Proceedings of the Estonian Academy of Sciences
2026, **75**, 3, 251–259

https://doi.org/10.3176/proc.2026.3.10

www.eap.ee/proceedings
Estonian Academy Publishers

AI-ASSISTED DECISION-MAKING

RESEARCH ARTICLE

Received 26 January 2026
Accepted 27 April 2026
Available online 13 August 2026

Keywords:
sustainable management, workplace risk assessment, large language models, automation bias, AI–human disagreement, time efficiency, occupational safety

Corresponding author:
Aleksandr Bosler
aleksandr.bosler@taltech.ee

Citation:
Bosler, A., Koppel, T. and Reinhold, K. 2026. The efficiency–verification trade-off in large language model-assisted workplace risk assessment. *Proceedings of the Estonian Academy of Sciences*, **75**(3), 251–259. https://doi.org/10.3176/proc.2026.3.10

© 2026 Authors. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

The efficiency–verification trade-off in large language model-assisted workplace risk assessment

Aleksandr Bosler, Tarmo Koppel and Karin Reinhold

Department of Business Administration, School of Business and Governance, Tallinn University of Technology, Akadeemia tee 3, 12616 Tallinn, Estonia

ABSTRACT

Large language models (LLMs) are increasingly adopted in occupational health and safety (OHS) to accelerate workplace risk assessment, yet their integration raises a dilemma between efficiency and verification. The study examines this trade-off in an LLM-assisted workflow using data from 111 student assignment files from an educational setting, analyzed at the session level, yielding 96–121 analytic sessions depending on measure availability. Efficiency was measured via perceived time savings versus an estimated manual baseline, while the verification-burden proxy was operationalized as an AI–human disagreement in quantitative scoring (probability and impact on 1–5 scales). User acceptance was captured via the utility and satisfaction / intention to reuse Likert scales.

The results show substantial perceived efficiency gains: the median LLM-assisted time was 20 minutes versus the median manual time of 120 minutes (median saving 0.84). The verification-burden proxy remained non-trivial: median disagreement was 0.50, and 26.8% of sessions exhibited disagreement ≥1.0. Acceptance remained high (mean satisfaction 4.06/5). Notably, 40.6% of the sessions showed a possible overreliance pattern, with high satisfaction coexisting with high disagreement. In a robust ordinary least squares (OLS) regression, utility was the strongest predictor of satisfaction, while disagreement reduced satisfaction; time saving was not significant.

Perceived speed gains may thus coexist with hidden assurance costs and possible trust miscalibration, highlighting the need for verification structures and targeted training for sustainable digital safety governance.

1. Introduction

The digital transformation of corporate governance has positioned generative AI as a practical tool for improving organizational resilience. In the context of sustainable management in the digital era, sustainability increasingly depends on maintaining high occupational health and safety (OHS) standards while optimizing limited time and human resources (Falegnami et al. 2024). Workplace risk assessment is central to this challenge, and large language models (LLMs) are now being integrated to support hazard identification and mitigation planning, with the promise of faster and more standardized safety workflows (Aladağ 2023; Nyqvist et al. 2024).

A primary managerial incentive for adopting LLMs is a potentially substantial increase in process efficiency. Recent field experiments have documented substantial productivity gains; for instance, Martin et al. (2025) found that AI-assisted groups completed construction risk evaluations in an average of 36 minutes, compared to 157 minutes for traditional methods. However, these gains may introduce an efficiency–verification trade-off: while LLMs can rapidly generate broad and plausible risk narratives, they may lack site-specific depth and struggle with precise quantitative scoring (e.g., probability and impact), which remains critical for prioritizing controls and allocating resources (Collier et al. 2024; Nyqvist et al. 2024).

A relevant concern for sustainable safety governance is automation bias (AB) – the human tendency to over-rely on automated suggestions even when they contradict contextual evidence (Romeo and Conti 2026). Empirical studies suggest that the "mere knowledge" of a suggestion being AI-generated

can induce overreliance, leading to inferior decision payoffs (Klingbeil et al. 2024). This phenomenon is exacerbated by a "gulf of impatience," where users prioritize the low latency of AI over the time-consuming process of manual verification (Qiao et al. 2025). Although users frequently report high satisfaction with AI integration in safety-related work (Choudhury and Shamszare 2024; Martin et al. 2025), there remains limited evidence on the "assurance cost" of perceived efficiency – namely, whether perceived time savings coincide with systematic AI–human divergence in quantitative risk metrics.

To address this gap, the present study analyzes structured workplace risk assessment data from an educational setting, linking perceived time savings to AI–human disagreement in quantitative risk scoring and user acceptance. The aim of this study is to examine the efficiency–verification trade-off in an LLM-assisted workplace risk assessment across three dimensions: perceived time savings relative to a manual baseline, quantitative disagreement between AI-generated and human-finalized risk scores, and the extent to which high user acceptance may coexist with substantial scoring divergence. This approach allows us to explore a possible overreliance pattern in which satisfaction and intention to reuse remain high despite substantial scoring disagreement.

To investigate this efficiency–verification trade-off, this study pursues three research questions:

1. What is the magnitude of perceived efficiency gains (time savings) when students use LLMs to assist in workplace risk assessment compared to their expected manual baseline?
2. How significant is the quantitative disagreement between initial AI-generated risk scores (probability and impact) and the final scores validated by human users?
3. To what extent do high user satisfaction and intention to reuse persist despite substantial disagreement with the AI's quantitative output, suggesting a possible overreliance pattern?

2. Background

LLMs are increasingly evaluated as safety support tools in workplace risk management, with evidence showing strong performance in early-stage risk identification. For example, Nyqvist et al. (2024) reported that GPT-4 produced more comprehensive hazard identification outputs than human experts (overall score 8.6 vs. 5.7). Similarly, Oral et al. (2026) found high internal consistency in ChatGPT-generated assessments (similarity 0.90–0.99), suggesting stable output structure under repeated prompting. At the same time, technical specificity and numerically grounded scoring remain weak points: in failure mode and effects analysis (FMEA)-style risk assessments, Collier et al. (2024) reported that experts rated AI-generated severity and likelihood scoring as "poor" or "fair" in over half of cases, indicating limitations in decision-critical quantification.

A major driver of adoption is the prospect of process efficiency gains. In construction risk assessment, Martin et al. (2025) documented substantial time reductions, with AI-assisted groups completing evaluations in 36 minutes compared to 157 minutes using traditional methods. However, these gains may introduce an efficiency–verification trade-off. Prior work argues that filtering irrelevant or inaccurate AI content can impose a non-trivial verification burden, including iterative prompting and "debugging" of AI artifacts (Collier et al. 2024; Sun and Kalar 2025). From a governance perspective, such acceleration may also increase the risk of expertise erosion if users reduce reflective deliberation in safety-critical decisions (Martin et al. 2025).

This trade-off may become particularly consequential under AB – the tendency to over-rely on automated suggestions despite contradictory evidence. Parasuraman and Riley (1997) first systematically described how misuse of automation through overreliance leads to monitoring failures and decision errors. Subsequent reviews confirmed that erroneous automated advice increases incorrect decision-making by over 25% compared to unaided conditions (Goddard et al. 2012), and that verification complexity is a central factor determining susceptibility to AB – the harder it is to verify automation correctness, the greater the risk of overreliance (Lyell and Coiera 2017). In AI-assisted workflows, recent evidence confirms that merely knowing a recommendation was AI-generated can induce overreliance, leading to decision payoffs roughly 22% lower than control conditions (Klingbeil et al. 2024; Romeo and Conti 2026). Low-latency outputs may further intensify this effect through a "gulf of impatience," where users accept "fast" results to avoid cognitive effort (Qiao et al. 2025). While satisfaction with LLM integration is often high (Choudhury

and Shamszare 2024; Martin et al. 2025), there remains limited evidence linking perceived time savings to objective AI–human disagreement in quantitative risk metrics. This study addresses this gap by analyzing educational workplace risk assessment data and examining whether a possible overreliance pattern emerges when high acceptance coexists with substantial scoring divergence.

3. Methods

This study examines the efficiency–verification trade-off in an LLM-assisted workplace risk assessment using data from a structured educational workflow. The dataset comprised 111 submitted student assignment files collected across multiple course cohorts. Although all assignments involved AI-assisted workplace risk assessment, some task formats required students to complete more than one tool- or interface-specific workflow. Accordingly, one file could contribute one or more analytic sessions. The 111 submitted assignment files yielded 121 analytic sessions in total. For analysis, the primary unit was an assessment session, defined as one completed risk assessment workflow under a given tool condition. Within a session, time measures were recorded separately for each workplace and then aggregated to session totals. Hazard-level AI–human scoring differences were averaged to obtain session-level disagreement. Likert-based perception measures were recorded once per session. Analytic sample sizes then varied by outcome due to incomplete time, scoring, or questionnaire data. Specifically, 113 sessions had valid time data, 97 had complete AI–human scoring pairs, 120–121 had questionnaire data depending on the scale, 96 had complete data for the overreliance zone analysis, and 91 had complete cases for regression analysis. The study does not aim to establish ground-truth occupational risk levels; instead, it analyzes perceived efficiency gains, AI–human quantitative scoring disagreement as a verification-burden proxy, and user acceptance.

Time measures were self-reported for each workplace within a session: time spent using the LLM-assisted workflow (T_{ai}) and the estimated manual baseline without AI support (T_{manual}). Session totals were computed by summing across workplaces where both values were available. Perceived time saving was operationalized as:

$$SavingPct = \frac{T_{manual} - T_{ai}}{T_{manual}}. \tag{1}$$

For each hazard, both the AI and the participant provided numeric scores on 1–5 scales for probability and impact. Session-level disagreement was computed as the mean absolute divergence across hazards:

$$Disagreement = \frac{|Prob_{ai} - Prob_{human}| + |Impact_{ai} - Impact_{human}|}{2}. \tag{2}$$

Disagreement is interpreted as a verification-burden proxy and does not imply that either AI or human scores represent a ground-truth reference.

The key measures used in this study should be interpreted as analytically useful proxies rather than direct measures of the underlying constructs. First, the disagreement index captures divergence between AI-generated and human-assigned scores but does not establish which party is more accurate; human scores may themselves reflect miscalibration, and high disagreement does not necessarily indicate AI error. Second, the manual-time baseline (T_{manual}) is a subjective self-estimate rather than an observed measurement, meaning that perceived time savings reflect the user's internal valuation of the tool rather than objective efficiency gains. Third, the overreliance zone is defined operationally as the co-occurrence of high disagreement and high satisfaction and should be interpreted as a possible overreliance pattern consistent with AB-related concerns rather than a direct measurement of overreliance or decision degradation.

Participants completed a 17-item Likert questionnaire (1–5) measuring risk-metric realism, competence/accuracy, utility, and satisfaction / intention to reuse. Scale scores were computed as mean item ratings, and internal consistency was assessed using Cronbach's alpha. No questionnaire items were dropped from the predefined scales. An overreliance zone was defined as the intersection of high disagreement ($D \geq 0.5$) and high satisfaction ($S \geq 4.0$) and is interpreted as an operational indicator of a possible overreliance pattern.

Descriptive statistics were reported using medians and interquartile ranges for time and disagreement measures, and means for perception scales. Associations were assessed using Spearman correlations. Group differences were tested with the Mann–Whitney U test, reporting Cliff's delta as the effect size. Finally, an ordinary least squares (OLS) regression with heteroskedasticity-robust standard errors (HC3) examined whether satisfaction could be jointly explained by perceived efficiency gains (SavingPct), disagreement, and perceived utility.

The study was conducted in an educational setting using anonymized student-generated materials, with results reported only in aggregated form.

4. Results

4.1. Perceived efficiency gains and AI–human scoring disagreement (RQ1–RQ2)

Across sessions with a valid manual baseline (n = 113), participants reported substantial perceived efficiency gains when completing workplace risk assessments with LLM assistance. Median LLM-assisted time was 20 min, compared to 120 min for the estimated manual baseline, corresponding to a median proportional time saving of 0.84. Positive time savings were reported in 97.3% of sessions with valid baselines (Table 1). Figure 1 visualizes the strong separation between the distributions of LLM-assisted and manual baseline times.

As a verification-burden proxy, a session-level disagreement index was computed from absolute AI–human differences in probability and hazard impact scoring. Among sessions with valid scoring pairs (n = 97), median disagreement was 0.50, and 26.8% showed mean disagreement ≥ 1.0 (Table 1), indicating non-trivial divergence between AI-generated and human-assigned scores in a meaningful share of cases. At the hazard level, exact agreement was lower for probability (43.9%) than for hazard impact (59.9%), with visible off-diagonal dispersion in the mid-to-high scoring range (Fig. 2).

Table 1. Descriptive statistics for efficiency, disagreement, and user acceptance across session-level analytic samples

Construct	Measure	n	Statistic	Value
Efficiency	LLM-assisted time (T_{ai})	113	Median [IQR]	20 [10–35]
	Estimated manual time (T_{manual})	113	Median [IQR]	120 [90–240]
	Time saving (SavingPct)	113	Median [IQR]	0.84 [0.75–0.92]
Verification-burden proxy	Mean AI–human disagreement (1–5 scale)	97	Median [IQR]	0.50 [0.25–1.00]
	High disagreement (≥ 1.0)	97	% (count/n)	26.8% (26/97)
Perceptions (Likert 1–5)	Metrics realism (1–5)	121	Mean (SD)	4.18 (0.62)
	Competence/accuracy (1–5)	121	Mean (SD)	4.25 (0.67)
	Utility (1–5)	120	Mean (SD)	4.41 (0.70)
	Satisfaction / intention to reuse (1–5)	120	Mean (SD)	4.06 (0.88)
AB indicator	Overreliance zone share	96	% (count/n)	40.6% (39/96)

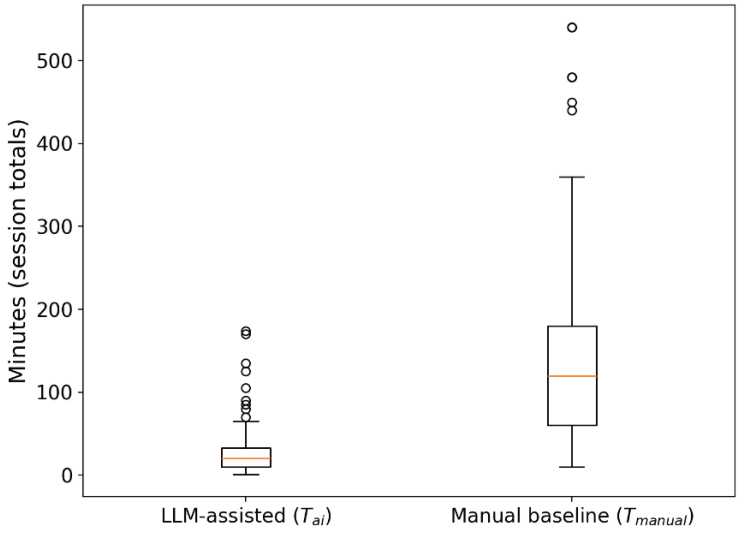

Fig. 1. Distribution of session-level LLM-assisted time and estimated manual baseline time (n = 113).

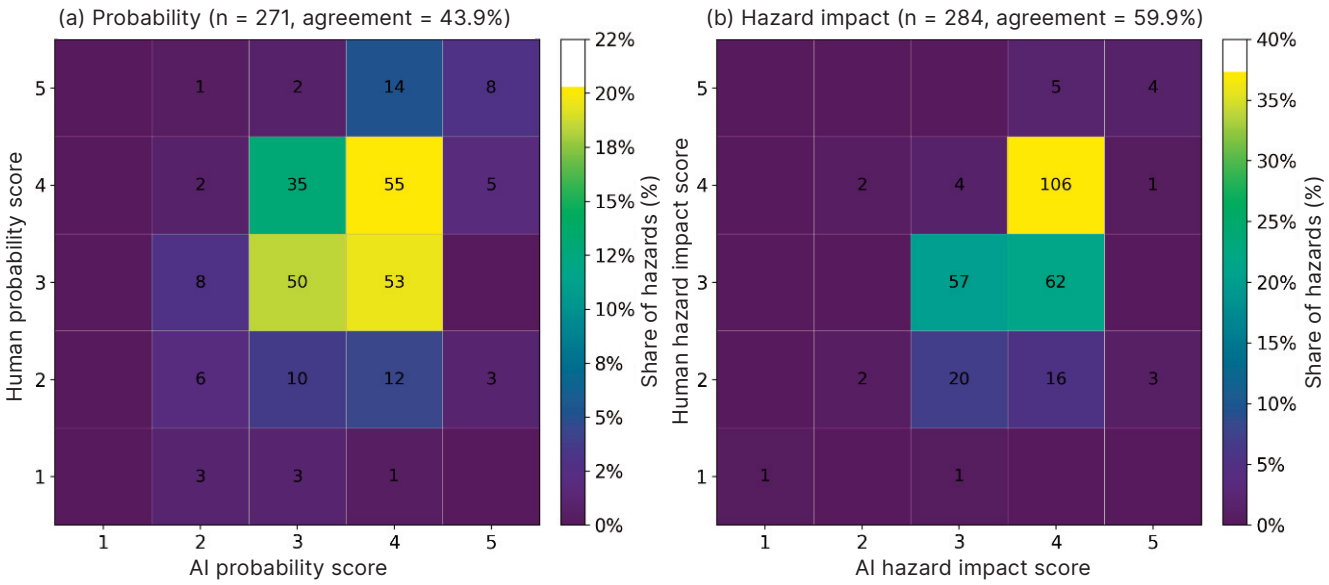

Fig. 2. AI–human agreement in quantitative risk scoring (hazard level): (a) probability and (b) hazard impact.

4.2. User acceptance and possible overreliance patterns (RQ3)

Despite scoring disagreements, participants reported high acceptance of the LLM-assisted workflow. All perception scales averaged above 4.0, including utility and satisfaction / intention to reuse (Table 1). Internal consistency was acceptable to high across the four composite scales: metrics realism ($\alpha = 0.896$, 7 items), competence/accuracy ($\alpha = 0.799$, 3 items), utility ($\alpha = 0.902$, 4 items), and satisfaction / intention to reuse ($\alpha = 0.862$, 3 items). Disagreement was negatively associated with satisfaction (Spearman $\rho = -0.26$, $p = 0.012$), whereas the relationship between time saving and satisfaction was small and marginal ($\rho = 0.20$, $p = 0.053$).

Figure 3 maps disagreement against satisfaction using the median disagreement (0.50) and a satisfaction cutoff of 4.0. Among sessions with complete data (n = 96), 40.6% fell into the high-disagreement/high-satisfaction quadrant (overreliance zone), indicating frequent cases of strong tool acceptance even when participants assigned scores that diverged substantially from AI-generated values.

Finally, an OLS regression model predicting satisfaction (Table 2) showed that perceived utility was the strongest positive predictor ($\beta = 0.98$, $p < 0.001$), while disagreement remained a significant

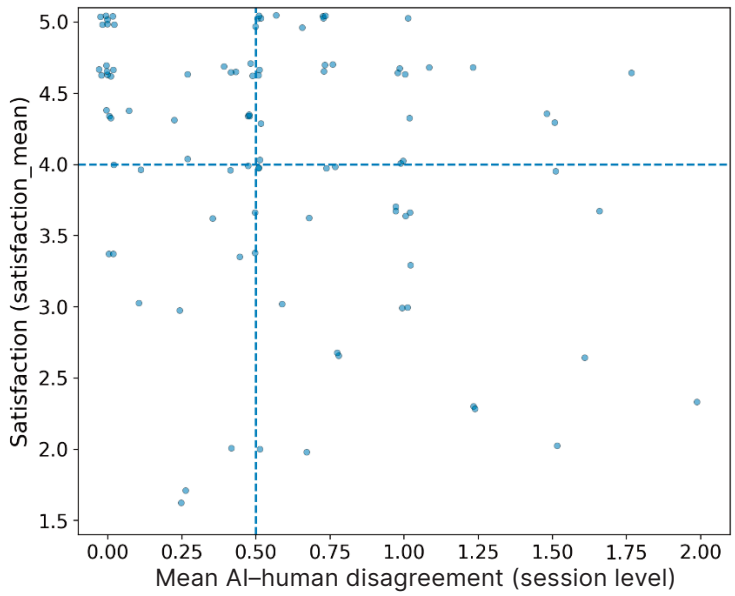

Fig. 3. Disagreement–satisfaction relationship and the overreliance zone. Dashed lines indicate the median disagreement threshold (0.50) and the satisfaction cutoff (4.0).

Table 2. Robust OLS model predicting satisfaction (1–5); N = 91, R^2 = 0.729, adj. R^2 = 0.719; HC3 robust standard errors; β* denotes standardized coefficients

Predictor	Coef. (β)	β*	Robust SE	t	p
Intercept	0.089	–	0.287	0.31	0.756
SavingPct	–0.134	–0.041	0.144	–0.93	0.350
Mean disagreement	–0.398	–0.211	0.126	–3.17	0.002
Utility	0.985	0.830	0.061	16.21	< 0.001

negative predictor (β = –0.40, p = 0.002). Perceived proportional time saving was not statistically significant once the other predictors were included (p = 0.35). The model explained 72.9% of variance in satisfaction (R^2 = 0.729). Standardized coefficients confirmed that utility was the dominant predictor (β* = 0.830), followed by disagreement (β* = –0.211), while time saving was negligible (β* = –0.041). A supplementary check for nonlinearity revealed no significant quadratic terms for any of the three predictors (SavingPct^2, mean disagreement^2, utility^2; all p > 0.30), supporting the adequacy of the linear specification.

5. Discussion

This study mapped the efficiency–verification trade-off in an LLM-assisted workplace risk assessment. The workflow appears to offer substantial perceived speed advantages, but the findings also suggest an assurance cost: quantitative risk scores often showed substantial AI–human disagreement, while acceptance remained high even when disagreement was non-trivial. This creates a governance dilemma for sustainable safety management: perceived time efficiency is attractive, but it does not automatically imply reliability or well-calibrated trust.

5.1. Perceived efficiency gains (RQ1)

The median perceived time saving of 84% documented in this study indicates a substantial perceived shift in OHS process management. A reduction from a 120-minute manual baseline to a 20-minute AI-assisted session mirrors the "efficiency leap" observed in construction risk assessment by Martin et al. (2025), who recorded a move from 157 to 36 minutes.

However, our findings suggest that this efficiency is primarily perceived. By using T_{manual} as a subjective baseline, we captured the user's internal valuation of the tool. This strong perceived reduction in effort may serve as an important driver of technology adoption, but as Sun and Kalar (2025) note, such gains often come with hidden costs. The lack of statistical significance for SavingPct in our regression suggests that while speed may act as an initial adoption driver, it is not the primary driver of long-term satisfaction once the user begins interacting with the tool's output.

5.2. AI–human disagreement as a verification-burden proxy (RQ2)

Despite strong efficiency perception, scoring disagreement suggests a non-trivial verification-burden proxy. Median AI–human disagreement was 0.50 (1–5 scale), and 26.8% of sessions showed mean disagreement ≥ 1.0, suggesting that scoring divergence – and by implication, potential verification demands – remains substantial in a meaningful subset of cases. Probability judgments showed lower exact agreement than hazard impact (43.9% vs 59.9%), which is consistent with probability being more context-dependent (exposure, local conditions) and therefore harder for LLMs to infer reliably from limited textual input (Collier et al. 2024; Lee et al. 2026). This aligns with the view that LLMs can support structured hazard narratives while still requiring active human calibration for numeric prioritization (Nyqvist et al. 2024).

5.3. The overreliance zone and a possible overreliance pattern (RQ3)

One notable finding is the identification of the overreliance zone, encompassing 40.6% of sessions. In these cases, users reported high satisfaction alongside substantial AI–human scoring disagreement. This pattern may pose a challenge for sustainable safety governance and may be interpreted through the lens of automation complacency (Klingbeil et al. 2024). When the perceived benefit of the system is high, partly due to the large perceived time saving observed here, users

may, as Kücking et al. (2024) suggest, develop a "blindness to potential errors." This may reflect the "gulf of impatience" (Qiao et al. 2025), where the desire to finalize the assessment quickly outweighs the cognitive strain of rigorous auditing. For sustainable management, this implies that high user satisfaction may be an insufficient proxy for process quality; a satisfied user is not necessarily a vigilant one.

5.4. What drives satisfaction: utility matters most, disagreement still matters (RQ2–RQ3)

The regression results are consistent with the interpretation above. Perceived utility was the strongest positive predictor of satisfaction, while higher disagreement significantly reduced satisfaction. Meanwhile, the time saving variable did not remain significant once utility and disagreement were included.

This suggests a plausible behavioral pattern. Perceived efficiency may support initial adoption, utility may sustain intention to reuse, and disagreement may act as a friction cost that reduces satisfaction but often does not eliminate acceptance – helping explain why a possible overreliance pattern may persist despite recognized scoring divergence. This is important because it helps clarify how such patterns may arise in ways consistent with AB-related concerns: people may recognize scoring divergence (hence, disagreement hurts satisfaction) but still judge the tool as "worth it" because it remains useful overall (Choudhury and Shamszare 2024; Kücking et al. 2024).

5.5. Implications for sustainable digital safety governance

From a management perspective, the results support a practical recommendation: organizations should treat LLM-based risk assessment as a potentially speed-enhancing system that still requires structured assurance controls, especially for numeric risk metrics.

Practical implications include:

- Do not equate satisfaction with assurance. Acceptance can remain high even when quantitative outputs show substantial AI–human disagreement.
- Track AI–human disagreement as a verification-burden indicator. Monitoring scoring divergence or disagreement rates can approximate verification burden and surface reliability issues.
- Design and train for verification. Workflow design should include explicit review steps (e.g., "why this score?" prompts) and user training focused on the most fragile component – probability calibration in context-dependent environments (Collier et al. 2024).

5.6. Limitations

This study has several limitations. First, the dataset originates from an educational setting and reflects student-based validation, which may differ from professional OHS practice in important ways. Compared to students, expert assessors likely bring domain-specific contextual knowledge that could reduce AI–human scoring divergence, as they may be better positioned to evaluate site-specific probability and impact estimates. At the same time, experts may exhibit different trust calibration patterns – either greater skepticism toward AI outputs due to professional standards, or conversely, higher complacency if time pressure is a dominant factor in real workplace settings. Whether the overreliance zone identified here would persist, shrink, or shift in composition among experienced practitioners remains an open empirical question. Second, the manual-time baseline was self-estimated rather than experimentally measured, so efficiency results should be interpreted as perceived savings rather than objective time-and-motion evidence. Third, the disagreement metric captures AI–human divergence, not ground-truth correctness; therefore, it should be interpreted as a proxy for verification effort and calibration rather than an accuracy benchmark. Finally, sessions varied in the number of workplaces and hazards assessed, which introduces heterogeneity in workload and opportunities for AI–human scoring divergence.

5.7. Future work

Future research should replicate the analysis in professional OHS contexts using a comparative design that directly contrasts student and expert assessors on disagreement magnitude, satisfaction, and overreliance patterns under matched task conditions. Such studies should also include observed manual baselines rather than self-estimated ones, to provide objective validation of time-effect

estimates. Additional work could examine whether disagreement patterns differ by hazard type, workplace context, or user expertise, and whether structured interface interventions (e.g., mandatory justification steps) reduce possible overreliance patterns. Finally, qualitative analysis of edits and reasoning could clarify why users accept AI outputs even when numeric scores diverge from human-assigned values.

6. Conclusion

This study suggests that the LLM-assisted workplace risk assessment delivers strong perceived efficiency benefits but introduces a potential verification burden, reflected in consistent AI–human disagreement in probability and impact scoring. A substantial share of sessions exhibited high satisfaction despite high disagreement, indicating a possible overreliance pattern consistent with AB-related concerns that challenges the use of acceptance metrics as a proxy for process reliability. For sustainable digital safety management, the results highlight that speed gains should be paired with explicit verification structures so that efficiency does not come at the expense of risk governance quality.

Data availability statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgment

The publication costs of this article were partially covered by the Estonian Academy of Sciences.

References

Aladağ, H. 2023. Assessing the accuracy of ChatGPT use for risk management in construction projects. *Sustainability*, **15**(22), 16071. https://doi.org/10.3390/su152216071

Choudhury, A. and Shamszare, H. 2024. The impact of performance expectancy, workload, risk, and satisfaction on trust in ChatGPT: cross-sectional survey analysis. *JMIR Human Factors*, **11**(1), e55399. https://doi.org/10.2196/55399

Collier, Z. A., Gruss, R. J. and Abrahams, A. S. 2024. How good are large language models at product risk assessment? *Risk Analysis*, **45**(4), 766–789. https://doi.org/10.1111/risa.14351

Falegnami, A., Tomassi, A., Corbelli, G., Nucci, F. S. and Romano, E. 2024. A generative artificial-intelligence-based workbench to test new methodologies in organisational health and safety. *Applied Sciences*, **14**(24), 11586. https://doi.org/10.3390/app142411586

Goddard, K., Roudsari, A. and Wyatt, J. C. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, **19**(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Klingbeil, A., Grützner, C. and Schreck, P. 2024. Trust and reliance on AI – an experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, **160**, 108352. https://doi.org/10.1016/j.chb.2024.108352

Kücking, F., Hübner, U., Przysucha, M., Hannemann, N., Kutza, J.-O., Moelleken, M. et al. 2024. Automation bias in AI-decision support: results from an empirical study. In *German Medical Data Sciences 2024*. IOS Press, Amsterdam, **317**, 298–304. https://doi.org/10.3233/SHTI240871

Lee, J., Park, S., Oh, S. and Ma, B. 2026. Can large language models automate the HAZOP process without human intervention? *Safety Science*, **194**, 107039. https://doi.org/10.1016/j.ssci.2025.107039

Lyell, D. and Coiera, E. 2017. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, **24**(2), 423–431. https://doi.org/10.1093/jamia/ocw105

Martin, H., James, J. and Chadee, A. 2025. Exploring large language model AI tools in construction project risk assessment: Chat GPT limitations in risk identification, mitigation strategies, and user experience. *Journal of Construction Engineering and Management*, **151**(9), 04025119. https://doi.org/10.1061/JCEMD4.COENG-16658

Nyqvist, R., Peltokorpi, A. and Seppänen, O. 2024. Can ChatGPT exceed humans in construction project risk management? *Engineering, Construction and Architectural Management*, **31**(13), 223–243. https://doi.org/10.1108/ECAM-08-2023-0819

Oral, M., Alboga, Ö., Aydınlı, S. and Erdis, E. 2026. Usability of large language models for building construction safety risk assessment. *Engineering, Construction and Architectural Management*, **33**(9), 7021–7048. https://doi.org/10.1108/ECAM-08-2024-1143

Parasuraman, R. and Riley, V. 1997. Humans and automation: use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **39**(2), 230–253. https://doi.org/10.1518/001872097778543886

Qiao, H., Vermeulen, J., Fitzmaurice, G. and Matejka, J. 2025. To use or not to use: impatience and overreliance when using generative AI productivity support tools. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan*. Association for Computing Machinery, 1–18. https://doi.org/10.1145/3706598.3714103

Romeo, G. and Conti, D. 2026. Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. *AI & SOCIETY*, **41**, 259–278. https://doi.org/10.1007/s00146-025-02422-7

Sun, N. and Kalar, D. 2025. Gemini at work: knowledge workers' perceptions and assessment of productivity gains. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference, Funchal, Madeira, Portugal*. Association for Computing Machinery, 3681–3695. https://doi.org/10.1145/3715336.3735679

Tõhususe ja usaldusväärsuse tasakaal LLM-toega töökoha riskihindamisel

Aleksandr Bosler, Tarmo Koppel ja Karin Reinhold

Suuri keelemudeleid (*large language model*, LLM) võetakse töötervishoius ja tööohutuses üha enam kasutusele töökoha riskihindamise kiirendamiseks, kuid nende lõimimine tekitab dilemma tõhususe ja usaldusväärsuse vahel. Uuringus vaadeldakse seda väljakutset LLM-toega töövoos, kasutades hariduskeskkonnast pärit 111 üliõpilastöö failide andmeid, mida analüüsiti sessioonitasandil. Sõltuvalt mõõdikute kättesaadavusest saadi 96–121 analüütilist sessiooni. Tõhusust mõõdeti tajutud ajasäästu kaudu võrreldes hinnangulise käsitsi tehtava lähteolukorraga, samas kui verifitseerimiskoormuse asendusnäitaja määratleti tehisintellekti ja inimese lahknevusena kvantitatiivses hindamises – tõenäosuse ja mõju skoorides 1–5-punktisel skaalal. Kasutajate omaksvõttu mõõdeti kasulikkuse ning rahulolu ja taaskasutamise kavatsuse Likerti skaalade abil.

Tulemused näitavad märkimisväärset tajutud tõhususe kasvu: LLM-toega töö mediaanaeg oli 20 minutit võrreldes 120 minutiga käsitsi tehtava töö korral, mediaanne ajasääst oli 0,84. Verifitseerimiskoormuse asendusnäitaja jäi siiski arvestatavaks: mediaanne lahknevus oli 0,50 ning 26,8%-l sessioonidest esines lahknevus ≥1,0. Omaksvõtt püsis kõrge – rahulolu keskmine oli 4,06/5. Märkimist on väärt asjaolu, et 40,6%-l sessioonidest ilmnes võimalik liigse usalduse muster, kus kõrge rahulolu eksisteeris koos suure lahknevusega. Robustses vähimruutude meetodi (*ordinary least squares*, OLS) regressioonis oli kasulikkus tugevaim rahulolu ennustaja, samas kui lahknevus vähendas rahulolu; ajasääst ei olnud statistiliselt oluline.

Seega võivad tajutud kiirusevõidud eksisteerida koos varjatud verifitseerimis- ja kontrollikuludega ning võimaliku usalduse väärkalibreerimisega. Sellest tulenevalt tekib vajadus kontrollistruktuuride ja sihipärase koolituse järele kestliku digitaalse ohutusjuhtimise tagamiseks.