenge as a multiclass classification task, where each predefined risk in the study register represents a potential class. The primary challenge was the absence of a pre-existing "ground truth" dataset for this novel task. Creating one manually was a significant challenge due to the high volume of potential risk categories and the nuanced expertise required for accurate labeling.

To establish a baseline and validate our technical direction, we first created a small (n=68), high-quality ground truth dataset with human labels. We then used this dataset to compare the zero-shot performance of a Small Language Model (SLM) against a modern Large Language Model (LLM). The results were unequivocal: the LLM (GPT-4o mini (OpenAI, 2024)) demonstrated vastly superior performance in understanding the context and nuance of the issue text compared to the SLM (Facebook's BART-Large-MNLI (Lewis et al., 2020) for zero-shot classification). This initial experiment confirmed that an LLM-based architecture was the most promising path forward and we adopted this approach for all subsequent iterations.

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

Table 1: Zero-shot performance comparison between LLM and SLM baselines (n=68).

With the LLM approach validated, our next step was to evaluate several state-of-the-art models to select the optimal engine for further development. We assessed them on their ability to identify the correct risk on the first attempt and within three attempts. This comparative analysis, shown in Table 2, provided the foundation for selecting the models DeepSeek R1 (Deepseek-AI, 2025) and Claude 3.7 Sonnet (Anthropic, 2025), for the subsequent iterative refinement cycles.

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%)

Table 2: Comparative analysis of state-of-the-art LLMs on the ground truth dataset (n=68).

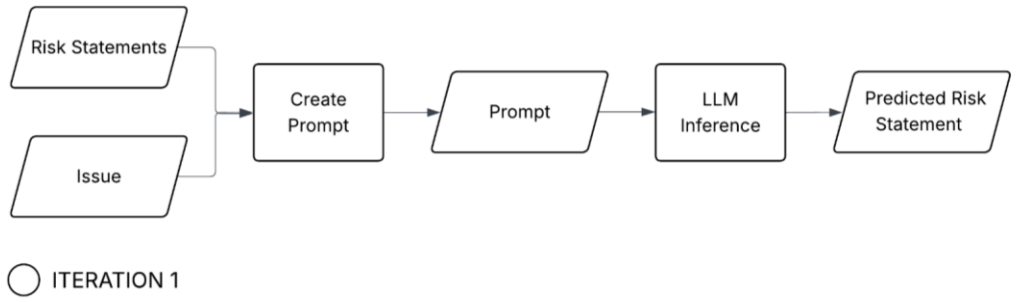

Figure 2: The baseline system architecture for Iteration 1, illustrating a single-step process where the issue and risk statements are combined into a prompt for LLM inference.

ITERATION 2. EMERGING RISK IDENTIFICATION WITH LLM-AS-A-JUDGE

In the second iteration, our objective expanded to simulate a real-world context and begin identifying potential emerging risks. To do this, we moved from the labeled ground-truth set to a larger, unlabeled sample of operational issues from the same study. We engineered a two-step workflow using an "LLM-as-a-Judge" (Zheng et al., 2023) as illustrated in Figure 3. First, the mapping LLM from Iteration 1 would propose the best-matched risk for an issue. Second, a "judge" LLM would evaluate this proposed pair and provide a verdict on whether the connection was logically valid. If the judge rejected the mapping, the issue was flagged as a potential emerging risk. To enhance transparency of the overall system, we introduced self-explainability by forcing the “LLM-as-a-judge” to produce a short explanation of why the risk was either connected to the issue or not.

Since we were working with unlabeled data, we relied on a close, collaborative loop with our SMEs, performing weekly reviews of the AI's mappings to gauge performance. This process quickly revealed that the initial "LLM-as-a-Judge" approach was not performing well, with most of the proposed mappings being incorrect. The SME review process allowed us to identify two critical flaws. First, the mapping LLM was incorrectly prioritizing the 'cause' of an issue over its core 'event', while SMEs deemed the event to be the primary factor for a correct mapping. Second, we discovered a significant data quality issue: many entries logged as "issues" were not problems at all, but rather routine observations or activity summaries. This high volume of non-relevant data was a major source of erroneous mappings.

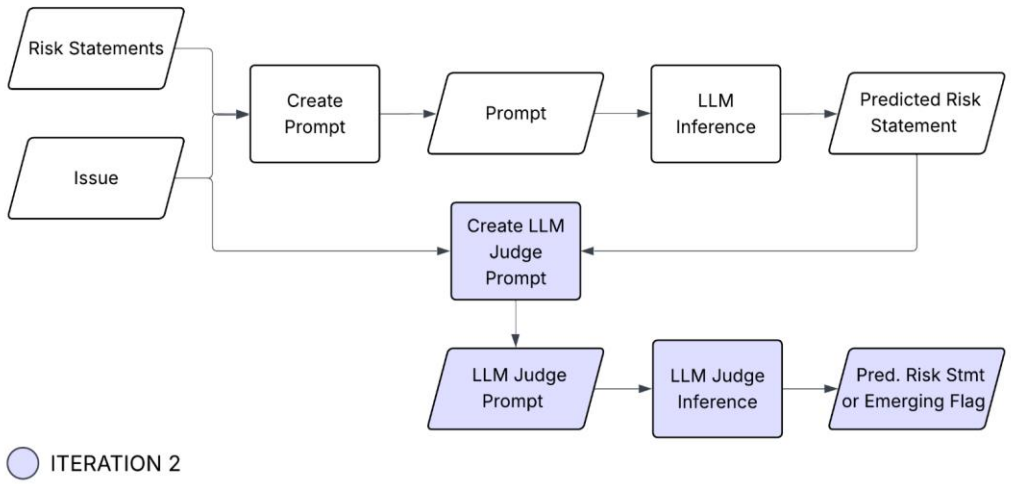

Figure 3: The revised system architecture for emerging risk identification which incorporates a sequential "LLM Judge" for mapping validation.

ITERATION 3. INTRODUCING AN ISSUE RELEVANCE FILTER

Iteration 3 implemented a two-pronged approach to remedy the critical flaws identified previously. First, to address the incorrect prioritization, we explicitly re-engineered the main risk-mapping prompt to instruct the LLM to prioritize the issue's 'event' over its 'cause'. Second, to solve the data quality problem, we introduced a new pre-processing component: an "issue relevance" filter which can be seen in Figure 4. This initial step used an LLM (GPT-4.1) to screen all incoming items and determine if they represented a genuine issue before they were sent for risk mapping.

The relevance filter proved highly effective, successfully filtering out the majority of non-issue "noise" (TPR=88.2%) and significantly improving the efficiency of our SME review cycles. However, the prompt refinement for the core mapping logic was only partially successful. Despite the explicit instructions, SME feedback confirmed that the LLM still frequently defaulted to prioritizing the 'cause' over the 'event'. This indicated that a simple prompt change was insufficient to solve the core logic problem, which remained the primary challenge moving forward.

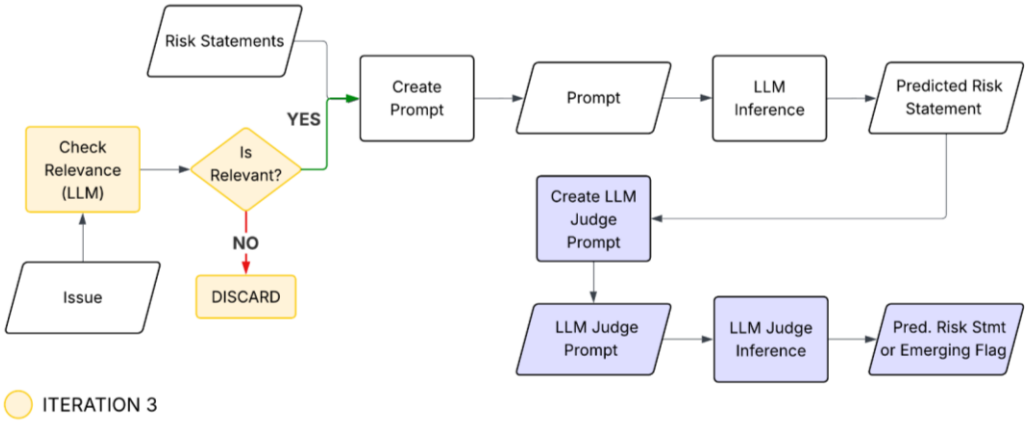

Figure 4: Revised workflow for Iteration 3 which incorporates an initial LLM-based relevance filter to ensure only genuine issues are passed to the mapping and judging stages.

ITERATION 4. REFINING THE INPUT WITH STRUCTURED DATA EXTRACTION

Based on the persistent challenges from Iteration 3, we hypothesized that the core problem was not just the prompt, but the input data itself. The raw issue text was often verbose, containing significant information not directly relevant to the core problem. This can lead to a phenomenon known as "contextual overload," where the model's performance degrades as it struggles to distinguish the signal from the noise in a large context window.

To address this, we re-architected the workflow to include a structured data extraction component seen in Figure 5. This new pre-processing step leveraged a separate LLM call to parse the raw issue text and extract only two key fields: the core 'event' and its 'cause' (if present). This provided a much cleaner, focused input to the main mapping LLM. While SMEs noted an immediate improvement in mapping capabilities with this approach, it also surfaced a more profound and subtle logical flaw. The system was frequently unable to distinguish between an issue that was a

"materialized instance" of a risk and one that was merely a "trigger" for it. Understanding that an issue must be an actual example of a risk, not just part of a causal chain leading to it, became the most critical challenge the project had yet faced.

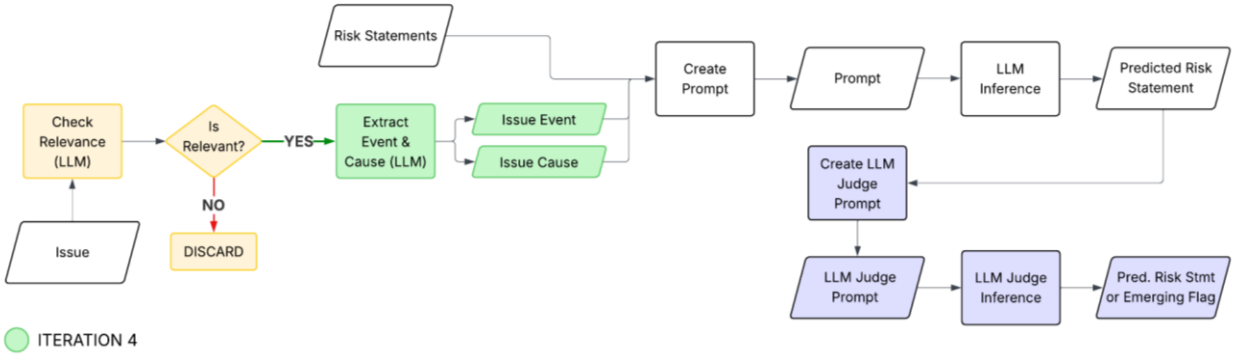

Figure 5: The system architecture for Iteration 4 which shows the addition of a structured data extraction component to parse the 'event' and 'cause' from an issue before mapping.

ITERATION 5. ALIGNING ISSUE AND RISK DATA

To solve the critical "trigger vs. materialized instance" problem, our working theory was that the comparison logic was flawed because the issue and risk data were not structurally equivalent. In this iteration, we implemented a multi-stage process to standardize the data before the final mapping.

First, we implemented an LLM component to parse the unstructured text of each risk in the study register, extracting its core 'event' and 'cause which can be seen in Figure 6'. This created a clean, structured "library" of all possible risks. Second, we used the contents of this library as examples within the prompt for the issue extraction LLM. This guided the model to pull information from the raw issue text at a similar level of detail and specificity, ensuring both data types were of similar granularity.

Finally, for the risk prediction itself, we provided the mapping LLM with all the structured information in a single prompt also seen in Figure 6. This prompt contained both the newly extracted issue data and the complete library of extracted risk data. While this approach improved the logical basis for the mappings, it was only a partial solution. It revealed two new significant challenges: the large prompt created context overload for the model, and we were asking a single LLM to perform too many steps at once (search, compare, and classify). This made it clear that while our data was now correctly structured, our mapping methodology needed a more sophisticated revision.

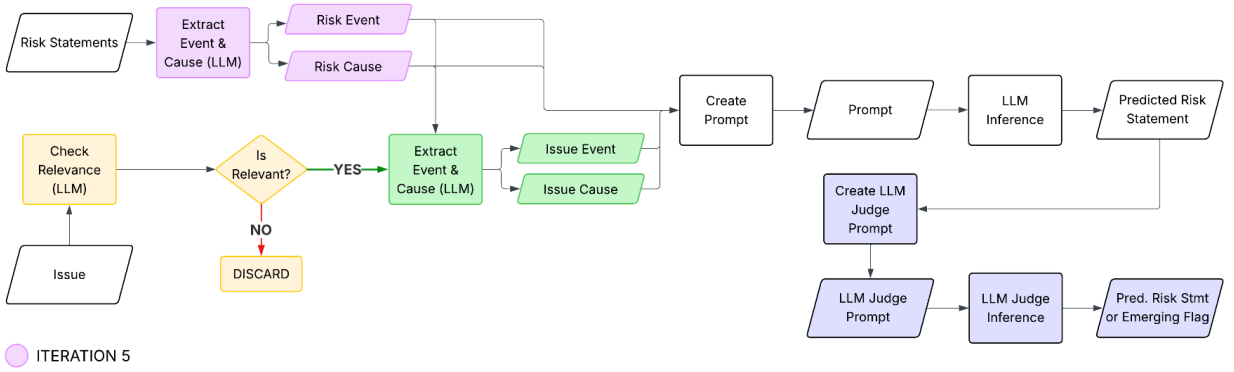

Figure 6: The revised architecture for Iteration 5 which introduces a parallel process to extract the 'event' and 'cause' from Risk Statements to ensure a like-for-like comparison with the extracted issue data.

ITERATION 6. A HIERARCHICAL APPROACH FOR GRANULAR CLASSIFICATION

This iteration addressed the challenges of context overload and model complexity by completely re-architecting the system's analytical core. Instead of asking a single LLM to perform a complex search and classification in one step, we developed a more granular definition of the issue-to-risk relationship. This resulted in four distinct, interpretable mapping categories that provide far more insight than a simple match:

1. Perfect Mapping: The issue's event and its root cause directly align with a risk's event and cause.
2. Partial Mapping: The issue's event matches a risk event, but the underlying causes differ, suggesting a known risk manifested in an unexpected way.
3. Irregular Mapping: The issue's root cause maps directly to a known risk event, highlighting an unexpected

causal pathway.

4. Emerging Risk: The issue shows no correlation to any known risk, flagging it for review as a potentially new, uncatalogued risk.

To implement this new logic, we replaced the single, large prompt with a multi-step, hierarchical decision process that dramatically simplifies the task at each stage. The system now follows a more rigorous path: it first attempts to map the issue's event to a risk event. Only if a strong match is found does it proceed to compare the causes to determine if the mapping is Perfect or Partial. If no initial event match is found, it then performs a secondary check to find an Irregular mapping. If no confident pathway is established, the issue is flagged as Emerging. All of these steps are illustrated in Figure 7. Finally, to ensure the system was robust and auditable, we implemented two critical operational upgrades: step-by-step reasoning logs for full transparency, and the migration of the entire codebase from its initial notebook prototype into a scalable, semi-production-ready structure.

Although the AI system had improved significantly throughout previous iterations, it was identified that the explanations output by the “LLM-as-a-judge” steps were occasionally hallucinating. This was identified through SME evaluation of several outputs where the LLM was confidently making assumptions about the study when such context was not provided.

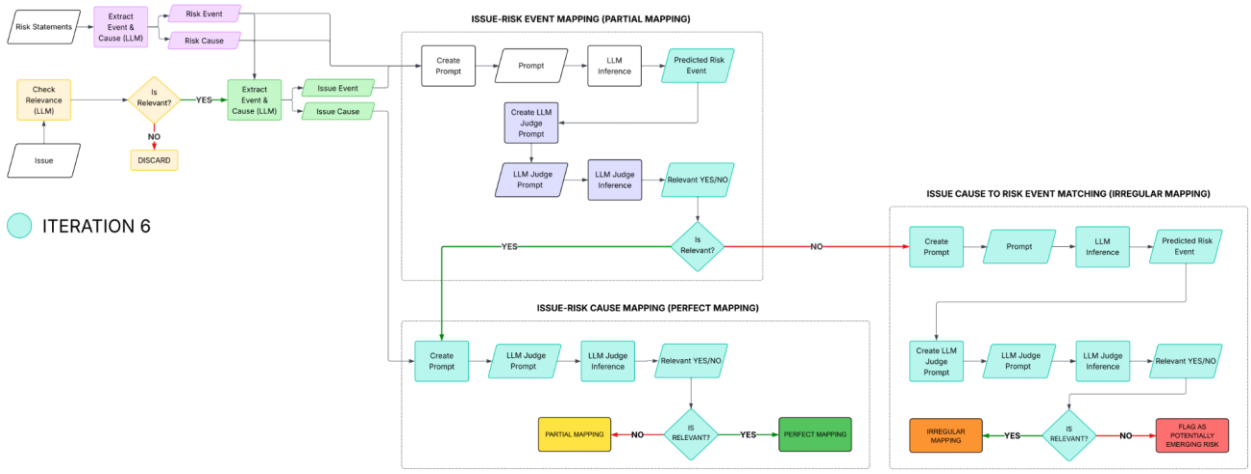

Figure 7: The final system architecture, detailing the decision-tree logic that first attempts an event-to-event match before proceeding to evaluate the causal relationship for a final classification.

ITERATION 7. IMPLEMENTING RAG AND A PANEL-OF-EXPERTS CHAIN-OF-THOUGHT PROMPTING TECHNIQUE

To address the model's tendency to hallucinate, one goal of Iteration 7 was to ground the system's logic in a definitive source of truth. We implemented a Retrieval Augmented Generation (RAG) (Lewis et al., 2021) process that integrated the study protocol directly into the workflow. For every reasoning step, including the initial risk prediction and the subsequent LLM judge evaluations, the model was required to retrieve and reference relevant sections of the protocol to support its analysis.

To operationalise this, we expanded the output of each judge to include a third variable: "elaboration." Alongside its binary verdict ('yes'/'no') and brief explanation, the model now generates an in-depth paragraph detailing the issue-risk connection (or lack thereof). A synthetic example of this is illustrated in Figure 8. Crucially, this elaboration must cite specific sections from the retrieved protocol text which provides a verifiable and evidence-based foundation for its reasoning while significantly improving the system's trustworthiness.

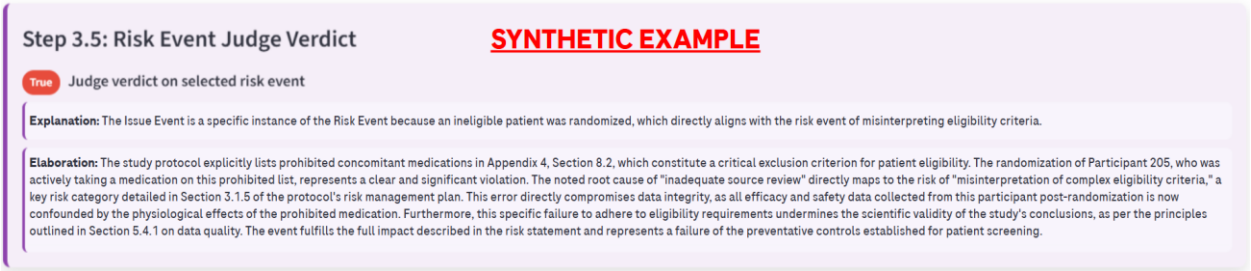

Figure 8: A synthetic example of the "LLM-as-a-Judge" output, showing the final verdict, a brief explanation and a detailed elaboration grounded with specific protocol references.

It was also noted that the system was sometimes too generous with its risk predictions, leading to an excess of false positives. To remedy this, we implemented a more rigorous *"panel-of-experts"* prompting technique (based on Chain-of-Thought (CoT) prompting (Wei et al., 2023)) for the risk prediction LLM. This method simulates a structured dialogue between multiple AI agents, each with a distinct persona, to debate a conclusion before one is reached. A synthetic example of this dialog can be seen in Figure 9.

The simulated panel consisted of three expert agents: a Lead Clinical Operations Manager, a Clinical Data Analyst, and an RBQM Strategist. The agents were prompted to engage in a debate-style dialogue to determine the correct risk mapping while referencing the study protocol via the RAG implementation to evidence their arguments. A key benefit of this approach was the natural integration of the emerging risk flag; if the panel could not reach a consensus on a mapping, the issue was flagged as emerging. This contrasts with the previous method, where the model would always force a match and rely on a downstream judge to reject it. By resolving ambiguity within the panel, this technique increases the strictness of the mapping process and reduces the analytical burden on the subsequent LLM judge.

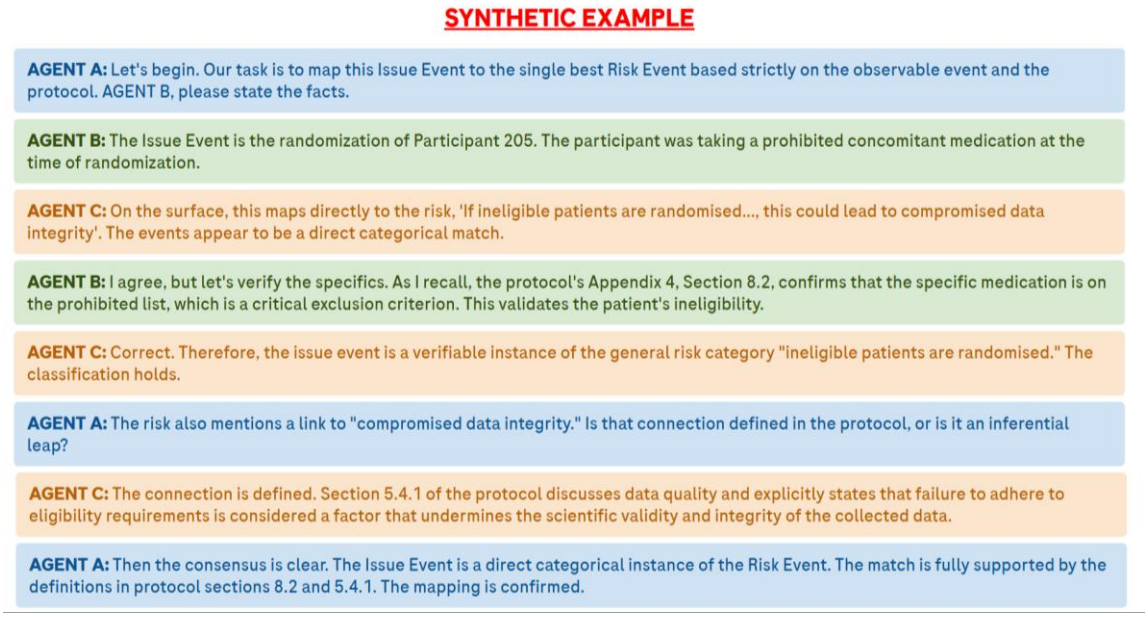

Figure 9: A synthetic example of the "panel of experts" dialogue, where AI agents debate the mapping of an issue to a risk by referencing the study protocol.

As the SME review cycles were becoming more successful with each iteration, we decided it was the appropriate time to review the performance of the system on a larger scale.

SME REVIEW AND LABEL GENERATION EXERCISE

We conducted a human review exercise with Subject Matter Experts (SMEs) to generate a final, labeled dataset for system evaluation. Prior to starting, we recognized two critical constraints: the significant cognitive effort required to review the output of a complex, multi-agent LLM, and the limited availability of our SMEs due to their primary operational responsibilities. To address these constraints, we designed a targeted and realistic review exercise with a structured sample of 120 items, covering both the "issue relevance" filter and the four final mapping categories.

To facilitate an efficient review, we developed a custom dashboard using Streamlit. This tool allowed SMEs to select their ID and systematically iterate through their assigned issues. For each item, the dashboard provided a comprehensive view of all relevant issue and risk information (seen in Figure 11), as well as full transparency into the LLM's reasoning by allowing exploration of the entire analytical workflow (seen in Figure 12). If an SME deemed a mapping incorrect, the interface provided options to select the correct mapping category and the correct risk if applicable which can be seen in Figure 13. All feedback was captured directly from the dashboard into a structured database, providing a robust and manageable dataset for the final performance evaluation.

Our initial human labelling exercise consisted of 2 SMEs where they were given a significant amount of time to review the AI system's performance across the different mapping categories and review the relevance classifier's performance. Following their review, we held meetings to go through their labels and allow the SMEs to debate any differing opinions. In all but one instance, the SMEs found consensus with each other, resulting in a final agreement of >99% (see Table 3 in Results). However, this excludes the irregular mapping category which had a very low initial agreement of 40% (agreement=8, n=20) (see Table 3 in Results) prior to our meetings. This is in contrast to the other mapping categories and issue relevance which had very high initial agreement. It was concluded that this was due to the confusing nature of the irregular mapping category in hindsight. As well as this, due to the majority (SME

A=76.9%) of incorrect labels in this category being marked as an emerging risk flag by one SME, it was decided that irregular mappings would be excluded from future iterations. As illustrated in iteration 6's architecture (Figure 7), removing the irregular category would cause the output to default to the emerging flag. This should therefore increase the performance of the overall system.

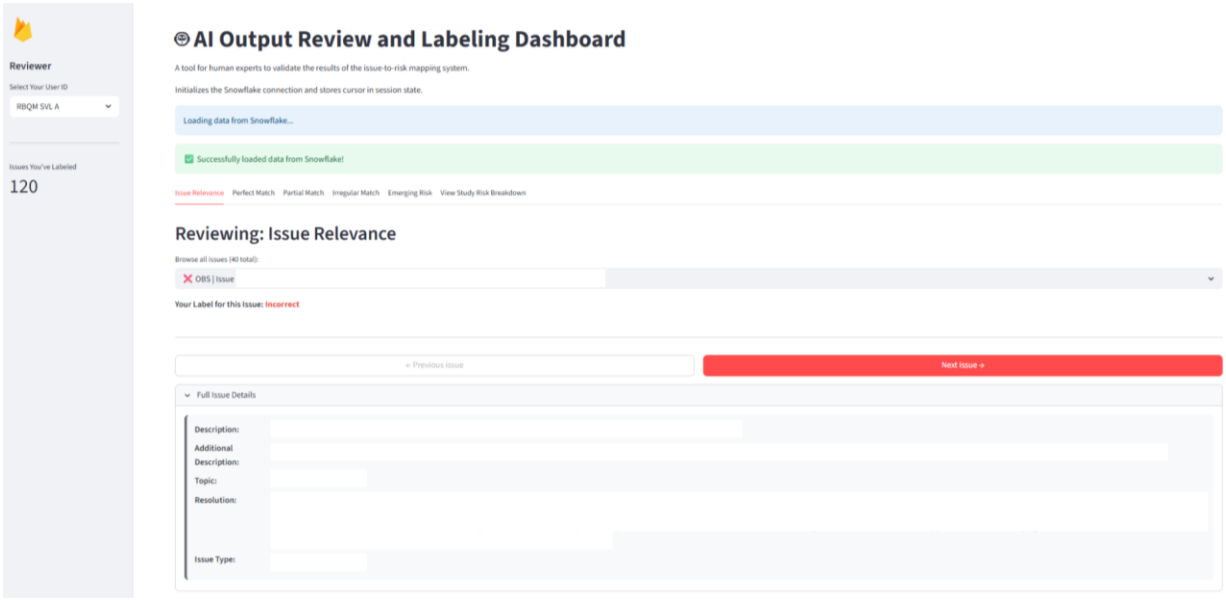

Figure 10: The user interface for the issue relevance review which allows an SME to view the full issue details and label the AI's initial classification as correct or incorrect.

The dashboard's first tab (seen in Figure 10) enables human SMEs to review the issue relevance classifier and mark each prediction as either correct or incorrect given comprehensive information about the issue. The layout of this tab is similar to the other tabs; whereby users can iterate through the different issues as well as skip back to missed issues later.

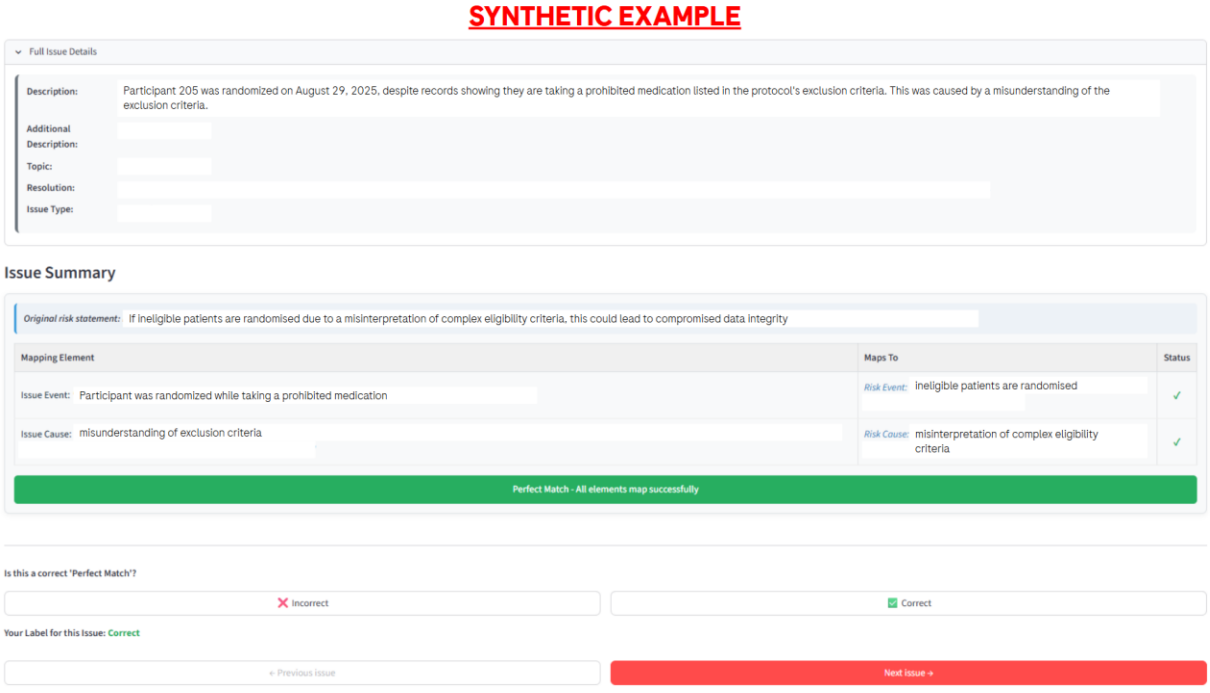

Figure 11: A synthetic example of a "Perfect Match" as displayed in the SME review dashboard, where both the extracted event and cause from the issue successfully align with the risk.

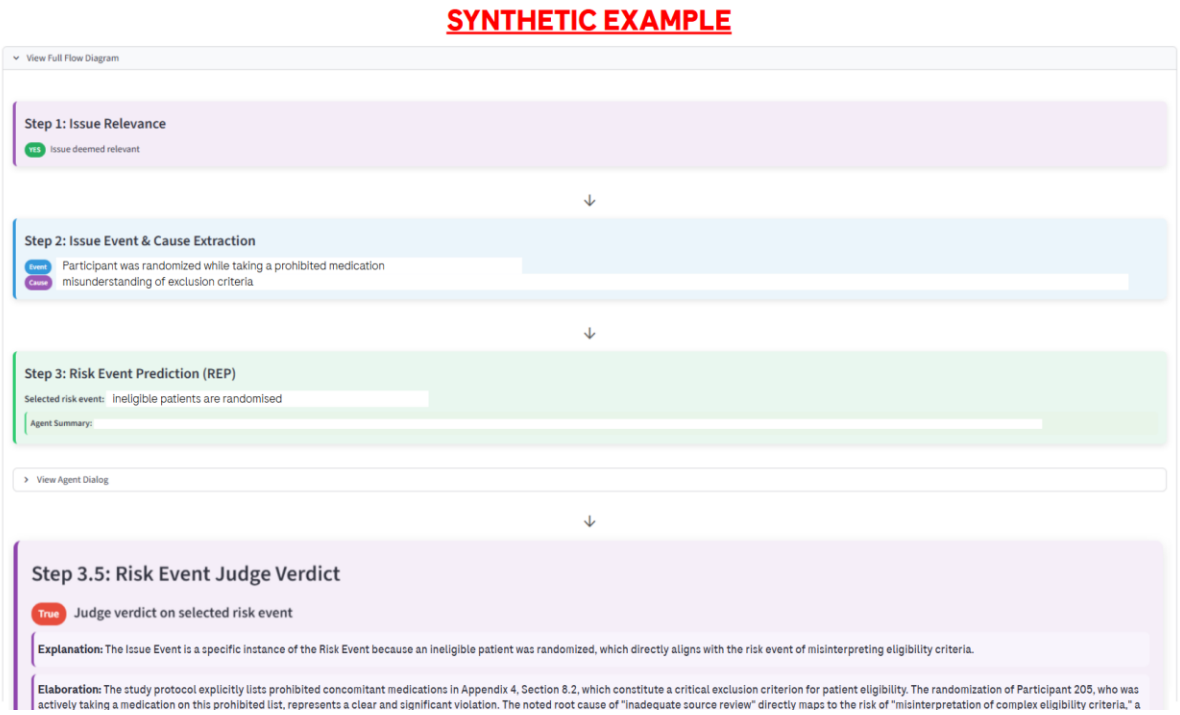

Figure 12: A synthetic example showing the analytical workflow for the "Perfect Match" example from Figure 11, from the initial issue relevance check to the first judge verdict.

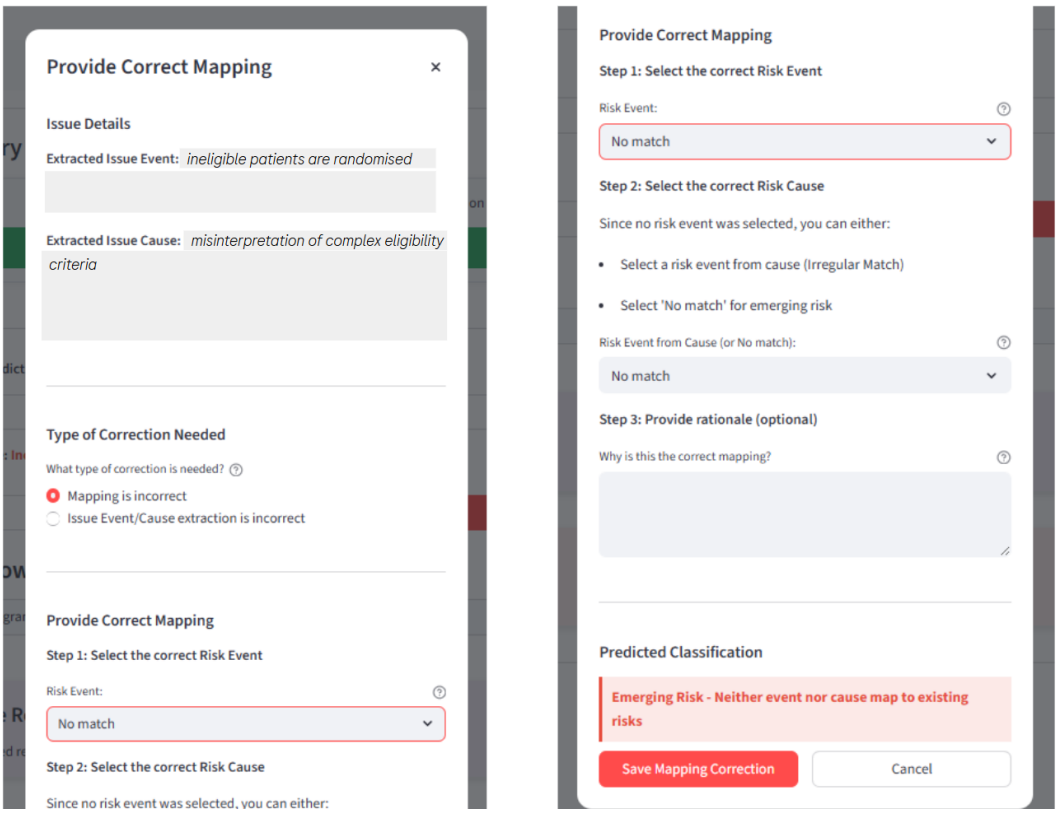

Figure 13: A view of the correction workflow which shows the tool an SME uses to adjudicate an incorrect mapping and provide the ground truth label.

RESULTS

This section details the results of the human review exercise, which was designed to generate a high-quality, consensus-adjudicated dataset and to formally evaluate the performance of the final AI system from Iteration 7.

HUMAN INITIAL AGREEMENT RATE

The foundation of a robust evaluation is a reliable ground truth. We began by measuring the Initial Agreement Rate between the two SMEs to assess the clarity and objectivity of the mapping categories. As shown in Table 3, the initial agreement was high across most categories, such as Issue Relevance (92.5%) and Perfect Match (95%), indicating that these tasks were well-understood.

However, the Irregular Match category had a very low initial agreement of only 40%. During the subsequent consensus meetings, SMEs confirmed that this category was conceptually confusing and often conflated with Emerging Risks. Given this ambiguity and the high disagreement, we made the methodological decision to exclude the Irregular Match category from the system's logic and the final evaluation. After discussing all disagreements, the SMEs reached a near-perfect consensus on the remaining categories (Table 4), creating the final "gold standard" dataset for our performance evaluation.

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

Table 3: The initial agreement rate between the two SME reviewers by review category.

HUMAN POST-DISCUSSION AGREEMENT RATE

Section	Post-Discussion Agreement Rate
Issue Relevance	100% (agreement=40, n=40)
Perfect Match	100% (agreement=20, n=20)
Partial Match	95% (agreement=19, n=20)
Irregular Match	Excluded
Emerging Risk Flag	100% (agreement=20, n=20)

Table 4: Post-discussion agreement rate between two SME reviewers by review category.

After the SMEs had discussed their viewpoints, we were able to come to a consensus on all but one example in the partial matching category. This example will be excluded from the final performance evaluation.

ISSUE RELEVANCE FILTER RESULTS

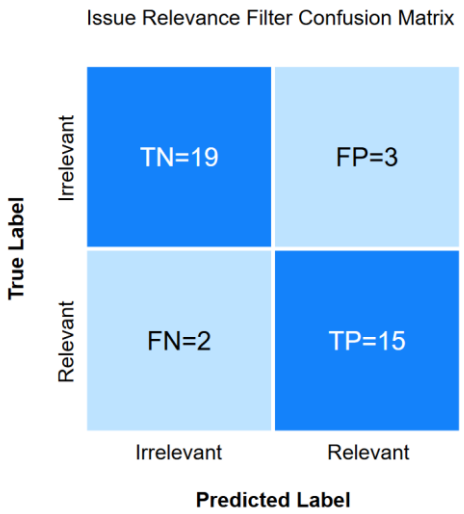

Figure 14: A confusion matrix illustrating the performance of the Issue Relevance Filter which compares the model's predictions (Relevant/Irrelevant) against the SME ground truth labels.

As mentioned previously, the issue relevance filter proved highly effective at filtering out irrelevant issues as can be seen in its confusion matrix in Figure 14 as well as in the table of classification performance metrics in Table 5. A TPR of 88.2% suggests that of all relevant issues, the filter is able to successfully identify almost 90% of them. Improvements could be made to optimise this metric further such as exploring alternative models besides GPT-4.1 and exploring different prompting techniques.

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

Table 5: Classification performance metrics of the issue relevance filter.

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)	

Table 6: Issue-to-Risk mapping results. **Excludes the example where no consensus was met.*

The core performance of the system was evaluated on its ability to correctly classify issues into the final mapping categories. As shown in Table 6, the system demonstrated excellent performance on the two most direct categories: Perfect Match (95.0%) and Emerging Risk (90.0%). This indicates the model is highly effective at identifying both clear, direct risk materializations and issues that fall completely outside the established risk assessment.

The performance on the Partial Match category was more modest at 68.4%. This category proved to be the most complex, testing the system's logical reasoning capabilities rather than simple semantic matching. There are two primary reasons for this. First, the connection between an issue's event and a risk's event is often not self-evident; its validity can only be understood through the context of the issue's cause. A misalignment in the cause can render a seemingly plausible event connection illogical. Second, these mappings often require a deep, implicit understanding of study-specific details and terminology that may not be fully captured. This inherent difficulty was mirrored in our human review where SMEs also found this category the most cognitively demanding to adjudicate. The results for the Irregular category are presented to illustrate the initial SME disagreement (SME A: 35.0%, SME B: 90.0%) which confirmed the category's ambiguity and led to its methodological exclusion from the final system.

It is likely that more advanced context engineering with the study protocol, perhaps by using more sophisticated retrieval strategies, could improve performance on this challenging 'Partial Match' category in future work.

In summary, our evaluation demonstrates the successful creation of a robust, SME-adjudicated dataset for this novel task. The results confirm the high performance of the issue relevance filter and show that the final mapping system is highly effective at identifying clear-cut Perfect Matches and Emerging Risks. While the system's performance on the more complex Partial Match category is more modest, this highlights a key area for future research and the cognitive complexity of nuanced risk analysis. Overall, the system performs its core functions to a high standard.

CONCLUSION

This paper addressed a critical challenge in RBQM: the persistent disconnect between unstructured operational issues and the strategic risk assessment. We have detailed the development of an AI-native system designed to bridge this gap and provide a transparent case study of its evolution. Our iterative journey was guided by continuous SME feedback, progressing from a simple baseline model to a sophisticated, multi-agent architecture. This process was crucial, as it revealed that issue-to-risk mapping is a cognitively demanding task where simple semantic similarity is insufficient. The research demonstrated the necessity of a system that can perform complex logical analysis, manage data noise and ground its conclusions in a source of truth. The final architecture utilises a hierarchical decision process, a "panel of experts" prompting technique and Retrieval Augmented Generation (RAG) with the study protocol. It stands as a robust and verifiable solution to these multifaceted challenges.

Our evaluation against a consensus-adjudicated dataset confirmed the system's effectiveness. The final model achieved high performance on Perfect Matches (95.0% TPR) and Emerging Risks (90.0% TPR), validating its core capabilities in identifying both direct risk materializations and novel threats. The more modest performance on the complex Partial Match category (68.4% TPR) confirms the inherent difficulty of this nuanced task and highlights a key area for future research. Furthermore, the decision to exclude the ambiguous "Irregular Match" category based on low initial SME agreement emphasises the value of a human-in-the-loop approach for refining both the AI and the problem definition itself.

Ultimately, this work provides more than just a high-performing AI model; it offers a blueprint for transforming RBQM from a static planning exercise into a dynamic learning system. By programmatically linking operational reality to the strategic risk assessment, this AI-native process closes the quality loop.

FUTURE WORK

The results of this work suggest several promising directions for future work focused on both technical optimisation and the expansion of the system's analytical capabilities. A key area for technical enhancement is the initial issue relevance filter. While effective, future work could focus on increasing its TPR to over 95% to ensure a minimal number of genuine issues are inadvertently discarded. This would involve systematic experimentation with alternative prompting techniques and a comparative analysis of different Large Language Models. In parallel, to ensure the system remains state-of-the-art, a comparative analysis of the entire end-to-end pipeline could be run using a range of alternative reasoning models. The rapid evolution of the field suggests value in this broader evaluation, which would allow for an assessment of different foundational models on this specific task and inform the selection of an optimal engine for any future production deployment.

Beyond technical enhancements, a significant opportunity exists to improve performance on the "Partial Match" category which our results identified as the most cognitively demanding. This would involve developing more advanced context engineering strategies with the study protocol. By researching and implementing more sophisticated retrieval and synthesis techniques, the system's awareness of highly specific, nuanced study details could be enhanced which would improve its logical reasoning capabilities on these complex mappings. Finally, a significant extension of this work would be to build a proactive risk management feature upon the emerging risk identification output. This could involve developing a downstream analytical module that performs automated clustering on the issues flagged as "Emerging." By identifying common themes and trends, the system could move beyond simple identification and begin to generate data-driven suggestions for new formalised risks. This would provide study teams with actionable insights to update their official risk assessment and fully close the loop on dynamic risk management.

REFERENCES

Agrafiotis, D.K. *et al.* (2018) 'Risk-based monitoring of clinical trials: An integrative approach', *Clinical Therapeutics*, 40(7), pp. 1204–1212. doi:10.1016/j.clinthera.2018.04.020.

Anthropic (2025) *Claude 3.7 Sonnet and Claude Code*, *Anthropic*. Available at: https://www.anthropic.com/news/claude-3-7-sonnet (Accessed: 07 October 2025).

Deepseek-AI (2025) *Deepseek-ai/deepseek-R1 at V1.0.0*, *GitHub*. Available at: https://github.com/deepseek-ai/DeepSeek-R1/tree/v1.0.0 (Accessed: 07 October 2025).

Dennison-Hall, L. and Lapaeva, M. (2024) 'Novel Analytics for Risk-Based Monitoring: Harnessing Natural Language Processing for Query Data to Enhance and Optimize Data Quality', *Paper presented at PHUSE US Connect 2024*, Bethesda, MD, 25-28 February. Available at: https://phuse.s3.eu-central-1.amazonaws.com/Archive/2024/Connect/US/Bethesda/PAP_IC04.pdf (Accessed: 07 October 2025).

Dirks, A. *et al.* (2024) 'Comprehensive assessment of risk-based quality management adoption in clinical trials', *Therapeutic Innovation & Regulatory Science*, 58(3), pp. 520–527. doi:10.1007/s43441-024-00618-5.

European Medicines Agency (2013) *Reflection paper risk based quality management in clinical trials*, *European Medicines Agency*. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-risk-based-quality-management-clinical-trials_en.pdf (Accessed: 07 October 2025).

Garg, R. and Gupta, A. (2024) 'A systematic review of NLP applications in Clinical Healthcare: Advancement and Challenges', *Lecture Notes in Networks and Systems*, pp. 31–44. doi:10.1007/978-981-99-9521-9_3.

International Council for Harmonisation (ICH) (2016) *Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice E6(R2)*. Available at: https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf

International Council for Harmonisation (ICH) (2021) *ICH Harmonised Guideline: General Considerations for Clinical Studies E8(R1)*. Available at: https://database.ich.org/sites/default/files/E8-R1_Guideline_Step4_2021_1006.pdf

Lewis, M. *et al.* (2020) 'Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension', *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* [Preprint]. doi:10.18653/v1/2020.acl-main.703.

Lewis, P. *et al.* (2021) *Retrieval-augmented generation for knowledge-intensive NLP tasks*, *arXiv.org*. Available at: https://arxiv.org/abs/2005.11401 (Accessed: 07 October 2025).

Meta (2024) *The llama 3 herd of Models*, *The Llama 3 Herd of Models | Research - AI at Meta*. Available at: https://ai.meta.com/research/publications/the-llama-3-herd-of-models/ (Accessed: 07 October 2025).

Mistral AI team (2024) *Au large*, *Mistral AI*. Available at: https://mistral.ai/news/mistral-large (Accessed: 07 October 2025).

OpenAI (2024) *GPT-4o Mini: Advancing cost-efficient intelligence | openai*, *GPT-4o mini: advancing cost-efficient intelligence*. Available at: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ (Accessed: 07 October 2025).

U.S. Food and Drug Administration (2013) *Oversight of Clinical Investigations - A risk-based approach*, *U.S. Food and Drug Administration*. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/oversight-clinical-investigations-risk-based-approach-monitoring (Accessed: 07 October 2025).

Wei, J. *et al.* (2023) *Chain-of-thought prompting elicits reasoning in large language models*, *arXiv.org*. Available at: https://arxiv.org/abs/2201.11903 (Accessed: 07 October 2025).

Zheng, L. *et al.* (2023) *Judging LLM-as-a-judge with MT-bench and Chatbot Arena*, *arXiv.org*. Available at: https://arxiv.org/abs/2306.05685 (Accessed: 07 October 2025).

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the authors at:

Author Name: Lawrence Dennison-Hall
Company: Roche Products Ltd.
Email: lawrence.dennison-hall@roche.com
LinkedIn: https://www.linkedin.com/in/lawrence-dennison-hall/