

Human-aligned AI for supply chain risk detection: validation and calibration of LLM-based early-warning assessments

Logistics
Research

Janis Purk
IHI Zittau, TU Dresden, Zittau, Germany

Received 30 December 2025
Revised 11 June 2026
Accepted 25 June 2026

Abstract

Purpose – This paper investigates how a large language model (LLM)-based early-warning system (EWS) for supply chain risk can be systematically aligned with human expert judgement. Building on prior work that introduced the overall system architecture, this study focuses on the detailed prompting logic underlying the automated analysis and on an empirical validation of the resulting AI-based risk assessments. The purpose is to examine how human evaluations can be used to validate and calibrate AI-generated classifications and numerical risk scores, thereby improving their accuracy, reliability and interpretability. In doing so, the paper contributes to a human-centric understanding of AI-supported decision-making in the context of Logistics 5.0.

Design/methodology/approach – The study draws on two datasets from a real-world validation conducted in January and March 2025, in which AI-generated risk indicators – including impact, likelihood, credibility, sentiment and risk category – were systematically compared with human expert assessments. Deviations between model and expert evaluations were analyzed using established performance metrics such as mean absolute error (MAE), classification accuracy and error distribution patterns. The methodological design follows a human-in-the-loop approach: expert assessments serve as an operational reference standard and are used to iteratively calibrate the prompting logic underlying the automated analysis. Quantitative evaluation is complemented by design-oriented insights derived from the iterative refinement of the system.

Findings – The findings demonstrate that a structured, iterative human-in-the-loop validation process successfully improves the quality of AI-based risk assessments. In the initial phase, limited contextual grounding led to conservative model behavior, particularly midpoint bias in numerical ratings such as severity, probability and temporal horizon. Prompt refinements systematically integrated supply-chain-relevant anchor examples, partly adapted to company-specific contexts. This enhanced the model's ability to interpret events in operational terms rather than abstract scales. Across the refined prompts, MAE decreased consistently, confirming that context-aware prompt calibration through expert feedback is an effective mechanism for aligning AI outputs with human judgement.

Research limitations/implications – The validation is based on two empirical evaluation rounds with a limited number of pilot companies, which constrains immediate generalization across industries and regions. Future research may extend the approach to larger datasets, additional sectors and alternative LLM variants. Nonetheless, the study provides a transferable methodological template for validating and calibrating AI-driven risk detection systems in logistics. More broadly, the findings underline the importance of embedding domain knowledge and expert judgement into AI-based classifications rather than relying solely on probabilistic model behavior, thereby supporting human-centric decision-making in line with Logistics 5.0 principles.

Practical implications – The results support logistics managers in assessing when and how AI-generated alerts can be trusted in operational risk monitoring. By documenting the prompting logic and workflow design, the study enables organizations to implement comparable EWSs using low-code environments. The findings show how structured human feedback reduces noise and misclassification, helping firms focus on truly critical disruptions, avoid unnecessary reactive measures and strengthen resilience and sustainability in supply chain operations.

Social implications – By aligning AI-generated assessments with human expertise, the proposed system reinforces human oversight rather than replacing it. This supports transparency, trust and responsible AI use in operational decision-making. In line with Logistics 5.0 principles, the approach reduces cognitive overload caused by information noise and enables more balanced, informed and socially responsible responses to supply chain disruptions.

© Janis Purk. Published in *Logistics Research*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) license. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this license may be seen at <http://creativecommons.org/licenses/by/4.0/>



Logistics Research
Emerald Publishing Limited
e-ISSN: 1865-0368
p-ISSN: 1865-035X

DOI 10.1108/LORE-12-2025-0184

Originality/value – This paper advances research on AI-based supply chain risk monitoring by providing a transparent and empirically grounded analysis of how LLMs can be calibrated through human expert judgement. Beyond presenting the prompting logic of an operational LLM-based EWS, the study demonstrates how structured expert feedback, context-specific anchor examples and iterative prompt refinement systematically improve model behavior across multiple risk dimensions. In contrast to prior conceptual, simulation-based or purely performance-driven studies, this work shows how and why human-in-the-loop calibration reduces systematic biases rather than merely reporting accuracy gains. By explicitly linking prompting logic, expert validation and operational decision logic, the paper contributes a transferable methodological approach and strengthens the practical and theoretical foundations of human-centric Logistics 5.0 risk management.

Keywords Large language models (LLMs), Prompt engineering, Business process automation, Early warning systems, Human-centric AI, Supply chain risk detection

Paper type Research article

1. Introduction

Global supply chains are increasingly exposed to complex and rapidly evolving risks arising from geopolitical conflicts, climate-related disruptions, infrastructure failures, regulatory shifts and socio-economic instability. Organizations must therefore process large volumes of weak signals originating from heterogeneous information sources such as news outlets, social media, governmental alerts and industry reports. Traditional early-warning mechanisms, which often rely on manual monitoring or static rule-based indicators, are no longer sufficient to detect emerging threats in a timely and consistent manner (Ivanov and Dolgui, 2021; Ghadge *et al.*, 2020). As a result, firms face growing challenges in distinguishing relevant disruptions from background noise and translating external events into internal implications for their supply chains.

Advances in natural language processing (NLP) and large language models (LLMs) offer new opportunities to automate parts of the risk detection process by extracting, classifying and evaluating information at scale (Li *et al.*, 2024; Kumar *et al.*, 2023). LLMs can interpret unstructured text, assign sentiment and detect event types without task-specific training, enabling significantly faster scanning and pre-assessment of potential risks. However, despite impressive semantic capabilities, LLM-generated classifications remain probabilistic and may diverge from domain-specific reasoning used by human experts (Bommasani *et al.*, 2021). Especially in high-stakes environments such as supply chain risk management, uncalibrated AI outputs can lead to false alarms, overlooked hazards or misinterpretations of sector-specific relevance.

Logistics 5.0 emphasizes a human-centric paradigm that integrates advanced automation with expert oversight, ensuring that technology enhances rather than replaces human judgement (Modica *et al.*, 2023; Caliskan *et al.*, 2024). In this context, early-warning systems (EWSs) must not only automate data processing but also align AI-generated risk signals with human interpretation standards. Achieving such alignment requires transparency regarding how LLMs evaluate risk dimensions, how prompts shape model behavior and how deviations from expert assessments can be systematically corrected.

Previous work laid the foundation for this research stream, as a conceptual framework for automated supply-chain risk analysis using LLMs, outlining core dimensions such as impact, likelihood, credibility and sentiment, was proposed (Purk *et al.*, 2025). Subsequently, the iterative development of an operational EWS including its architecture and Power Automate-based workflow integration, was presented (Purk and Vallée, 2026). However, both studies explicitly left two crucial elements unexplored:

- (1) the detailed prompting logic and evaluation heuristics embedded in the automation flows, and
- (2) a systematic empirical validation comparing LLM-generated risk assessments with expert judgement.

Addressing these gaps is essential for the scientific and practical advancement of human-centric AI in supply-chain risk management. Without validated scoring mechanisms and transparent prompting logic, organizations risk deploying AI systems that produce inconsistencies or biases that remain undetected until they have operational consequences.

This paper contributes to the Logistics 5.0 discourse by presenting a human-centric approach to the design and validation of an LLM-based EWS for supply chain risk monitoring. Rather than focusing on predictive performance alone, the study addresses how AI-generated risk assessments can be systematically aligned with expert judgement through structured prompting and validation procedures. Specifically, the paper:

- (1) documents the prompting logic and decision rules underlying an operational LLM-based EWS, thereby increasing transparency into how risk indicators are generated;
- (2) introduces a two-phase validation design in which AI outputs are compared with expert assessments from five industrial pilot companies using established evaluation metrics and
- (3) outlines an iterative calibration approach that integrates expert feedback to refine prompting logic and contextualization.

By linking prompt engineering, contextual information and human evaluation, the study provides a methodological foundation for human-in-the-loop AI in logistics risk management. The empirical analysis and its implications for operational robustness and decision support are presented in the subsequent sections.

2. State of the art

Research on AI-supported EWSs spans multiple streams, including supply chain risk analytics, human–AI collaboration and the technical design of automated decision-support systems. The following subsections synthesize these streams in order to identify conceptual and empirical gaps relevant to the validation and calibration of LLM-based risk assessments.

2.1 Early-warning systems and supply chain risk analytics

Supply chain EWSs traditionally rely on predefined indicators, threshold-based monitoring and deterministic rule sets (Trkman and McCormack, 2009). While such systems can detect known risks efficiently, they struggle with emerging, ambiguous or weak signals – especially when these are embedded in unstructured textual data (Sadeek and Hanaoka, 2023). As global disruptions increase in frequency and complexity, modern risk analytics must capture a broader set of information sources and process them with higher temporal resolution.

AI-driven approaches have increasingly been proposed to enhance risk detection (Ghadge et al., 2020; Ivanov and Dolgui, 2021). Traditional machine-learning models such as SVMs, random forests or topic models can classify events or detect anomalies but require structured training data and lack flexibility across domains. LLMs, by contrast, can interpret unstructured text, identify causal relations and evaluate sentiment or severity without task-specific retraining (Bommasani et al., 2021). These capabilities make them promising candidates for automated EWS components.

However, AI-generated evaluations are not inherently reliable. Studies show that LLMs can misinterpret context, overestimate severity, hallucinate facts or reflect source biases (Ji et al., 2023). In risk contexts, such deviations can have substantial operational consequences. Thus, the integration of LLMs into EWS requires mechanisms for calibration, supervision and alignment with domain-appropriate reasoning – an area where research remains sparse.

2.2 Human–AI collaboration, explainability and human-centric logistics

Human-centricity is a core principle of Logistics 5.0 (Modica *et al.*, 2023; Nicoletti and Appolloni, 2024). Rather than fully delegating cognitive tasks to machines, organizations are encouraged to design socio-technical systems where humans provide domain expertise, contextual reasoning and ethical oversight, while AI contributes to speed, scalability and pattern recognition.

Human–AI collaboration is especially valuable in risk assessment, where human experts excel at evaluating plausibility, contextual constraints, organization-specific risk appetite and cross-dependencies (Caliskan *et al.*, 2024). Research shows that hybrid decision-making systems outperform both purely automated and purely manual approaches when human feedback is integrated systematically (Seeber *et al.*, 2020).

Explainability and interpretability also play a critical role:

- (1) Experts must understand how AI arrives at scores.
- (2) Systems must provide transparency around evaluation criteria.
- (3) Decision rules must be adjustable to reflect organizational needs.

Despite increasing attention to HITL systems in logistics (Nicoletti and Appolloni, 2024; Seeber *et al.*, 2020), empirical studies on human–AI alignment in risk detection remain limited. Few works examine how expert annotation can improve LLM behavior or how prompting logic can be iteratively refined to reduce deviations from human judgement. Existing human-in-the-loop AI and decision-support studies provide important foundations for combining algorithmic outputs with human judgement, interactive feedback and iterative system evaluation (Amershi *et al.*, 2014; Holzinger, 2016; Hevner *et al.*, 2004; Gregor and Hevner, 2013). However, these studies rarely examine how such principles can be operationalized in LLM-based EWSs for supply-chain risk monitoring. This study therefore differs by focusing on the joint analysis of prompting logic, company-specific context, low-code workflow integration and expert-based validation within a deployed system. Recent work on LLM evaluation and trustworthy AI further underlines that model performance cannot be assessed by output accuracy alone. Trustworthy AI frameworks emphasize reliability, transparency, accountability, explainability and the management of harmful bias as central requirements for AI systems (NIST, 2023; OECD, 2024). This is particularly relevant for LLM-based decision support, as recent studies show that both human and LLM-based evaluations may be affected by systematic judgement biases and prompt-sensitive inconsistencies (Chen *et al.*, 2024; Wei *et al.*, 2024). In operational risk contexts, plausible but misaligned classifications may still appear linguistically convincing, which increases the need for structured validation and human oversight. Accordingly, human-in-the-loop validation can be understood not only as a practical calibration mechanism, but also as a governance mechanism that makes deviations visible, supports accountability and reduces the risk of uncritical reliance on probabilistic model outputs.

2.3 Prompt engineering and low-code automation workflows

Prompt engineering is a central lever for controlling LLM behavior, affecting consistency, bias, reliability and interpretability (White *et al.*, 2023). Structured prompts, explicit scoring instructions and multi-step reasoning can significantly improve output quality (Chang *et al.*, 2023). However, most academic work focuses on general NLP tasks rather than domain-specific risk assessments. Supply chain research has yet to document detailed prompting strategies for multi-dimensional risk scoring.

Parallel to this, low-code environments such as Microsoft Power Automate, n8n or Zapier enable scalable workflow orchestration for ingesting data, triggering LLM evaluations, validating outputs and notifying stakeholders (Microsoft, 2025). Yet the logistics research literature rarely describes concrete automation flows or their integration with risk analytics tools.

As outlined in [Section 1](#), the architecture of an EWS using Power Automate was introduced in previous work as an important step in this research stream ([Purk and Vallée, 2026](#)); however, the following aspects were not detailed:

- (1) the prompting logic,
- (2) evaluation heuristics or
- (3) decision rules underlying the flows.

Thus, current research lacks both methodological transparency and empirical evidence regarding how LLM-based risk detection should be implemented and validated.

2.4 Research gap

The literature reveals three critical gaps:

- (1) There is limited empirical evidence comparing LLM-generated risk assessments with expert judgement in supply chain risk contexts.
- (2) Existing studies provide limited transparent insight into how prompting logic and scoring heuristics shape multi-dimensional risk evaluations in operational LLM-based EWSs.
- (3) Existing work does not empirically evaluate how iterative expert feedback can be used to systematically calibrate and stabilize AI-generated risk assessments in logistics contexts.

This paper addresses these gaps by presenting a comprehensive human-aligned validation study of an operational LLM-based EWS for supply chain risk detection. In doing so, it shows how expert judgement can be operationalized as a calibration mechanism for multi-dimensional LLM-based risk assessments in an operational supply-chain setting.

3. System architecture and automation flows

The developed EWS is based on a modular architecture in which several Power Automate flows operate sequentially ([Figure 1](#)). The structure follows the principle of automatically capturing large volumes of unstructured external information, evaluating them through multiple LLM-supported analytical steps and informing stakeholders only when necessary. The architecture thus combines automation with human decision support, as required in the context of Logistics 5.0: automation performs the information compression, while experts retain interpretative control.

3.1 System overview

The system consists of three central data structures and four functional flows.

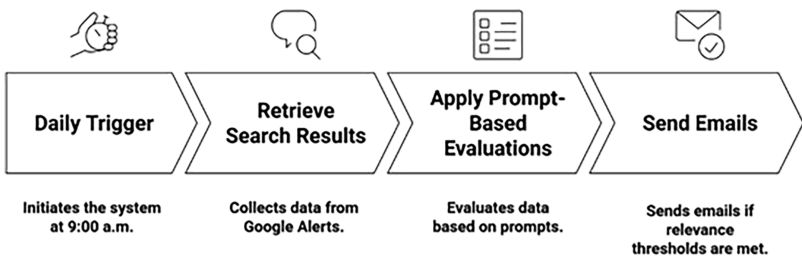


Figure 1. High-level workflow of the early-warning system. *Source:* Author's own work

LORE

Each data structure fulfils a distinct role:

- (1) *Company data* contain firm-specific characteristics such as material dependencies, supplier locations or production sensitivities.
- (2) *Search configurations* define company-specific monitoring keywords and activation parameters.
- (3) *Search results* store the incoming articles and all subsequent LLM evaluations.

This structure clearly separates configuration, context and incoming signals, ensuring that an article is transformed step by step into a complete risk assessment rather than being evaluated in one monolithic step.

The overarching workflow, shown in [Figure 1](#), is divided into four stages:

- (1) the system is triggered daily,
- (2) new search results are retrieved,
- (3) a multi-stage LLM evaluation is performed and
- (4) notifications are sent when thresholds are exceeded.

A daily trigger initiates the retrieval of new search results. These are subsequently processed through a multi-stage LLM evaluation. If predefined thresholds are exceeded, automated notifications are sent.

The workflow is implemented using a low-code automation stack built on Microsoft Power Automate. All steps shown in [Figure 1](#) – daily triggering, retrieval of search results, prompt-based evaluation and notification – are executed as cloud-based flows. SharePoint serves as the system's structured data backbone, storing company profiles, search configurations, intermediate model outputs and final evaluation results. The analytical components are executed via API calls to the ChatGPT 3.5 Turbo model, which processes each article together with the contextual information prepared in Flow 2. This combination of workflow automation and LLM-based analysis enables reproducible execution, transparent auditability and rapid adaptability without the need for custom software development.

3.2 Flow 1: automated retrieval of search results

The first flow serves the sole purpose of information acquisition. It is triggered once per day and reads all monitoring keywords stored in the search configurations. For each active keyword, new items from external sources (e.g. Google Alerts or comparable feeds) are retrieved and stored as individual entries. Each entry includes structural metadata such as title, link, timestamp and the reference to the respective search configuration and associated company.

This flow produces the raw data foundation. No content-related analysis takes place at this stage; the focus lies on robust and repeatable collection of potentially relevant information.

3.3 Flow 2: context enrichment and data preparation

Before an article can undergo a substantive risk assessment, contextual information must be consolidated. This task is performed by the second flow, which follows Flow 1 immediately after a new entry has been created in the search results.

The flow loads the associated company profile, enriches the structural metadata of the article and prepares the text for model-based analysis (e.g. normalization, removal of technical artefacts). Through this enrichment, subsequent model evaluations rely not only on textual evidence but also on company-specific exposure patterns.

The second flow therefore generates the complete input required for the actual LLM analysis, without carrying out an evaluation itself.

3.4 Flow 3: multi-prompt LLM evaluation (core analytical flow)

The third flow constitutes the analytical core of the system and forms the basis for the empirical investigations presented in [Chapter 5](#). Through a sequential multi-prompt design, each article is evaluated along several risk-relevant dimensions. Splitting the assessment across multiple prompts enhances interpretability and consistency, as the model workload is distributed across clearly defined subtasks.

The evaluation begins with a brief content-based interpretation of the article, resulting in an initial justification of relevance. This is followed by a classification of the event type, in which the LLM categorizes the article as risk-related, opportunity-related or irrelevant, and assigns it to an appropriate risk category.

Only after this initial classification does the following justification generation, risk-categorizing and fine-grained numerical scoring of risk dimensions take place. These dimensions include:

- (1) risk impact (operational consequences),
- (2) time horizon,
- (3) likelihood of occurrence,
- (4) credibility of the source and the described event and
- (5) sentiment, reflecting the communicative tone of the article.

Each of these dimensions is rated on a scale from 1 to 5, where 5 is always the indicator for the highest score tending to a report of the event as a risk at the end. All outputs are stored in structured fields, resulting in a complete risk profile for each article, which is statistically analyzed in [Chapter 5](#).

3.5 Flow 4: threshold evaluation and notification (operational layer)

The fourth flow represents the operational escalation layer of the EWS. While it is essential for practical deployment, it does not generate additional AI-based assessments and is therefore not part of the empirical validation presented in [Chapter 5](#). Instead, Flow 4 aggregates the structured outputs produced by Flow 3 and evaluates them against configurable, company-specific escalation criteria.

In this step, the numerical risk indicators generated by the multi-prompt evaluation – risk impact, time horizon, likelihood, credibility and sentiment – are combined into a composite risk score. Each dimension is multiplied by a weighting factor that reflects the focal company's internal prioritization logic. Typical configurations assign higher weights to operational impact and likelihood, while credibility and sentiment serve as moderating factors. The specific weighting scheme and threshold values are adjustable and differ across participating companies.

If the aggregated score exceeds a predefined escalation threshold, the system automatically notifies the responsible risk manager via email or Power App. The notification includes the individual indicator values as well as the qualitative justification generated earlier in the pipeline, enabling rapid human review and decision-making.

Because both the weighting factors and threshold values are organization-specific and subject to confidentiality agreements with the pilot companies, they are not disclosed in detail in this paper. Importantly, this aggregation step does not alter or override the underlying AI evaluations; it merely operationalizes them in accordance with firm-level risk tolerance. All scientific analyses therefore focus exclusively on the analytical behavior of Flow 3, which constitutes the core contribution of this study.

3.6 Architectural rationale

The architecture of the system deliberately follows a threefold logic: First, separating the flows enables clear functional structuring, ensuring that LLM evaluations remain transparent and

individually optimizable. Second, the system explicitly integrates company-specific context into the evaluation logic, substantially improving output quality. Third, the complete storage of all intermediate and final results creates a comprehensive audit trail that supports both scientific validation and operational traceability.

With this architecture, the system provides the foundation for the validation analyses described in [Chapter 4](#), which focus on the evaluation logic of the third flow.

4. Methodology and validation design

The methodological approach for evaluating the EWS followed a structured, two-phase validation design. The objective was to assess the internal consistency, interpretability and robustness of eight LLM-based analytical prompts and to iteratively refine the prompting logic where necessary. This chapter describes the validation setup, the overall design of both phases and the analytical procedures used to align human and model-based evaluations. No empirical results are reported in this section; all findings are presented in [Chapter 5](#).

4.1 Validation framework and system context

The EWS processes unstructured news articles through a modular analytical pipeline that uses a series of eight prompts to evaluate each article along distinct risk dimensions. These dimensions include:

- (1) criticality classification,
- (2) justification generation,
- (3) risk category assignment,
- (4) impact severity,
- (5) probability,
- (6) time horizon,
- (7) credibility and
- (8) sentiment.

Each prompt represents a discrete analytical task and produces one structured output per article.

The validation involved five participating pilot companies. These companies were manufacturing-oriented, industrial B2B firms located in Germany, mainly in North Rhine-Westphalia, with procurement- and supply-chain-related risk exposure. Each company provided a structured company profile, including supplier network information, relevant locations and material dependencies, as well as human reference evaluations for the analyzed articles. The empirical validation was based on article-level observations retrieved through the companies' monitoring configurations in two validation phases. In the January 2025 baseline phase, 2,059 incoming news items were processed as valid article-level observations for the initial criticality classification, of which 234 were classified as risk-relevant and passed to the downstream prompts. In the March 2025 re-validation phase, 2,467 incoming news items were processed, of which 318 were classified as risk-relevant and entered the follow-up assessments. In both phases, incoming news items were generated through company-specific monitoring configurations defined by the participating firms. Each configuration consisted of predefined search terms or alerts linked to the respective company profile, and all retrieved items were automatically pushed into the system database for subsequent LLM-based evaluation. Duplicate, technically incomplete or non-interpretable entries were excluded before validation. These human assessments served as the comparative benchmark for evaluating the LLM's outputs and reflected operational decision criteria

rather than objective ground truth. For each participating company, two domain experts from procurement, supply chain risk management or closely related operational functions were involved in the evaluation. The final human reference values were submitted as consolidated company-level assessments after internal expert alignment. Therefore, the validation did not generate independent individual ratings for each case and no statistical inter-rater reliability coefficient was calculated. To reduce anchoring effects, the expert assessments were entered before the corresponding AI-generated results were made available. Accordingly, the expert evaluations are not interpreted as objective ground truth in a strict statistical sense. Rather, they represent an operational reference standard that captures how experienced company representatives assess risk relevance under firm-specific conditions. The validation therefore evaluates the degree to which the LLM-based system aligns with human operational decision logic, not whether it reproduces an externally observable objective truth.

Figure 2 provides an overview of the two-phase validation design and its human-centric logic. The figure illustrates how AI-based assessments and human expert evaluations interact across the baseline assessment, prompt refinement and focused re-validation phases. Using human expert assessments as a validation reference is well-established in risk analysis and decision-support research, particularly in contexts where objective ground truth is unavailable or ill-defined. Prior studies emphasize that expert judgement provides a suitable benchmark for evaluating AI-based risk assessments when operational relevance, contextual interpretation and managerial decision criteria are central to the task (Aven, 2016; Shanteau et al., 2002; Saaty and Vargias, 2012).

4.2 Phase 1 (January 2025): baseline assessment

The baseline evaluation follows established iterative validation principles commonly applied in decision-support systems and design science research. Such approaches recommend an initial comprehensive assessment to identify systematic deviations and calibration needs prior to targeted refinement in subsequent iterations (Hevner et al., 2004; Gregor and Hevner, 2013). Phase 1 served as a comprehensive baseline evaluation of all eight prompts. Each article retrieved by the EWS was assessed in parallel by:

- (1) the LLM (based on the established prompting pipeline), and
- (2) human evaluators from the participating companies.

The purpose of Phase 1 was to generate a complete mapping of human–LLM alignment across all risk dimensions. This included categorical prompts (e.g. criticality, risk category) and numerical prompts (e.g. severity, probability).

Phase 1 thus provided an initial measurement of consistency and interpretability and formed the conceptual basis for identifying which prompts would be subject to refinement in the subsequent iteration.

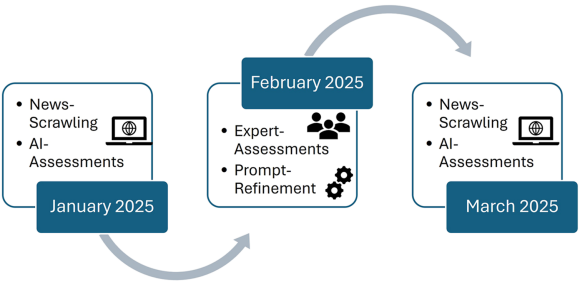


Figure 2. Validation Process. Source: Author's own work

4.3 Prompt refinement approach

Between Phase 1 and Phase 2, selected prompts were refined to improve clarity, interpretability and operational grounding. The prompt refinement strategy aligns with human-in-the-loop approaches for calibrating AI systems through structured expert feedback. Human-centered AI research highlights that systematic interaction between human judgement and model behavior is essential for improving interpretability, robustness and operational reliability, particularly in high-stakes decision contexts (Amershi *et al.*, 2014; Holzinger, 2016; Ribeiro *et al.*, 2020). The refinement process followed a set of general methodological principles rather than targeting specific numerical results. These principles describe the *design space* for refinement and do not imply that all adjustments were applied uniformly:

- (1) Clarification of semantic boundaries: Numeric scales can be restructured to include clear verbal anchor points, enabling more consistent interpretation by the LLM.
- (2) Integration of supply-chain-relevant examples: Illustrative patterns (e.g. operational disruptions, sourcing delays, regulatory signals) can be embedded into the prompt descriptions to reduce ambiguity.
- (3) Sharpening of the model's analytical role: Prompts can be reframed to emphasize the perspective of a supply chain-focused analyst, ensuring that the model incorporates operational context into its decision-making.
- (4) Tightening of output constraints: Output formats can be standardized (e.g. single-label outputs, numeric-only responses), reducing variance and ensuring structured comparability across evaluations.
- (5) Removal of vague or non-operational phrasing: Generalized instructions can be replaced with domain-specific terminology aligned with supply-chain risk logic.

These possible adjustments are optional. Prompts that already exhibited stable logical consistency were retained in their original form. This selective adjustment should enable Phase 2 to focus specifically on the dimensions where methodological enhancement is conceptually justified.

4.4 Phase 2 (March 2025): focused re-validation

Conducting a focused re-validation after targeted design adjustments reflects recommended practice in iterative evaluation cycles, where controlled follow-up assessments are used to isolate the effects of methodological refinements while keeping the company context, preprocessing logic and evaluation criteria comparable (Peffer *et al.*, 2007; Gregor and Hevner, 2013). Phase 2 reassessed only those prompts that should be conceptually refined.

The validation setup intentionally mirrors Phase 1 to ensure full methodological comparability:

- (1) the same participating companies,
- (2) the same company-specific profiles,
- (3) the same set of risk dimensions,
- (4) the same article-based evaluation procedure,
- (5) and the same quantitative metrics (see [Section 4.5](#)).

This is intended to ensure that the effects of prompt refinement could be evaluated under comparable company-specific and procedural conditions. Phase 2 is therefore designed as a controlled follow-up iteration, aimed at assessing whether the revised prompting logic improved internal consistency and interpretability relative to the structured human evaluations.

4.5 Evaluation metrics

To compare human and LLM outputs systematically, the following established evaluation metrics are applied using the variables listed in [Table 1](#):

- (1) *Accuracy, Precision and Recall* for qualitative prompts, defined as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

- (2) *Mean Absolute Error (MAE)* for numerical prompts, defined as

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - y_i|$$

These metrics provide a consistent and interpretable basis for comparing the structured human and machine evaluations across both validation phases. The selected evaluation metrics are widely used in comparable methodological contexts. Accuracy, precision and recall represent standard measures for categorical classification performance, while MAE is particularly suitable for evaluating ordinal and interval-scaled predictions due to its interpretability and robustness against outliers ([Stehman, 1997](#); [Willmott and Matsuura, 2005](#); [Hand and Christen, 2018](#)). Given the exploratory and design-oriented nature of the validation, the analysis relies on descriptive performance metrics rather than inferential significance testing. The objective is not to test a population-level hypothesis, but to examine whether systematic deviations between LLM outputs and expert assessments can be identified and reduced through iterative prompt refinement under comparable operational conditions. The results are therefore interpreted in terms of direction, magnitude and practical relevance of observed alignment improvements.

5. Human-centric AI-validation

This chapter presents all empirical findings derived from the validation metrics introduced in [Chapter 4](#). The analysis focuses on the alignment between AI-generated assessments and human expert judgement across the individual prompts that constitute the analytical core of the EWS.

Table 1. List of variables

Variable	Description
(<i>TP</i>)	True positives: cases correctly classified as positive by the model
(<i>TN</i>)	True negatives: cases correctly classified as negative by the model
(<i>FP</i>)	False positives: cases incorrectly classified as positive by the model
(<i>FN</i>)	False negatives: cases incorrectly classified as negative by the model
(<i>x_i</i>)	Model-generated score for observation (<i>i</i>)
(<i>y_i</i>)	Human expert reference score for observation (<i>i</i>)
(<i>n</i>)	Number of evaluated observations
Source(s): Author's own work	



The validation was conducted in close cooperation with industrial pilot companies under strict confidentiality agreements. Consequently, certain operational details – such as absolute alert volumes, company-specific thresholds and individual firm profiles – cannot be disclosed. All results are therefore reported in aggregated or anonymized form. This does not affect the validity of the comparative analyses, as all human and AI evaluations were performed under comparable conditions across both validation phases. In interpreting the improvements, the analysis considers not only relative changes in MAE, but also the practical relevance of reduced deviations. Since all numerical prompts use a five-point scale, even moderate MAE reductions can be meaningful when they reduce systematic midpoint bias, improve differentiation between low-, medium- and high-risk cases or decrease the frequency of larger deviations. From an operational perspective, such improvements support more reliable prioritization of risk signals and reduce the likelihood that relevant alerts are either overestimated, underestimated or treated as indistinct medium-risk cases.

The following sections examine each prompt in the sequential order of the evaluation pipeline, describing its analytical role, observed deviation patterns and – where applicable – refinements introduced between the two validation phases.

5.1 Criticality classification (prompt 1)

The first prompt (Figure 3) in the evaluation pipeline determines whether an incoming news article constitutes a *Risk*, *Opportunity* or *No impact* for the respective company. Because only items

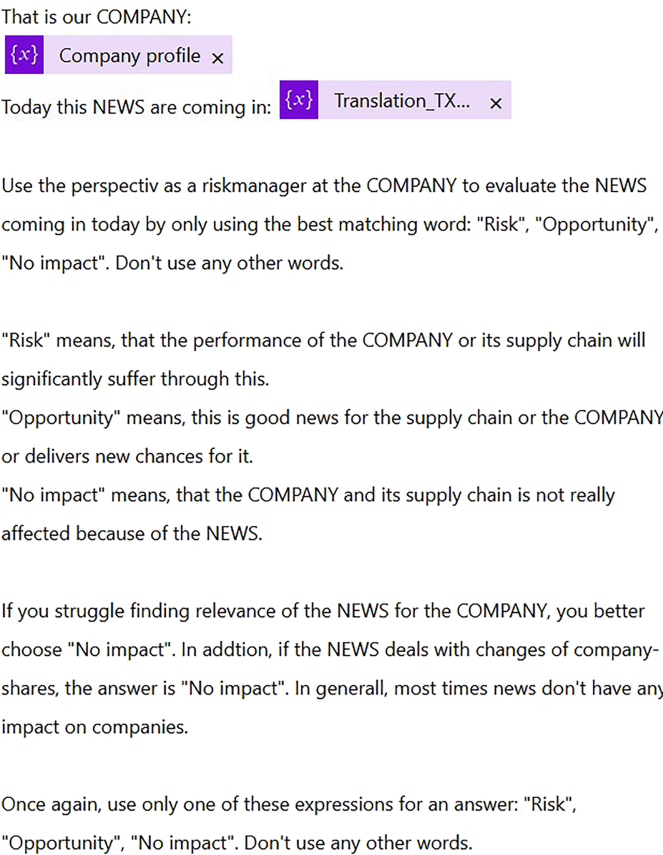


Figure 3. Criticality classification (prompt 1)

classified as *Risk* proceed to further analytical steps, this stage functions as a decisive filter for the entire EWS. Its design therefore emphasizes consistency, contextual grounding and controlled sensitivity.

The prompt follows a strictly constrained instruction format. The model is required to adopt the perspective of a supply-chain risk manager at the focal company and may output only one of three labels. These labels are explicitly defined within the instruction to reduce interpretative drift. The prompt also contains a preference structure: in cases of ambiguity, the model should default to *No impact* to minimize false alarms. Crucially, all evaluations are context-conditioned. Through Flow 2, the entire company profile – including material dependencies, supplier regions and critical infrastructures – is inserted into every invocation of the prompt. As a result, the classification does not rely on textual signals alone, but on the intersection between article content and documented exposure patterns.

This combination of restricted output space, explicit definitions and contextual anchoring ensured a highly stable behavior in Phase 1 (January 2025). The evaluation results are summarized in Table 2.

The quantitative pattern confirms that Prompt 1 behaves deliberately conservative. High accuracy combined with comparatively low recall for *Risk* indicates that the model avoids escalating weak or ambiguous signals – an outcome explicitly preferred by the participating firms. Importantly, all high-salience disruptions (e.g. supplier insolvencies, factory shutdowns, strikes, critical infrastructure failures) were correctly classified as *Risk*. Missed cases were consistently weakly specified, speculative or only indirectly related to the focal company and did not include events perceived by experts as operationally relevant.

False positives were limited and did not pose systemic risk, as downstream prompts (e.g. severity, probability, credibility) further contextualize and moderate the overall risk assessment.

Given this behavior, no adjustments were applied to Prompt 1 for Phase 2. Increasing sensitivity at this stage would have generated additional false positives, increased the burden on human evaluators and conflicted with the preference for lean escalation. Prompt 1 therefore fulfils its role as a robust first relevance filter and provides a stable foundation for all subsequent analytical stages.

5.2 Justification generation (prompt 2)

Following the initial classification, Prompt 2 (Figure 4) generates a concise justification for cases labelled as *Risk*. The purpose of this step is to increase transparency and interpretability

Table 2. Performance of prompt 1 (January 2025)

Metric	Value
Sample size (valid)	2,059
Accuracy (valid entries)	0.72
Precision (Risk)	0.63
Recall (Risk)	0.26
Source(s): Author's own work	

Evaluate as a riskmanager why the following TEXT could mean a risk to our COMPANY.

This is our COMPANY:  Company profile ×

TEXT:  Translation_TX... ×

Only use one sentence for your answer!

Figure 4. Justification generation (prompt 2)

of the automated pipeline by ensuring that each risk decision is accompanied by a short, context-aware rationale. The justification is not used for quantitative scoring but serves as a qualitative audit trail that helps analysts understand why a particular article was deemed relevant.

The prompt used in Phase 1 was deliberately simple yet tightly structured. It instructed the model to provide one sentence, clearly stating *why* the article represents a risk for the company, grounded in both the article content and the previously injected company profile. The instruction included strict output formatting constraints such as: “Write exactly one sentence. Do not add explanations, hedging language or formatting.” This approach prevents overlong or interpretively diffuse responses and keeps justifications operationally useful.

Because the company context is embedded before execution – as described in Flow 2 – the model is able to connect textual events to concrete supply-chain exposures. For example, a strike in a port becomes relevant only if this port appears in the company’s logistics dependencies. Conversely, if the article describes an event unrelated to the company’s materials or supplier geographies, the justification must reflect its irrelevance or focus on the precise element that triggers concern.

This structure led to highly consistent outputs in Phase 1 (Table 3). Across all articles classified as *Risk*, human evaluators assessed 222 out of 234 justifications as correct and appropriate, resulting in a success rate of 95%. The remaining discrepancies were minor and mostly involved stylistic deviations or slight overgeneralization. None of the incorrect justifications altered the core interpretation of risk or masked critical information.

The qualitative evaluation further indicated that the LLM consistently followed the intended logic: it derived its reasoning from explicit textual cues and matched these with exposure information contained in the company profile. This combination produced justifications that were specific rather than generic, for example “because the port affected is crucial for importing Material X from Country Y” rather than “because it may affect the supply chain.”

Given the high accuracy and the absence of consequential errors, Prompt 2 required no modification for Phase 2. The participating companies expressed strong approval for the justification outputs, noting that they supported human interpretive oversight without introducing noise. Furthermore, the consistency and brevity of the generated sentences aligned well with operational needs: analysts could verify risk decisions at a glance without reviewing entire articles.

In summary, Prompt 2 effectively fulfils its intended function of generating precise, contextually grounded rationales for risk-relevant events. Its structure supports transparency and auditability within the automated pipeline, and its near-perfect alignment with human judgement in Phase 1 rendered further refinement unnecessary. Prompt 2 was therefore not re-evaluated in the second validation cycle.

5.3 Risk category assignment (prompt 3)

Once an article has been classified as risk-relevant, Prompt 3 (Figure 5) assigns it to a predefined risk category. This categorization step is essential because it translates textual information into operationally meaningful risk types. It relies on the same contextual input as

Table 3. Performance of prompt 2 (January 2025)

Metric	Value
Valid risk cases evaluated	234
Correct justifications	222
Accuracy	95%
Nature of errors	minor phrasing issues; no operational misinterpretation
Source(s): Author’s own work	

This is our COMPANY: {x} Company profile x
NEWS: {x} Translation_TX... x

Previously to this prompt, it was decided, that the NEWS mean some kind of RISK to our COMPANY.
Use the perspective as a riskmanager at our COMPANY to find a CATEGORY the NEWS are described
the best by only using the best matching expression: Financial Risk, Nature-related Risks, Reputational
/ Image Risks, Geopolitical Risks, Local / Man-made Risks, Cyber Risks, Human Resources Risks. Don't
give out any other words, expressions or explanations.

Figure 5. Risk category assignment (prompt 3)

the preceding steps – namely, the combination of the enriched company profile and the cleaned article text. Unlike Prompt 1, however, Prompt 3 requires more interpretive depth, as the model must identify the operational mechanism through which the event could affect the company. The prompt therefore provides a clear taxonomy with short definitions and instructs the model to select exactly one category.

This taxonomy was intentionally kept broad to ensure interpretability and consistency across different companies. The model's output was also evaluated qualitatively by human experts during the January validation phase.

Across all risk-classified articles in Phase 1, the model demonstrated highly consistent behavior (Table 4). Of 234 risk cases, 217 category assignments were judged correct by human evaluators, corresponding to an accuracy of 92.7%. Errors primarily occurred in articles where multiple risk mechanisms were plausible or where the article text remained vague about the causal pathway of the disruption.

Despite these occasional ambiguities, no misclassification resulted in a fundamentally incorrect understanding of the article's operational relevance. The disagreements were concentrated in borderline scenarios where even human evaluators acknowledged that more than one category could apply. This pattern indicates that category-level interpretation is inherently fuzzy in certain cases, yet the model consistently reproduced the dominant operative mechanism.

With this level of accuracy – and because misclassifications did not propagate into harmful downstream effects – no adjustments were introduced to Prompt 3 for Phase 2. As with Prompt 2, the participating companies expressed no need for refinement, emphasizing that the high-level categorization was “operationally intuitive” and consistent with human expectations.

Prompt 3 thus provides a reliable structuring layer in the analytical pipeline. By mapping articles to macro-level risk mechanisms, it enhances interpretability and contributes to trend analysis. Its strong Phase-1 performance fully justified its exclusion from Phase 2 testing.

5.4 Severity assessment (prompt 4)

The severity prompt (Figure 6) quantifies the operational impact of a risk event on a scale from 1 to 5. Because this dimension directly influences prioritization and escalation behavior in the EWS, the alignment between human and model assessments is particularly important. Severity is evaluated after category assignment and uses both the article text and the full company

Table 4. Performance of prompt 3 (January 2025)

Metric	Value
Valid risk cases evaluated	234
Correct category assignments	217
Accuracy	92.7%
Nature of errors	ambiguous causal mechanisms; dual-category cases
Source(s): Author's own work	

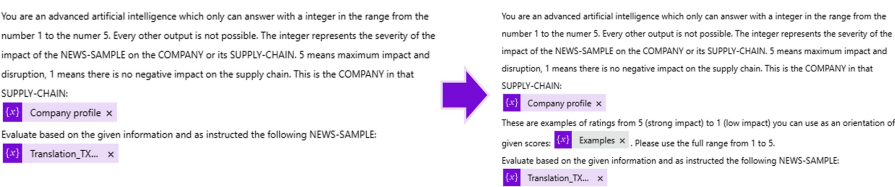


Figure 6. Severity assessment (prompt 4)

profile injected via Flow 2. This contextualization ensures that the impact estimate refers to the specific company rather than general industry dynamics.

In the January validation, the prompt followed a simple structure with coarse definitions for the five scale points (Table 5). The model was instructed to consider operational consequences for the focal company and to output a single numerical value. Despite this clarity, human reviewers quickly observed a strong central tendency: the model overwhelmingly preferred the value 3, producing only limited variation around this midpoint. This pattern led to a MAE of 0.75, noticeably higher than the error values observed for the categorical prompts. Although the direction of impact was usually correct, the model struggled to differentiate between moderate and more substantial disruptions. Extreme values such as 1 and 5 were rarely used, even if human assessors considered them appropriate. The error structure was therefore systematic rather than random, with most deviations falling within ± 1 and only a small number of ± 2 errors occurring in articles with ambiguous or incomplete signals.

Because this structural bias limited the analytical value of the severity output, the prompt was revised before the March validation. The revised version introduced two main changes. First, each value on the 1–5 scale received a precise operational definition. These definitions distinguished clearly between negligible, mild, moderate, substantial and severe disruptions, anchored in typical supply-chain impact patterns. Second, the prompt explicitly encouraged use of the full scale where justified. This addition counteracted the model’s tendency to collapse uncertain cases into the midpoint.

The effects of these changes became clearly visible in the March results (Table 6). The distribution of severity values broadened: 2 and 4s were used substantially more often, and 1

Table 5. Performance of prompt 4 (January 2025)

Metric	Value
Mean absolute error (MAE)	0.75
Most frequent model value	3
Typical deviation	± 1
Extreme deviations	rare; ± 2
Primary issue	central tendency bias
Source(s): Author’s own work	

Table 6. Performance of prompt 4 (March 2025)

Metric	Value
Mean absolute error (MAE)	0.59
Dominant values	3, followed by 2 and 4
Improvement vs January	–0.16 MAE (–21%)
Main effect	greater dispersion; improved sensitivity
Source(s): Author’s own work	

and 5s appeared in cases where article content and company context indicated either very low or very high impact. The recalibration improved alignment with human assessments and reduced the MAE to 0.59, corresponding to a 21% improvement over the January prompt. Most remaining deviations were minor, and extreme misalignments nearly disappeared.

The refined severity prompt thus increased both differentiation and interpretive quality without compromising robustness. Importantly, the improvements did not lead to over-sensitivity; rather, the model became more capable of expressing relative differences in operational relevance while maintaining conservative behavior in ambiguous cases. The severity results therefore indicate that clearer scale anchors can reduce systematic rating bias and improve alignment with expert assessments.

5.5 Probability assessment (prompt 5)

The probability prompt (Figure 7) estimates how likely it is that a risk described in a news article will materialize for the focal company or its supply chain. As in all prompts, the model receives the fully enriched company profile from Flow 2, ensuring that probability is assessed with reference to company-specific exposure patterns rather than general industry behavior.

In the January validation, the prompt provided only brief textual anchors for the five probability levels (Table 7). Although the instructions were clear, the definitions did not sufficiently differentiate between adjacent scale points. As a result, the model exhibited a pronounced central tendency: most outputs were assigned the value 3, even in cases where human reviewers considered the event clearly more or less likely. The distribution was narrow, deviations were typically small (± 1) and extreme deviations were rare. This produced a Mean Absolute Error (MAE) of 0.45, signaling moderate alignment but insufficient discriminatory power.

To address this limitation, the prompt was refined before the March validation. All five scale points were given clearer operational definitions, each linked to representative examples (e.g. accidents for short-term high likelihood, restructuring for medium-term moderate likelihood, geopolitical changes for low likelihood). The revised version also included explicit instructions to use the full scale where context justified it, counteracting the model's tendency to collapse uncertain cases into the midpoint.

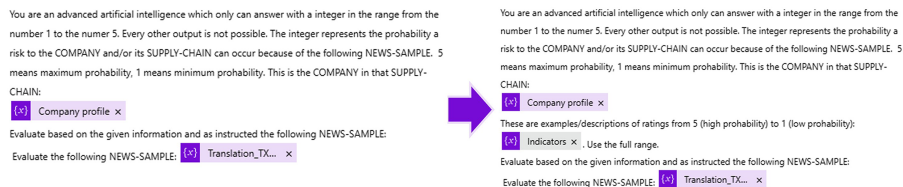


Figure 7. Probability assessment (prompt 5)

Table 7. Performance of prompt 5 (January 2025)

Metric	Value
Mean absolute error (MAE)	0.45
Most frequent model value	3
Typical deviation	± 1
Extreme deviations	very rare; ± 2
Primary issue	central-tendency bias
Source(s): Author's own work	

The March results demonstrate that these adjustments broadened the output distribution (Table 8). While the value 3 remained common, 2 and 4s were now assigned more frequently and 1 and 5s appeared in cases where the article content or company exposure clearly supported them. The MAE improved to 0.38, a reduction of roughly 15% relative to January. Deviations greater than ± 1 became extremely rare, indicating that the prompt changes enhanced sensitivity without increasing noise or false alarms.

Overall, the refined probability prompt produced outputs that were better aligned with human judgement, more responsive to contextual cues and less prone to defaulting to the midpoint. Crucially, the system remained conservative in ambiguous situations but became more capable of signaling very likely or very unlikely events when textual and contextual evidence supported such conclusions. This behavior enhances the EWS’s prioritization logic without increasing the number of false positives – because any over-flagging in earlier steps is subsequently filtered out by the downstream prompts.

5.6 Temporal horizon assessment (prompt 6)

The temporal-horizon prompt (Figure 8) estimates how soon a described event is likely to affect the company or its supply chain. Because this dimension informs the escalation timing, a consistent alignment between human and model judgements is essential. Like all analytical prompts, the temporal evaluation operates on the full company profile injected via Flow 2, ensuring that the LLM anchors its judgement in firm-specific exposure patterns rather than in abstract notions of time.

During the January validation shown in Table 9, the initial version of the prompt relied on a simple 1–5 scale with coarse and relatively abstract descriptions (e.g. “1–2 weeks from now on” vs “1 year from now on”). Although the model generally detected whether an event was immediate or deferred, human reviewers observed substantial variability in how the LLM interpreted these categories. The absence of concrete temporal anchors led to inconsistent differentiation between short-, medium- and long-term consequences.

Table 8. Performance of prompt 5 (March 2025)

Metric	Value
Mean absolute error (MAE)	0.38
Dominant values	3, followed by 2 and 4
Improvement vs January	–0.07 MAE ($\approx -15\%$)
Main effect	broader dispersion; reduced midpoint bias
Source(s): Author’s own work	

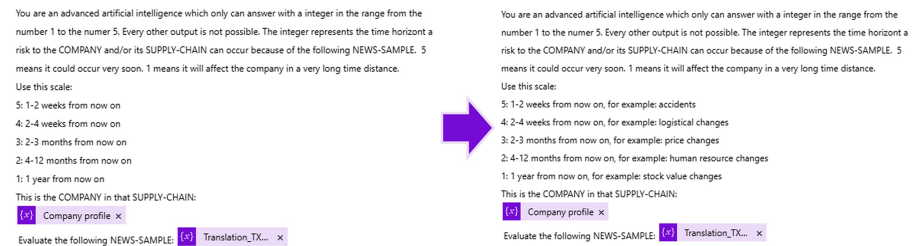


Figure 8. Temporal horizon assessment (prompt 6)

Table 9. Performance of prompt 6 (January 2025)

Metric	Value
Mean absolute error (MAE)	0.71
Most frequent model value	3
Typical deviation	± 1 to ± 2
Extreme deviations	present (up to ± 3)
Primary issue	ambiguous scale; inconsistent temporal mapping
Source(s): Author's own work	

Quantitatively, this resulted in a MAE of 0.71, indicating limited calibration. While many predictions were directionally correct, deviations frequently occurred within ± 1 to ± 2 scale points, and extreme deviations of up to ± 3 were present. These discrepancies were most pronounced in articles containing multiple temporal layers (e.g. immediate disruptions combined with long-term strategic implications) or vague temporal framing.

Because these weaknesses limited the interpretive value of the temporal-horizon output, the prompt was revised before the March validation. The revised version introduced explicit, supply-chain-relevant anchor examples for each scale point. Instead of relying only on abstract time intervals, each value was linked to concrete logistics-related examples. For example, value 5 was associated with acute disruptions such as accidents; value 3 with medium-term effects such as price adjustments or ongoing negotiations and value 1 with long-term developments such as stock value changes. This reformulation significantly reduced interpretive ambiguity and contributed to a more stable temporal calibration of the model outputs.

The effects of these refinements became clearly visible in the March results (Table 10). The LLM produced a much more coherent distribution of values, and the proportion of exact matches with expert assessments increased substantially. Quantitatively, the MAE decreased from 0.71 to 0.51, corresponding to an improvement of approximately 28%.

Most remaining deviations were within ± 1 , representing normal interpretive variation even among human assessors. Larger deviations (± 2) became rare and occurred almost exclusively in articles combining near-term effects with long-term trajectories or speculative future developments. Extreme deviations (± 3) nearly disappeared, indicating that the revised anchor examples effectively constrained the model's temporal interpretation.

Together, these improvements demonstrate the relevance of domain-grounded anchor examples for temporal reasoning in LLM-based risk assessment. The refined prompt enhanced both the stability and interpretive quality of the model's judgements without producing oversensitivity, enabling more consistent differentiation of operationally meaningful temporal differences.

5.7 Credibility assessment (prompt 7)

The credibility prompt (Figure 9) estimates the reliability of the information source referenced in a news article. As with the other numerical dimensions, credibility is evaluated on a 1–5 scale, with 5 representing highly trustworthy sources and 1 indicating low reliability.

Table 10. Performance of prompt 6 (March 2025)

Metric	Value
Mean absolute error (MAE)	0.51
Dominant values	3, followed by 2 and 4
Improvement vs January	–0.20 MAE ($\approx$ –28%)
Main effect	improved calibration; clearer temporal mapping
Source(s): Author's own work	

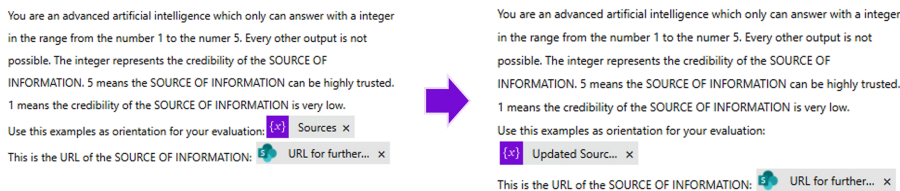


Figure 9. Credibility assessment (prompt 7)

In the January validation, this prompt operated with only minimal guidance (Table 11). The model received a list of exemplary outlets for orientation, but the descriptions for the scale points remained imprecise. While the model did not exhibit a strong central tendency as observed in the severity prompt, its behavior showed a different pattern: credibility estimates were generally reasonable but displayed asymmetry, particularly an over-use of value 4 for moderately reliable outlets. Human evaluators reported that the model distinguished well between high-credibility and mid-credibility sources but occasionally underrated or overrated borderline cases.

Despite this, the numerical alignment between model and human assessors was comparatively strong. Most deviations were limited to ± 1 , with only isolated ± 2 discrepancies, typically caused by ambiguous or composite sources (e.g. aggregator sites referencing original reporting). The January MAE was 0.38, making credibility the strongest-performing numerical dimension in the initial validation phase. Nevertheless, the error structure indicated room for refinement, particularly regarding calibration for mid-tier sources such as specialized blogs or trade magazines, whose reliability varies substantially by context.

To address this, the March prompt revision strengthened the scale anchors and sharpened source categories. Most importantly, the prompt was expanded by explicit examples for each score level, differentiating not only between “trusted” and “untrusted” media but also between public broadcasters, established private news outlets, regional reporting, specialized trade portals and user-generated sources. This contextualization allowed the model to make more consistent distinctions between medium-credibility outlets, which had previously produced ambiguous output patterns. The instruction to use the full range of the scale where justified further reduced clustering around the upper mid-range.

The March results showed a clear improvement (Table 12). The distribution broadened: 3s were used more selectively, 4s no longer dominated and 5s appeared only in cases where human reviewers also agreed. Low scores were applied appropriately for sources lacking transparent journalistic standards. The MAE decreased to 0.31, marking a relative improvement of roughly 18%. As with the earlier prompts, most remaining deviations fell within ± 1 , and no systematic bias remained visible. In short, the refined prompt produced more balanced and interpretable credibility assessments without increasing volatility.

Table 11. Performance of prompt 7 (January 2025)

Metric	Value
Mean absolute error (MAE)	0.38
Most frequent model value	3 and 4
Typical deviation	± 1
Extreme deviations	over-use of value 4
Primary issue	inconsistent treatment of mid-credibility sources
Source(s): Author’s own work	

Table 12. Performance of prompt 7 (March 2025)

Metric	Value
Mean absolute error (MAE)	0.31
Dominant values	3 and 4 with more balanced distribution
Improvement vs January	−0.07 MAE ($\approx$ −18%)
Main effect	improved calibration of mid-credibility sources; reduced over-reliance on value 4
Source(s): Author’s own work	

Overall, credibility stands out as a dimension where even the initial prompt produced relatively strong alignment, yet targeted refinements achieved noticeable improvements in consistency and interpretative value. Importantly, credibility acts as a stabilizing factor in the multi-prompt evaluation pipeline: even when earlier prompts classify a news item as potentially relevant, a low source credibility score often prevents unwarranted escalation, while high-credibility signals reinforce the reliability of detected risks. The March iteration therefore strengthened both analytical robustness and downstream decision support.

5.8 Emotional tone assessment (prompt 8)

The emotional-tone prompt (Figure 10) measures how emotionally charged a news article is, using a scale from 1 (very emotional) to 5 (highly neutral). This dimension functions as a noise-filter in the EWS: emotional or sensational reporting can distort risk perception and should therefore be detected explicitly. As with the other numerical prompts, the model receives both the article and the company context, but the evaluation refers strictly to linguistic tone, not to the risk level or operational impact.

During the January validation, the model showed a clear structural bias (Table 13). It overwhelmingly preferred the value 3, rarely selecting values above or below this midpoint. Extreme ratings such as 1 or 5 were almost absent. As a result, the model produced a highly compressed distribution with limited differentiation. Deviations between human and AI evaluations clustered around ± 1 , with a few ± 2 errors when emotional language was present but not recognized by the model.

This behavior paralleled the patterns observed in the severity and probability prompts: the model generally avoided both low and high scale values, even when strong affective wording (e.g. “devastating”, “dramatic”, “massive criticism”, “intensifying protests”) clearly indicated

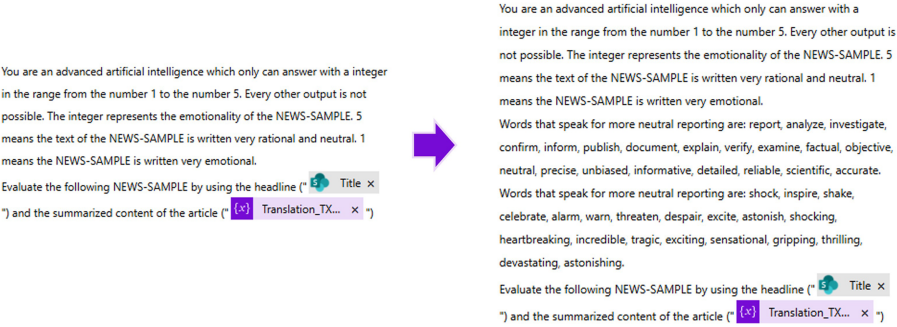


Figure 10. Emotional tone assessment (prompt 8)

Table 13. Performance of prompt 8 (January 2025)

Metric	Value
Mean absolute error (MAE)	0.68
Most frequent model value	3
Typical deviation	±1
Extreme deviations	rare; ±2
Primary issue	strong central-tendency bias
Source(s): Author’s own work	

emotional reporting. The underlying cause was the lack of precise linguistic anchors combined with a broadly phrased instruction.

The prompt was therefore revised before the March validation (Table 14). The updated version introduced explicit linguistic markers for each point on the scale (e.g. evaluative adjectives, alarmistic verbs, modal intensifiers) and provided concrete examples of neutral journalistic phrasing. It also explicitly instructed the model to use the full range of the scale whenever justified by the text. This revision increased granularity and reduced the model’s reluctance to depart from the midpoint. The resulting distribution widened noticeably: 2 and 4s became more frequent, and highly emotional articles were finally assigned the value 1 more reliably. Accuracy improved accordingly.

The refined prompt increased sensitivity without sacrificing stability. Remaining deviations occurred mostly in borderline cases – articles that were broadly factual but contained isolated emotionally loaded terms. Overall, the results demonstrate how linguistic scale anchors can reduce systematic biases and significantly improve human–AI alignment for tone classification.

5.9 Reporting behavior and threshold configuration

While this study validates the analytical behavior of the LLM prompts, the decision whether an article is displayed to users depends on a separate, company-specific reporting configuration. This layer aggregates the validated prompt outputs into a composite risk score using configurable weights and applies an escalation threshold.

Weighting schemes reflect firm-level priorities, typically emphasizing impact severity, likelihood and temporal proximity, while credibility and sentiment act as moderating factors. Thresholds and weights differ across companies and represent operational risk appetite rather than model performance.

Accordingly, reporting behavior is treated as an application layer and is not included in the empirical evaluation. Absolute reporting volumes and configurations are not disclosed due to confidentiality agreements. The analytical results presented in this chapter therefore remain fully comparable and independent of firm-specific escalation settings.

Table 14. Performance of prompt 8 (March 2025)

Metric	Value
Mean absolute error (MAE)	0.54
Dominant values	3, with increased use of 2 and 4
Improvement vs January	–0.14 MAE (≈21%)
Main effect	broader dispersion; better detection of emotional language
Source(s): Author’s own work	

5.10 Summary of prompt-level validation outcomes

After the prompt-specific analyses, Table 15 summarizes the validation outcomes across all prompts. The table synthesizes the detailed results reported above and highlights the main finding for each prompt.

6. Discussion

The validation results demonstrate that LLMs can be systematically aligned with expert judgement when supplied with structured contextual information and carefully engineered prompts. Across all evaluated dimensions – categorizations, justification, severity, temporal horizon, probability, credibility and sentiment – the model exhibited stable behavior and identifiable deviation patterns that could be addressed through targeted prompt refinement. This supports the view that LLMs can meaningfully contribute to early-warning tasks when system design externalizes domain knowledge and constrains model ambiguity.

A central insight of the study is that deviations between AI and human assessments are not random but highly structured. In the initial validation phase, conservative biases dominated, most notably central-tendency effects in ordinal ratings such as severity and probability. These effects occurred primarily when scale definitions were vague. Introducing explicit semantic anchors substantially reduced midpoint bias. The practical relevance of these improvements lies in the improved discriminatory capacity of the system. Lower MAE values are not only statistical indicators of closer alignment; they also indicate that the system became better at distinguishing operationally minor, moderate and more severe risk signals. This is particularly important in early-warning contexts, where prioritization quality determines whether human attention is directed toward the most relevant disruptions. Nevertheless, remaining errors indicate that prompt refinement cannot fully eliminate interpretive ambiguity. Residual deviations occurred particularly in borderline cases, such as articles with vague causal links to the focal company, mixed short- and long-term implications or ambiguous source credibility. These cases show that the system improves structured pre-assessment, but does not replace human judgement in final risk interpretation. The approach should therefore be understood as a decision-support mechanism that reduces information noise and improves prioritization, while still requiring expert oversight for ambiguous or high-impact cases. Similarly, borderline cases with weak or implicit signals revealed the importance of contextual grounding; embedding company-specific exposure profiles proved decisive in preventing generic, industry-level reasoning and improving interpretability.

Table 15. Consolidated overview of prompt-level validation outcomes

Prompt	Key metric	Main finding
1. Criticality classification	Accuracy 0.72; Precision 0.63; Recall 0.26	Stable first-stage filter; deliberately conservative risk selection
2. Justification generation	95% correct	Contextual explanations were largely consistent with expert judgement
3. Risk category assignment	92.7% correct	Risk mechanisms were assigned reliably; errors mainly occurred in ambiguous cases
4. Severity	MAE 0.75 → 0.59	Scale anchors improved differentiation of operational impact
5. Probability	MAE 0.45 → 0.38	Refined scale definitions reduced midpoint tendency
6. Time horizon	MAE 0.71 → 0.51	Domain examples improved temporal interpretation
7. Credibility	MAE 0.38 → 0.31	Source categories improved calibration of medium-credibility sources
8. Sentiment	MAE 0.68 → 0.54	Refined anchors improved detection of emotional tone

Source(s): Author's own work

Taken together, the findings indicate that misalignment can be addressed through systematic prompt calibration rather than ad-hoc correction. This shifts analytical focus away from isolated model performance towards the controllability and transparency of the surrounding evaluation design – a key requirement for operational deployment in risk-sensitive environments.

Beyond conceptual design, system behavior was also shaped by the underlying technical architecture. The early-warning workflow was implemented using a low-code stack comprising Microsoft Power Automate, SharePoint and API-based calls to ChatGPT 3.5 Turbo. This setup enabled rapid prototyping, stable execution and transparent storage of intermediate results, which supported iterative validation. At the same time, the architecture also imposed constraints, including limited process granularity, restricted data modelling flexibility and model-specific sensitivities in fine-grained numerical distinctions. These observations underline that prompt engineering alone is insufficient; robustness and scalability are co-determined by the orchestration environment and toolchain.

From an implementation perspective, the approach also requires organizational capabilities beyond technical deployment. Companies need staff who can maintain search configurations, update company profiles, interpret AI-generated assessments and periodically validate model behavior. Although low-code tools reduce the need for custom software development, they do not eliminate maintenance efforts related to workflow monitoring, API changes, prompt updates and data-quality checks. Governance responsibilities must therefore be clearly assigned, particularly regarding who reviews escalated alerts, who approves changes to prompting logic and how accountability is handled when AI-supported assessments influence operational decisions.

The results further show that reliable escalation does not emerge from any single indicator but from the aggregation of multiple independent risk dimensions. Articles that consistently received elevated scores across several dimensions reached higher composite risk values, while weak or ambiguous signals remained below reporting thresholds. As a result, only a small number of genuinely risk-relevant articles were escalated – an outcome that expert reviewers assessed as operationally appropriate and supportive of focused decision-making.

It is important to distinguish the analytical contribution of this study from the operational configuration of reporting thresholds. While the validation examines the alignment and calibration of the analytical prompts, escalation thresholds reflect company-specific risk appetite and governance structures and are therefore not evaluated comparatively. This separation preserves methodological validity while allowing practical configurability. Building on the empirical setting described in [Section 4.1](#), transferability appears most direct for manufacturing-oriented, industrial B2B firms with procurement- and supply-chain-related risk exposure. The approach is particularly suitable when supplier structures, material dependencies and operational locations can be documented in structured company profiles and used as contextual input for the LLM-based evaluation. In contrast, companies with highly fragmented supplier structures, weakly documented dependencies or rapidly changing risk exposure may require additional calibration before comparable alignment can be expected. At larger scale, the main challenge is less technical processing than the continuous maintenance of search configurations, company profiles, prompting logic and validation routines.

A related limitation is the potential risk of prompt overfitting to the participating companies and their expert interpretations. Although the refinements were based on general principles such as clearer scale anchors, domain-specific examples and stricter output constraints, they were derived from a limited validation context. Future applications should therefore treat the prompting logic as a transferable starting point rather than a fixed universal template and should include periodic re-validation with new companies, sectors and expert groups.

From a Logistics 5.0 perspective, the study reinforces the role of human expertise as a governance and calibration layer. Automation reduces information overload and structures signals at scale, while expert judgement provides a reference frame that constrains AI outputs and supports accountability.

Finally, several limitations must be acknowledged. The validation focused on a limited number of pilot companies, a finite sample size, German-language business news and a single LLM model, which may limit generalizability of the findings. Model behavior may change with future LLM versions, requiring periodic re-validation. Absolute reporting volumes cannot be disclosed due to confidentiality agreements, but this does not affect the comparative alignment analyses conducted under comparable conditions. A further limitation concerns the use of expert judgements as the validation reference. Although the consensus-based assessments reflect the operational reasoning of the participating companies, they remain subjective and may be influenced by company-specific risk tolerance, prior experience or internal prioritization logic. The results should therefore be interpreted as evidence of alignment with structured expert judgement rather than as proof of objective predictive accuracy.

7. Conclusion

This paper presented and validated a multi-stage, LLM-based EWS for supply chain risk monitoring that explicitly integrates human expertise as a calibration and governance mechanism. Building on a real-world deployment and a two-phase validation with domain experts, the study demonstrated that LLMs can be systematically aligned with human judgement when company-specific context and clearly defined evaluation heuristics are provided.

Beyond demonstrating feasibility, the validation showed that initial deviations followed consistent patterns, such as midpoint bias in ordinal ratings or overly generic interpretations in borderline cases. These deviations were reduced through clearer scale anchors, domain-grounded examples, tighter output constraints and company-specific contextualization. The contribution of this study therefore lies in establishing a replicable validation and calibration logic for LLM-based EWSs. The results show that reliable AI-supported risk monitoring emerges from the controlled interaction between structured prompting, contextual information and expert reference evaluations. Overall, the study demonstrates that human-centered calibration is not an optional enhancement but a necessary condition for deploying LLM-based EWSs in operational supply chain contexts, and that systematic prompt validation provides a viable pathway for translating Logistics 5.0 principles into robust, practice-ready AI applications.

8. Future work and outlook

While this study demonstrates how human evaluations can be used to validate and calibrate LLM-based risk assessments within an operational EWS, several avenues for future research remain.

First, future work should examine the scalability of the validation approach under higher data volumes, longer operational horizons, multilingual data sources and additional company contexts. Future studies should investigate whether the observed error patterns remain stable over time and whether the calibration logic remains robust when the system processes larger and more heterogeneous information streams. Second, future research should test the robustness of the prompting logic across different LLM versions and model architectures. Since model behavior may change over time, future studies should examine when recalibration is required and whether the proposed heuristic structures remain stable across models. Third, recent developments in more autonomous, agent-based AI architectures suggest potential extensions for EWSs. However, the primary strength of the system presented in this paper lies in its high degree of automation, reproducibility and

controllability. Structured prompt pipelines enable reliable large-scale screening of unstructured information with minimal human intervention – an essential requirement for operational early-warning use cases. At present, agentic AI approaches remain difficult to integrate at this level of scale and transparency, particularly in risk-sensitive logistics contexts. Any future exploration of such approaches should therefore preserve the system's core advantage: stable, auditable and human-aligned automation.

Finally, future work may extend the human-centric validation logic beyond news-based signals. While this study focused exclusively on unstructured textual information from external reporting, EWSs increasingly combine multiple data sources, such as transactional indicators, sensor data or internal operational logs. Research is needed to understand how human expertise can be integrated consistently across heterogeneous data types without overburdening expert reviewers.

Overall, future research should build on the methodological template presented in this paper by testing its transferability, scalability and durability across contexts, model generations and organizational settings. Doing so will be essential for advancing reliable, automated and human-aligned AI-based EWSs within the broader Logistics 5.0 paradigm.

References

- Amershi, S., Cakmak, M., Knox, W.B. and Kulesza, T. (2014), "Power to the people: the role of humans in interactive machine learning", *AI Magazine*, Vol. 35 No. 4, pp. 105-120, doi: [10.1609/aimag.v35i4.2513](https://doi.org/10.1609/aimag.v35i4.2513).
- Aven, T. (2016), "Risk assessment and risk management: review of recent advances on their foundation", *European Journal of Operational Research*, Vol. 253 No. 1, pp. 1-13, doi: [10.1016/j.ejor.2015.12.023](https://doi.org/10.1016/j.ejor.2015.12.023).
- Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., . . . , and Liang, P. (2021), "On the opportunities and risks of foundation models", *arXiv preprint*, arXiv:2108.07258.
- Caliskan, H., Stentoft, J. and Dubey, R. (2024), "Human-centric logistics and supply chain management: a systematic literature review", *International Journal of Logistics Management*, Vol. 35 No. 1, pp. 1-25, doi: [10.1108/IJLM-05-2023-0194](https://doi.org/10.1108/IJLM-05-2023-0194).
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q. and Xie, X. (2023), "A survey on evaluation of large language models", *arXiv preprint*, arXiv:2307.03109.
- Chen, G.H., Chen, S., Liu, Z., Jiang, F. and Wang, B. (2024), "Humans or LLMs as the judge? A study on judgement bias", *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA, pp. 8301-8327, doi: [10.18653/v1/2024.emnlp-main.474](https://doi.org/10.18653/v1/2024.emnlp-main.474).
- Ghadge, A., Er Kara, M., Moradlou, H. and Goswami, M. (2020), "The impact of Industry 4.0 implementation on supply chains", *Journal of Manufacturing Technology Management*, Vol. 31 No. 4, pp. 669-686, doi: [10.1108/JMTM-10-2019-0368](https://doi.org/10.1108/JMTM-10-2019-0368).
- Gregor, S. and Hevner, A.R. (2013), "Positioning and presenting design science research for maximum impact", *MIS Quarterly*, Vol. 37 No. 2, pp. 337-355, doi: [10.25300/misq/2013/37.2.01](https://doi.org/10.25300/misq/2013/37.2.01).
- Hand, D.J. and Christen, P. (2018), "A note on using the F-measure for evaluating record linkage algorithms", *Statistics and Computing*, Vol. 28 No. 3, pp. 539-547, doi: [10.1007/s11222-017-9746-6](https://doi.org/10.1007/s11222-017-9746-6).
- Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004), "Design science in information systems research", *MIS Quarterly*, Vol. 28 No. 1, pp. 75-105, doi: [10.2307/25148625](https://doi.org/10.2307/25148625).
- Holzinger, A. (2016), "Interactive machine learning for health informatics: when do we need the human-in-the-loop?", *Brain Informatics*, Vol. 3 No. 2, pp. 119-131, doi: [10.1007/s40708-016-0042-6](https://doi.org/10.1007/s40708-016-0042-6).
- Ivanov, D. and Dolgui, A. (2021), "OR-methods for coping with the ripple effect in supply chains during COVID-19 pandemic: managerial insights and research implications", *International Journal of Production Economics*, Vol. 232, 107921, doi: [10.1016/j.ijpe.2020.107921](https://doi.org/10.1016/j.ijpe.2020.107921).

- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A. and Fung, P. (2023), "Survey of hallucination in natural language generation", *ACM Computing Surveys*, Vol. 55 No. 12, Article 248, pp. 1-38, doi: [10.1145/3571730](https://doi.org/10.1145/3571730).
- Kumar, S., Sureka, A. and Goyal, P. (2023), "Large language models for text classification: a comparative evaluation", *Expert Systems with Applications*, Vol. 213, 118863, doi: [10.1016/j.eswa.2022.118863](https://doi.org/10.1016/j.eswa.2022.118863).
- Li, J., Li, X., Wu, Y. and Zhang, Y. (2024), "Large language models in supply chain management: opportunities and challenges", *Computers and Industrial Engineering*, Vol. 187, 109814, doi: [10.1016/j.cie.2023.109814](https://doi.org/10.1016/j.cie.2023.109814).
- Microsoft (2025), "Microsoft Power Automate documentation", Microsoft Learn, available at: <https://learn.microsoft.com/en-us/power-automate/>
- Modica, M., Caliskan, H. and Stentoft, J. (2023), "Logistics 5.0: the role of human-centricity, sustainability and resilience", *Logistics Research*, Vol. 16 No. 1, pp. 1-14, doi: [10.23773/2023_1](https://doi.org/10.23773/2023_1).
- National Institute of Standards and Technology (NIST) (2023), *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, US. Department of Commerce, Gaithersburg, MD, doi: [10.6028/NIST.AI.100-1](https://doi.org/10.6028/NIST.AI.100-1).
- Nicoletti, B. and Appolloni, A. (2024), "Green Logistics 5.0: a review of sustainability-oriented innovation with foundation models in logistics", *European Journal of Innovation Management*, Vol. 27 No. 9, pp. 542-561, doi: [10.1108/EJIM-07-2024-0787](https://doi.org/10.1108/EJIM-07-2024-0787).
- OECD (2024), "OECD AI principles", OECD.AI Policy Observatory, available at: <https://www.oecd.org/en/topics/sub-issues/ai-principles.html>
- Peffer, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S. (2007), "A design science research methodology for information systems research", *Journal of Management Information Systems*, Vol. 24 No. 3, pp. 45-77, doi: [10.2753/MIS0742-1222240302](https://doi.org/10.2753/MIS0742-1222240302).
- Purk, J. and Vallée, F. (2026), "Development of an AI-driven risk analysis tool for transport supply chains", *Transportation Research Procedia*, Forthcoming, Paper Presented at the World Conference on Transport Research (WCTR 2026), Toulouse, 6–10 July.
- Purk, J., Vallée, F., Herrmann, F. and Claus, T. (2025), "Developing an automated approach of risk analysis in supply chains using AI", *Proceedings of the 39th ECMS International Conference*, pp. 412–418.
- Ribeiro, M.T., Wu, T., Guestrin, C. and Singh, S. (2020), "Beyond accuracy: behavioral testing of NLP models with CheckList", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902-4912, doi: [10.18653/v1/2020.acl-main.442](https://doi.org/10.18653/v1/2020.acl-main.442).
- Saaty, T.L. and Vargas, L.G. (2012), *Models, Methods, Concepts & Applications of the Analytic Hierarchy Process*, Springer, New York.
- Sadeek, S.N. and Hanaoka, S. (2023), "Assessment of text-generated supply chain risks considering news and social media during disruptive events", *Social Network Analysis and Mining*, Vol. 13, Article 96, doi: [10.1007/s13278-023-01100-0](https://doi.org/10.1007/s13278-023-01100-0).
- Seeber, I., Bittner, E., Briggs, R.O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A.B., Oeste-Reiß, S., Randrup, N., Schwabe, G. and Söllner, M. (2020), "Machines as teammates: a research agenda on AI in team collaboration", *Information & Management*, Vol. 57 No. 2, Article 103174, doi: [10.1016/j.im.2019.103174](https://doi.org/10.1016/j.im.2019.103174).
- Shanteau, J., Weiss, D.J., Thomas, R.P. and Pounds, J.C. (2002), "Performance-based assessment of expertise: how to decide if someone is an expert or not", *European Journal of Operational Research*, Vol. 136 No. 2, pp. 253-263, doi: [10.1016/s0377-2217\(01\)00113-8](https://doi.org/10.1016/s0377-2217(01)00113-8).
- Stehman, S.V. (1997), "Selecting and interpreting measures of thematic classification accuracy", *Remote Sensing of Environment*, Vol. 62 No. 1, pp. 77-89, doi: [10.1016/S0034-4257\(97\)00083-7](https://doi.org/10.1016/S0034-4257(97)00083-7).
- Trkman, P. and McCormack, K. (2009), "Supply chain risk in turbulent environments—a conceptual model for managing supply chain network risk", *International Journal of Production Economics*, Vol. 119 No. 2, pp. 247-258, doi: [10.1016/j.ijpe.2009.03.002](https://doi.org/10.1016/j.ijpe.2009.03.002).

- Wei, H., He, S., Xia, T., Wong, A., Lin, J. and Han, M. (2024), "Systematic evaluation of LLM-as-a-judge in LLM alignment tasks: explainable metrics and diverse prompt templates", *arXiv preprint*, arXiv:2408.13006.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J. and Schmidt, D.C. (2023), "A prompt pattern catalog to enhance prompt engineering with ChatGPT", *arXiv preprint*, arXiv:2302.11382.
- Willmott, C.J. and Matsuura, K. (2005), "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE)", *Climate Research*, Vol. 30 No. 1, pp. 79-82, doi: [10.3354/cr030079](https://doi.org/10.3354/cr030079).
-

Corresponding author

Janis Purk can be contacted at: janis.purk@mailbox.tu-dresden.de