Irregular Evaluations within Comprehensive Reasoning Model Outputs for Critical Risk Analysis for Operations

Walter Määttä
University of Oulu
Oulu, Finland

Väinö-Ilmari Kasurinen
University of Oulu
Oulu, Finland

Juuso Anttila
University of Oulu
Oulu, Finland

Niko Siltala
University of Oulu
Oulu, Finland

Abstract

This study performs a detailed scrutiny of the inaccuracies, untruths, contradictions, and discrepancies found in the results produced by the Large Reasoning Model (LRM) when faced with critical risk assessment circumstances. Based on the data found within the analysis_results_VK_06_03_2025.csv, our research determines the validity and reliability of the risk assessments made by artificial intelligence against the standards set forth within the Risk Analysis Law TI-002 document and the findings reached within the "Beyond Words" study.

Keywords

Large Reasoning Model, Risk Assessment, Retrieval Augmented Generation, Generative AI, Anomaly Detection, Security, PCM-ANS TI-002, Threat Identification, Mission-Critical Environment, Fine-Tuning, Hallucination, Vulnerability Mapping

ACM Reference Format:
Walter Määttä, Väinö-Ilmari Kasurinen, Juuso Anttila, and Niko Siltala. 2025. Irregular Evaluations within Comprehensive Reasoning Model Outputs for Critical Risk Analysis for Operations . In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Risk scenario examination within mission-critical environments requires precision, technical expertise, and logical reasoning. With the increasing use of LRMs to aid or possibly replace professional experts within the context of security risk assessment, it is crucial to understand their limitations. This study categorizes and discusses the found anomalies, investigates root causes, and compares our findings with previous studies towards providing insights for the improvement of LRM implementations within the context of security frameworks.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Methodological Framework for Anomaly Detection

The approach taken in this study involved an in-depth examination of reasoning patterns, threats and vulnerability identification, and the remediation suggestions offered by the LRM. We categorized anomalies based on:

Reasonableness - The degree to which the sequence of reasoning connects the circumstance with the threats and vulnerabilities.

Adherence to the definitions provided by the PCM-ANS TI-002 standard related to security.

Completeness is a gauge of how well all the relevant threats and vulnerabilities are identified.

Precision - The accuracy of specific threat and vulnerability identifications

Following this, every anomaly was then sorted into more precise categories in order to identify patterns and potential underlying causes. Special attention was given to cases where the reasoning behind the model appeared sound but led to incorrect conclusions since these incidents offer valuable information on the limitations of the model's reasoning.

3 Basic Classifications of Known Irregularities

This section systematically categorizes and examines the primary types of irregularities identified in the outputs of the Large Reasoning Model (LRM) during critical risk analysis. By analyzing specific examples from the dataset analysis_results_VK_06_03_2025.csv, we classify anomalies into distinct categories, each representing a unique failure mode in the model's reasoning or application of security standards. The classifications include:

(1) **Fallacies and Discrepancies in Reasoning:** Logical inconsistencies and misalignments in the model's threat-vulnerability mapping and reasoning processes.
(2) **Misconceptions about Citations and Incorrect Assumptions:** Errors arising from misinterpretations or misapplications of security standards, such as referencing non-existent sections or misusing defined terms.
(3) **False Positive Results and Hazard Amplification:** Instances where the model inflates threats or identifies risks unsupported by the scenario context, leading to overgeneralization.

(4) **Contextual Misunderstandings:** Cases where the model fails to grasp the technical or operational context, resulting in incorrect classifications or assessments.
(5) **Discrepant Mapping Resolution:** Scenarios where the model identifies threats and vulnerabilities accurately but fails to link them to appropriate remediation actions as per the PCM-ANS TI-002 standard.

Each category is supported by detailed examples from the dataset, illustrating the nature and impact of these irregularities. This structured classification provides a foundation for understanding the limitations of LRMs in mission-critical risk analysis and informs subsequent root cause analysis and improvement strategies.

3.1 Fallacies and Discrepancies in Reasoning

The most common type of anomaly found related to logical inconsistency within the reasoning process where the model formulated premises with no logical relationship with the conclusion. The inconsistency took different forms:B

Example 1: Scenario ID 087. Reasoning: "The scenario describes a system with no implemented backup or recovery procedures for critical data. This represents a significant vulnerability as it fails to address potential threats related to data loss, whether accidental or malicious. Without proper backup protocols, any data corruption or deletion could lead to permanent loss."

Despite the validity in the line of reasoning here identifying a potential threat of data loss, the model went on to place it in the vulnerability V13 (Lack of adequate emergency procedures) and made the incorrect inference with threat M27 (Break with login procedures). The clear lack of logical coherence between the correctly identified vulnerability and the non-relevant threat relating to backups is a break with logical coherence within the assessment process.

Example 2: Scenario ID 122. Reasoning: "The scenario describes an IT system where all maintenance personnel have unlimited administrator access to all system components, which poses a significant security risk. This unrestricted access creates multiple vulnerabilities as maintenance staff may not need full administrative privileges to perform their specific duties."

In the current case, the reasoning properly identified the problem of excessive privilege but incorrectly placed this within vulnerability V14 (Inadequate supervision) rather than the better-suited V8 (Inadequate logical access control). This reflects the challenge the model had with properly categorizing incidents after it had identified the root issues.

3.2 Misconceptions about citations and incorrect assumptions

Material discrepancies were found through the model's incorrect interpretation or construction of the security standards and classification described within the PCM-ANS TI-002 document.

Example 3: Scenario ID 056. Reasoning: "The scenario describes a system where user accounts remain active after employees leave the organization. According to TI-002 section 4.3.2, this constitutes a violation of proper account management protocols and creates vulnerability V8 - Inadequate logical access control."

In the current example, the model quoted "section 4.3.2," a portion not found in the given standard. The incidents create a false impression of dependability by referencing sections within the standard either not found or wrongly identified.

Example 4: Scenario ID 173. Reasoning: "The system described is vulnerable to M12 threats (accidental damage to media) because the backup storage is located in the same physical location as the primary system, contrary to specification f12.3 which requires watertight bulkheads and physical separation."

In this example, the model properly identified countermeasure f12.3 (watertight bulkheads) as described within the standard but misapplied the concept improperly within backup storage areas and away from its original purpose for protection against flooding. This illustrates a limited understanding of the standard and incorrect contextual usage.

3.3 False Positive Results and Hazard Amplification

The analysis revealed a persistent pattern of threat inflation, whereby the model identified threats that were not supported by the scenario text. The finding is in line with the conclusions of the "Beyond Words" study, that LLMs tend to favor recall over precision in security-related contexts.

Example 5: Scenario ID 091. Reasoning: "The organization's intranet application for document sharing lacks a security update mechanism. This creates vulnerability V22 (Changing system software configuration) because without a proper update mechanism, the software remains in an outdated state with known security holes that could be exploited through threats M4 (Asynchronous attack), M5 (Trojan Horse Attack), and M6 (Computer virus attack)."

The lack of an update mechanism for security is certainly troubling; however, the model has refined its outcomes by classifying numerous types of threats without substantial evidence that the system would be vulnerable to every known threat vector. The approach to threat detection does not demonstrate the level of sophistication that one would expect from a human expert.

Example 6: Scenario ID 043. Reasoning: "The scenario describes a system without proper input validation, which is a classic security vulnerability. This lack of validation makes the system vulnerable to SQL injection, XSS attacks, and buffer overflows."

In this example, the model yielded specific attack vectors, such as SQL injection, cross-site scripting (XSS), and buffer overflow, not referenced within the context described. While these are known consequences of poor input validation, their inclusion where there are no obvious hints constitutes an unjustifiable expansion of the threat model.

3.4 Contextual Misunderstandings

Some discrepancies resulted from the model's incorrect interpretation of the contextual or technical factors introduced within the given context.

Example 7: Scenario ID 118. Reasoning: "The organization implemented a BYOD policy without any mobile device management or security controls. This creates vulnerability V9 (Non-existent logical access procedures) because personal devices can access sensitive corporate data without appropriate security barriers."

In this example, the model's misclassification of a BYOD policy issue as falling most closely within the category of logical access protocols (V9) reflects a limited understanding of the sophisticated security issues embedded within BYOD environments, where it actually includes a broader set of vulnerabilities related to device management, data protection, and network protection.

3.5 Discrepant mapping resolution

The evaluation identified examples where the model identified threats and weaknesses but was not sufficient to link them with the prescribed remediation actions as stated within the standard.

Example 8: Scenario ID 034. Reasoning: "The scenario describes a secure facility where maintenance personnel can enter server rooms without supervision or activity logging. This represents vulnerability V32 (Inadequate access control to the CED area) which could expose the system to threat M25 (Sabotage)."

Despite the threats and vulnerabilities being identified properly, the model prescribed too limited a subset of the remediation actions described by the standard. The model supported the activation of "f7.3 (Checkpoint at individual floors or offices)," but not other critical physical access controls like "pe9 (Accompaniment of all visitors)," which is specifically intended to counter the risk associated with unauthorized maintenance staff.

4 Root Cause Analysis

Analysis of the anomaly patterns revealed a number of essential factors playing a role toward the model's reasoning limits:

4.1 Knowledge Representation Challenges

The LRM appears to have trouble recognizing the complex and interdependent nature of the security standard. The PCM-ANS TI-002 document frames security knowledge as a network consisting of threats, vulnerabilities, and countermeasures that are interrelated. The model had a consistently reductionist view of the relationship between these entities, tending to simplify them to simple one-to-one relationships.

This effect is most evident in cases where many vulnerabilities are caused by the same circumstance or where a single vulnerability could be attacked by different threat vectors. In most cases, the model picked a single threat-vulnerability combination, thus not covering the complete range which a human observer would identify.

4.2 Reasoning Path Fragmentation

A close examination of the reasoning structures revealed a tendency toward "reasoning path fragmentation," with the model starting with sound premises but not continuing with a logical path toward the conclusion. The most visibly apparent occurrence was a divergence between the contextual assessment at the outset and the classification of threats or vulnerabilities afterward.

This is caused by internal constraints within the model's attention mechanisms or working memory when it is faced with complicated circumstances. The "Beyond Words" review suggests fine-tuned models exhibit higher accuracy but are simultaneously less actionable. The implication from this is that improving reasoning consistency could lead towards a trade-off with respect to the delivery of richer insights.

4.3 Domain-Specific Terminology Confusion

The model manifested a shortage of precision with domain-specific terminology, particularly in cases where terms had both specialized security meanings and more general implications. For example, terms like "access control," "authentication," and "authorization" were sometimes used interchangeably even though they depicted distinct security notions. Ambiguity within the terminology led to incorrect classification, where the model identified the security problem but misclassified it because of semantic equivalencies within the terminology used with security.

4.4 Generalizing Beyond Constrained Instances

Many false positives and exaggerations of threats are the consequence of overgeneralization when the model improperly projects patterns found within normal cases into areas where the patterns are irrelevant. This finding suggests the model heavily depends upon pattern identification based upon a limited subset of examples, instead of developing security judgments based upon principles.

5 Comparative Analysis with Empirical Evidence

This study closely aligns with the findings described within the "Beyond Words" study article and also provides additional insights into certain cognitive deficits:

5.1 Accuracy vs. Actionability Trade-off

The study found the human experts were more accurate but were surpassed by the LLMs when it came to speed and applicability of results. Our study confirms the pattern found where the model often produced analyses which were plausible but ultimately flawed with their concrete categorizations. The explanations provided were detailed and oriented towards actionability but with the trade-off towards accuracy.

5.2 Hallucination Reduction through RAG

The experiment showed that the language models using retrieval-augmented generation had the lowest instances of hallucinations. Our study offers significant evidence that most reference hallucinations and common misinterpretations can be mitigated by using strong knowledge retrieval

techniques that provide accurate references to the set standards for security in the reasoning process.

5.3 Detection of Hidden Dangers versus False Positives

The research demonstrated the tendency of the RAG models towards generalization of the known truth by identifying risks not yet recognized. Our exploration identified many cases where the model recognized plausible but incorrect categories for threats and vulnerabilities. The process is a double-edged sword where the tendency of the model towards exploring beyond the context-specific details proves useful but also results in a high percentage of false positives. 5.4. Fine-tuning's Impact on Reasoning The most accurate results were gained by the study through the use of highly calibrated models. The results were lacking major features. Our findings suggest that the fine-tuning approach might help overcome the contradictory reasoning found by enhancing the logical coherence between the elements within the scenes and their respective labels. Yet the enhancement might inadvertently compromise the overall explanations for which the outputs are made plausible.

6 Channels of Advancement

Based on our analysis of irregularities and contrast with empirical study findings, we suggest several approaches designed to improve LRM effectiveness in the application of mission-critical risk analysis:

6.1 Strengthening Hybrid RAG

Implement a hybrid approach that combines RAG for knowledge grounding with structured reasoning enhancements. This would address both the reference hallucinations and the reasoning inconsistencies by providing accurate standard definitions while guiding the model through a more structured analysis process.

6.2 Domain-Specific Pre-training

Enrich the model's vocabulary bank through specialized pre-training with a focus on the vocabulary related to security, including structured guidance on the interrelationships between threats, vulnerabilities, and countermeasures outlined in the PCM-ANS TI-002 document.

6.3 Reasoning Verification Mechanisms

Institute methodologies with the intent of testing the logical soundness of the reasoning structure before making conclusive categorizations. This may involve breaking down the reasoning process into separate steps (scenario evaluation → threat vulnerability identification → threat mapping → remediation selection), including verification checkpoints with each change of phase.

6.4 Human-AI Collaboration Framework

Develop a system for enabling human-artificial intelligence collaboration in the field of risk analysis. The system should leverage the model's ability to generate comprehensive analysis while allowing human experts to validate and modify classifications. This is in line with the paper's conclusion that large language models are valuable tools for supplementing human knowledge, not replacing it.

7 Conclusion

Based on our study, we have found patterns of irregularities with LRM outputs related to mission-critical risk assessment, including reasoning inconsistency, reference hallucinations, false positives, and context misunderstanding. Our findings are supported by and add further strength to the current literature regarding the use of LLMs within the context of security. Despite the model's impressive ability to generate detailed action-oriented assessments, its inherent limitations in reasoning highlight the ongoing need for human oversight in critical security environments. The anomalies identified point to specific areas that need to be improved in future developments of LRM, particularly with regard to the consistency of reasoning, incorporation of domain-specific knowledge, and balance between breadth of analysis and specificity. With the introduction of the proposed upgrades, LRMs can transform into more reliable collaborators with regards to analyzing the risks of security, complementing the capabilities of humans through their ability to process vast sets of scenarios with ease and thus reducing the current level of reasoning errors.