Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26610-cd

OVERCOMING THE "GLASS CEILING" OF LLMS FOR AUTOMATED RISK ANALYSIS IN OCCUPATIONAL SAFETY

Gregory Forte

PREVENTEO, France. Email: gregory.forte@preventeo.com

Large language models (LLMs) are increasingly explored for automating expert reasoning tasks, including risk analysis in occupational health and safety (OHS). This study evaluates their reliability and reproducibility in this safety-critical context. Preventeo conducted 600 systematic tests across 25 representative scenarios, comparing the outputs of GPT-4o and Gemini Pro with expert-validated reference grids assessing hazard identification and risk evaluation accuracy. Results show that both models achieve high precision (~0.95), confirming their ability to produce relevant and contextually coherent assessments. However, their recall remains limited (~0.75), leading to an F1-score below industrial acceptance thresholds. Moreover, significant reproducibility issues were observed, arising from the stochastic nature of LLMs and unsignaled updates of proprietary models. These factors create a "glass ceiling" that prevents direct, autonomous deployment of LLMs in domains where reliability, traceability, and compliance are essential. This work provides two main contributions. To our knowledge, this work provides one of the first large-scale quantitative evaluations of LLM-based OHS risk-grid generation against expert-validated reference grids, highlighting the gap between linguistic fluency and domain reliability in safety-critical contexts. Second, it introduces a reproducible evaluation framework combining quantitative metrics (precision, recall, F1-score) and expert validation. Based on these insights, the study advocates for hybrid modular architectures integrating LLMs with knowledge graphs, retrieval-augmented generation (RAG), and confidence scoring to improve robustness, explainability, and compliance with industrial standards.

Keywords: AI, LLM, Risk assessment

1. Introduction

Occupational health and safety risk analysis constitutes a fundamental mechanism for accident prevention, regulatory compliance, and organizational resilience. Across industrial sectors characterized by complex socio-technical systems and high hazard exposure, organizations rely on structured methodologies to identify risks, assess their severity and likelihood, and define appropriate preventive measures.

In France, the Single Document for Occupational Risk Assessment (called in French "Document Unique d'Evaluation des Risques Professionnels (DUERP)) is mandatory (Audiffren et al., 2012), while internationally recognized frameworks such as HACCP and EBIOS structure risk management in food safety and information security. These practices are supported by standards such as ISO 31000 and ISO 45001 (ISO, 2018a; ISO, 2018b).

Despite their methodological maturity, these approaches remain predominantly manual and expert-driven. Producing comprehensive risk analysis grids requires synthesizing heterogeneous data sources, including regulatory texts, operational procedures, technical documentation, field observations, and historical incident reports (Bourreau et al., 2012; Vigneron et al., 2013). This process is resource-intensive and subject to variability depending on expert experience and organizational practices.

The increasing complexity of regulatory frameworks and industrial systems further amplifies these challenges. Organizations must continuously update risk assessments in response to regulatory changes, technological innovation, and evolving operational contexts. Simultaneously, stakeholders increasingly demand traceability, auditability, and

demonstrable risk control, making efficient and reliable risk analysis a strategic requirement.

Artificial intelligence has emerged as a potential lever to address these constraints (Russell, 2021). Advances in large language models have demonstrated strong capabilities in natural language understanding, information synthesis, and scenario generation (Bommasani et al., 2021). Complementary progress in natural language processing enables automated extraction of relevant concepts from textual data, while computer vision techniques support hazard detection in operational environments (Fang et al., 2018).

However, the application of LLMs to structured occupational risk analysis remains insufficiently validated. Existing research highlights promising use cases in accident report analysis and safety monitoring but also raises concerns regarding reliability, contextualization, transparency, and reproducibility (Li, 2023). The probabilistic nature of generative models is particularly problematic in safety-critical domains, where omissions or inconsistencies can have severe consequences.

This study therefore aims to quantitatively evaluate the performance of LLMs for automated OHS risk analysis, identify their technological limits, and explore organizational implications. The central research question is: to what extent can large language models reliably and reproducibly automate occupational risk analysis, and what conditions are required to overcome their current limitations?

2. Research methodology

The research followed an experimental development approach grounded in real operational scenarios. The methodological stance was deliberately empirical: rather than relying on synthetic benchmarks or generic question-answering evaluations, the study aimed to assess LLMs in conditions resembling professional risk assessment practices. The underlying hypothesis was that LLMs could generate structured risk analysis grids by synthesizing information from heterogeneous data sources, provided that model prompting and system architecture were designed to reflect domain constraints and expected deliverables.

A prototype inference engine was developed as a modular architecture integrating multiple technological components. The first component was a domain-specific knowledge graph designed to formalize relationships between job roles, equipment, environments, hazardous situations, and risk families. The motivation for integrating a knowledge graph is that occupational risk analysis requires structured domain knowledge and systematic coverage. While LLMs may generate plausible hazards, they do not inherently enforce completeness across hazard families or ensure consistent mapping from context to risk categories. By encoding expert knowledge in an ontology-like representation, the system aimed to support contextual reasoning and to provide an explicit scaffold for risk generation and validation.

The second component consisted of a natural language processing module used to extract regulatory obligations, hazard-related concepts, and operational information from textual sources such as procedures, job descriptions, safety instructions, and regulatory documents. This module served two roles. First, it reduced the cognitive burden on the generative model by pre-structuring relevant information. Second, it provided a pathway toward grounding and traceability: extracted concepts could be linked to source documents, thereby supporting auditability and helping experts verify whether

the generated risk statements were properly justified.

The third component was a computer vision module aimed at analyzing images and videos from worksites to detect unsafe conditions, improper protective equipment use, and hazardous environments. While the present study focuses primarily on LLM outputs and their evaluation against expert reference grids, the integration of computer vision reflects an important methodological assumption: occupational risk analysis often depends on observations of actual working conditions, and visual signals can complement textual descriptions by providing evidence of deviations between prescribed procedures and real practices.

These components were integrated with the APIs of two proprietary LLMs: GPT-4o and Google Gemini Pro. The choice of two models served both comparative and robustness purposes. If both models demonstrate similar performance patterns, particularly in the precision–recall relationship and reproducibility issues, the results can be interpreted as reflecting limitations of current LLM paradigms rather than idiosyncrasies of a single model.

To evaluate performance, a corpus of twenty-five representative operational scenarios was constructed with occupational safety experts across industrial, construction, and tertiary sectors. The objective was to capture a realistic diversity of contexts and hazard configurations, including different combinations of tasks, equipment, environmental constraints, and organizational arrangements. For each scenario, experts produced reference risk analysis grids serving as gold standards. These expert grids represented the expected output of a professional assessment, including identified hazards and associated preventive measures. Their function was not only to provide a target list but also to reflect domain-specific expectations regarding specificity, actionability, and relevance.

An experimental campaign comprising six hundred systematic tests was conducted. Each test combined a scenario, a structured prompt, and one of the LLMs. Prompting strategies were designed to standardize the form of outputs across models and to reduce prompt-induced variability. Model outputs were evaluated using precision, recall, and F1-score metrics. Precision measured the proportion of AI-generated risks that matched expert expectations, recall measured the proportion of expert-identified risks recovered by the AI, and F1-score captured the balance between relevance and completeness. Processing time was measured to estimate operational efficiency relative to manual methods, recognizing that industrial adoption is partly driven by productivity and scalability constraints.

Table 1. Research design

Element	Value
Scenarios	25
Models	2
Total runs	600
Evaluation metrics	Precision, Recall, F1-score
Additional assessment	Expert review, processing time

Finally, qualitative expert review complemented quantitative evaluation. Experts analyzed outputs to identify contextual errors, overly generic mitigation measures, systematic omissions, and reproducibility issues. This mixed evaluation approach was necessary because quantitative metrics may fail to capture important safety-related dimensions, such as whether a mitigation measure is actionable, context-appropriate, and aligned with regulatory and organizational constraints.

Although the proposed architecture includes knowledge-graph, NLP, and computer-vision components, the evaluation reported in this paper focuses on the LLM-generated risk grids compared against expert reference grids. The hybrid architecture is discussed as a design perspective for future improvement rather than as a fully benchmarked configuration.

3. Results

The quantitative analysis revealed consistent performance patterns across scenarios and models. Both GPT-4o and Gemini Pro achieved high average precision of approximately 0.95, indicating that most AI-generated risk items were relevant and contextually coherent when compared with expert reference grids. In practical terms, the models often used appropriate occupational safety terminology, produced plausible risk statements, and proposed prevention measures broadly aligned with standard safety logic. However, recall remained around 0.75, meaning that approximately one quarter of expert-identified risk items were omitted from the AI outputs. The corresponding F1-score was therefore approximately 0.84, confirming that strong relevance did not translate into exhaustive coverage in this safety-critical setting.

This limitation persisted across operational contexts and prompt configurations, suggesting that the issue is not merely a prompt-engineering defect but a persistent constraint of baseline LLM-based hazard identification. The missed hazards were not uniformly distributed; qualitative analysis suggested that omissions were more frequent in categories requiring specialized knowledge, in organizational and interface-related hazards, and in low-frequency but high-severity scenarios that are not easily inferred from generic context cues.

These results remain problematic for safety-critical applications, where the cost of a missed hazard may outweigh the benefit of multiple correct low-impact items. While the models demonstrated strong relevance, the lack of exhaustiveness constitutes a major barrier to autonomous deployment. In occupational safety, a single omitted hazard can undermine prevention planning and may have legal implications. Therefore, even a relatively high F1-score may not be acceptable if the remaining error profile includes omissions of high-consequence hazards.

Table 2. Main results

Metric	**Main observation**
Precision	~0.95
Recall	~0.75
F1-score	~0.84
Time reduction	~30%
Main limitation	Omission of expert-identified risks
Deployment implication	Human validation remains necessary

From an operational efficiency perspective, the automated system reduced preparation time by approximately 30 percent compared to traditional manual risk analysis processes. This gain reflects LLM strengths in rapidly producing structured text outputs and synthesizing information into a grid-like representation. Nevertheless, the productivity improvement was moderated by the continued necessity of expert validation and correction. Experts frequently needed to review AI-generated grids to identify missing risks, refine generic mitigation measures, and ensure that recommendations were actionable and aligned with organizational constraints and regulatory requirements. In other words, the automation reduced drafting and initial

structuring time, but it did not eliminate the expert workload required for validation and completion.

Qualitative analysis revealed recurring issues beyond omissions. The LLMs often generated mitigation measures that were overly generic, for example recommending training or PPE usage without specifying the relevant operational conditions, the type of equipment, or the practical enforcement mechanisms. Such genericity is problematic in professional risk management because prevention measures must be specific enough to be implemented, monitored, and audited. The models also struggled with the consistent identification of chemical, organizational, and ergonomic hazards, which often require explicit domain knowledge and systematic exploration of exposure pathways.

Reproducibility issues were also evident. Identical prompts sometimes produced different sets of identified risks due to the stochastic nature of generative inference. Additionally, provider-managed updates and alias-based routing can introduce performance drift over time unless model versions are explicitly pinned, which complicates traceability in regulated environments. In a safety-critical context where audit trails matter, these reproducibility constraints are not secondary; they affect whether organizations can justify that their risk assessment process is stable, repeatable, and compliant.

4. Scientific contributions and technological limits

This study provides, to our knowledge, one of the first large-scale quantitative evaluations of LLM-based occupational risk analysis using expert-validated reference grids. By comparing two leading models across 600 tests and 25 operational scenarios, it contributes empirical evidence to a field where many discussions remain based on qualitative demonstrations rather than systematic benchmarking.

The main scientific finding is the persistence of a precision–recall gap: the models generate largely relevant risk items, but they do not achieve the level of completeness required for autonomous use in safety-critical contexts. This result highlights a key distinction between linguistic fluency and domain reliability. More broadly, the study shows that evaluation in this domain cannot rely on plausibility alone, but requires quantitative metrics, expert validation, and reproducibility assessment as conditions for meaningful comparison and responsible deployment.

5. Organizational implications

From an organizational perspective, these findings support a human-in-the-loop deployment model rather than autonomous use. LLMs may accelerate documentation and structuring tasks, but expert validation remains necessary to ensure completeness, justify decisions, and maintain regulatory traceability. In practice, the value of these systems lies less in replacing experts than in supporting them through faster drafting, structured outputs, and decision support.

This also implies governance and training requirements. Organizations must define where AI can be used, which outputs remain provisional, and which validation steps are mandatory before a risk grid is accepted. They must also ensure traceability by documenting prompts, model versions, and supporting evidence, while training professionals to detect omissions, interpret uncertain outputs, and avoid automation bias.

6. Discussion: toward hybrid architectures

The results suggest that improving model scale alone is unlikely to satisfy safety-critical requirements. Occupational risk analysis is not only a language task, but also a systematic exploration problem requiring structured

coverage across hazard families, explicit grounding in evidence, and stable, reproducible outputs. Generative models can produce plausible syntheses, but they do not inherently guarantee exhaustive coverage.

A more realistic path therefore lies in hybrid architectures combining LLMs with structured knowledge sources, retrieval mechanisms, and uncertainty management. Knowledge graphs can support coverage by making hazard families explicit, retrieval-augmented generation can improve grounding and traceability, and confidence signals can help experts focus validation efforts on uncertain or high-risk outputs. In this perspective, LLMs should not be deployed autonomously, but integrated as linguistic and synthesis components within decision-support systems designed for reliability.

References

Audiffren, T., Rallo, J. M., & Guarnieri, F. (2012). The contribution of case law to compliance management in Occupational Health and Safety (OHS) in France. In *PSAM11 & ESREL 2012* (Vol. 2, pp. 1320-1328).

Bommasani, R. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bourreau, L., Audiffren, T., Rallo, J. M., & Guarnieri, F. (2012). The contribution of knowledge bases to compliance assessment: a case study of industrial maintenance in the gas sector. In *PSAM11 & ESREL 2012*.

Fang, W., Cho, Y. K., Zhang, S., & Perez, E. (2018). BIM and cloud-enabled real-time RFID indoor localization for construction management. Automation in Construction, 89, 31–43.

ISO. (2018a). ISO 31000: Risk management – Guidelines. International Organization for Standardization.

ISO. (2018b). ISO 45001: Occupational health and safety management systems – Requirements with guidance for use. International Organization for Standardization.

Russell, S. J., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

Vigneron, J., Guarnieri, F., & Rallo, J. M. (2013). The contribution of ontologies to the creation of knowledge bases for the management of legal compliance in occupational health and safety. In *22nd European Safety and Reliability Conference-ESREL*.

Li, J., & Wu, C. (2023). Deep learning and text mining: Classifying and extracting key information from construction accident narratives. *Applied Sciences*, *13*(19), 10599.