

Trust-aware evaluation frameworks for large language model reliability in enterprise AI platforms

Swaroop Suresh Borukar *

Independent Researcher, Workday Inc., San Jose, CA.

International Journal of Science and Research Archive, 2026, 20(01), 889–898

Publication history: Received on 05 May 2026; revised on 05 July 2026; accepted on 07 July 2026

Article DOI: <https://doi.org/10.30574/ijrsra.2026.20.1.1330>

Abstract

Trust-aware evaluation is an emerging but rapidly consolidating research area for assessing the reliability of large language models in enterprise AI systems. Generative AI brings a new set of reliability challenges that extend beyond traditional accuracy metrics: An incorrect answer could affect key business decisions, customer interactions, knowledge management, security setups, or organizational accountability. This review summarizes peer-reviewed journal articles published between 2015 and 2025 related to key aspects of trust-aware LLM evaluation, including hallucination and factuality assessment, LLM evaluation methods, trustworthy AI governance, and human trust calibration. As indicated in the literature, current evaluation practice is still fragmented, both in terms of the various technical metrics and in terms of the documentation instruments, as well as on the interpretation of the data and the user-centred design of the trust cues, organizational governance. The most critical gaps are weak alignment between benchmark outcomes and enterprise risk, limited post-deployment monitoring, insufficient context-specific trust calibration, and limited validation of evaluation frameworks in operational platforms. The article argues for an evidence-based approach that connects model behaviour, platform controls, user reliance, and auditable governance in enterprise reliability assessment.

Keywords: AI Governance; Enterprise AI Platforms; Large Language Models; Reliability Evaluation; Trust Calibration; Trustworthy AI

1. Introduction

Evaluating the reliability of large language models in a trust-aware manner is challenging because enterprise use involves uncertainty, prompt variation, diverse user needs, and business-critical knowledge. Trustworthy-AI research shows that reliability depends not only on technical performance but also on transparency, accountability, fairness, robustness, privacy, and human oversight. This proves especially important for enterprise AI platforms, where language models are not used as standalone prediction tools. These models reside in workflow systems, retrieval pipelines, compliance procedures, customer interfaces, software engineering platforms, and decision-support systems.

A second constraint arises from human trust research. Trust is not synonymous with satisfaction, liking a model, or perceiving it as intelligent. Empirical studies of trust in AI also indicate that trust in AI can lead to inappropriate reliance when systems are fluent, confident or social in their performance, but lack task-specific reasoning capability [2]. The difficulty is exacerbated because large language models can express uncertainty, lack of evidence, and sometimes inconsistent reasoning in natural language. A model evaluation approach must account not only for whether an output is accepted, rejected, escalated, or over-applied, but also for the conditions under which enterprise users make those decisions.

* Corresponding author: Swaroop Suresh Borukar

The field is interdisciplinary. Computer science contributes benchmark design, hallucination taxonomies, robustness testing, and explanation methods. Information systems scholarship emphasizes enterprise capability, enterprise governance, enterprise platform integration and enterprise risk. Human–Computer Interaction research examines calibrated trust, interface-mediated dependence, and the communication of uncertainty. Management and organizational studies address responsibility, sociotechnical control, and AI implementation. An accuracy-driven review would not adequately examine the central issue of reliability in relation to organizational purpose, control mechanisms, and warranted human reliance on model behavior.

There is a lack of literature that is cohesive and easily accessible in a single research stream. Trustworthy AI reviews [3] and generative AI scholarship [4] frames new opportunities and risks for business and information systems at a high level. However, relatively few studies directly examine trust-aware reliability assessment of large language models in enterprise AI systems. Most LLM-specific assessment efforts focus on benchmark coverage, factuality, hallucination or capability comparison. Most enterprise AI studies are either on implementation, governance or capability building, and lack evaluation of the reliability of the language model at the prompt, retrieval, user, and audit levels.

The objective of this review is to examine the current state of peer-reviewed journal literature and propose a lightweight trust-aware evaluation framework for enterprise LLM reliability. The review focuses on research pertinent to the assessment of model dependability, human trust calibration, and organizational governance and platform-level assurance. It does not equate trust with a nebulous moral goal. Rather, trust is treated as an assessable relationship among evidence, model behaviour, user reliance, and enterprise controls.

2. Literature Review

Peer-reviewed studies on hallucination and LLM evaluation provide the most direct technical and empirical basis for trust-aware reliability assessment. Research on hallucination shows that generation failures can arise from the training data, modelling goal, decoding process, external knowledge mismatch, and task ambiguity [5]. Hallucination is a problem that cannot be considered a rare fault in the output of enterprise AI platforms. It is a persistent reliability risk, which varies by domain, prompt style, retrieval setup, and tolerance for uncertainty. A second challenge revealed by evaluation surveys of LLMs is that model capability is often assessed using large families of benchmarks, whereas for enterprise reliability, it is necessary to have more focused evidence regarding the factual grounding, consistency, robustness, safety, instruction following, and domain-specific performance of models [6]. Therefore, it is justifiable to move from a single-score benchmark to multidimensional assurance, as this is supported by literature.

Generative AI scholarship in information systems further highlights enterprise-level issues that are not fully captured by typical NLP benchmarks. Generative AI is increasingly treated as a platform technology that will transform knowledge work, service interaction, content creation and decision making [7]. But platform value is dependent on organizational safeguards: scope of work, validation cycles, management duties, data stewardship, and controlled integration with business processes. Management-oriented AI research also calls for the management of AI as a sociotechnical capability and not a mere technical artefact [8]. From this perspective, the LLM reliability evaluation should focus on processes of acquisition, configuration, monitoring, documentation, and updating within enterprise systems.

Table 1 provides an overview of the eight peer-reviewed studies selected from the literature to inform the proposed trust-aware evaluation framework of enterprise LLM reliability. The table shows that there is no single method that is used in the field. Rather, the field integrates evidence from LLM assessment, hallucination analysis, generative-AI platform research, governance research, capability measurement, transparency documentation, responsible-trust design, and explainable AI.

There's a recurring disconnect between technical assessment and organizational assurance, as described in the table. Evidence relating to model behaviour comes from hallucination and benchmark studies, but evidence to inform procurement, monitoring, responsibility and user escalation is lacking. In contrast, enterprise-AI governance and capability studies define organizational needs but are less likely to include specific measures for LLMs. The challenge is to connect both forms of evidence without collapsing their distinct purposes: technical performance measurement and managerial assurance.

Table 1 Summary of key findings

Ref	Focus	Key Findings
[5]	Hallucination in natural language generation	Hallucination is linked to data, modelling, decoding, and grounding failures; factuality evaluation must distinguish intrinsic inconsistency from unsupported external claims.
[6]	LLM evaluation methods	LLM assessment spans knowledge, reasoning, robustness, safety, and alignment; benchmark fragmentation limits direct enterprise comparability.
[7]	Generative AI as an information systems phenomenon	Generative AI creates platform-level value but introduces reliability, accountability, and control problems in knowledge-intensive organizations.
[8]	Management of AI in organizations	Enterprise AI requires governance across data, models, users, and organizational routines rather than isolated algorithmic performance evaluation.
[9]	AI capability in organizations	Organizational value depends on AI resources, skills, coordination, and business integration; evaluation must connect model performance to capability maturity.
[10]	AI service documentation through FactSheets	Supplier declarations can structure transparency about purpose, training data, metrics, robustness, and governance controls for AI services.
[11]	Responsible trust in AI systems	Trust should be designed as calibrated reliance through communication of competence, uncertainty, process, and boundary conditions.
[12]	Explainable AI concepts and taxonomy	Explainability supports accountability and user understanding, but explanation quality depends on context, audience, and evaluation purpose.

Explainability research adds an important qualification. The general conclusions that have been drawn from broad surveys of explainable AI suggest that explanations have the capacity to aid transparency, contestability, and debugging, but that there is a wide variety of explanation methods, with varying fidelity, stability, and audience fit [12]. The concept of “interpretable-model scholarship” warns that post hoc explanation cannot be used to replace reliable modelling in high-stakes settings [13]. Previous surveys of black-box explanation methods also reveal that local explanations, feature attribution, example-based and rule-based explanations answer different questions and cannot be freely interchanged [14]. In LLM platforms, explanations, citations, rationales, retrieval traces, confidence signals, and refusal messages may support reliability assessment only when they are faithful, evidence-based, and appropriate to the user and task context. In LLM platforms, citations, rationales, retrieval traces, confidence signals, and refusal messages support reliability assessment only when they are faithful, evidence-based, and appropriate to the user and task.

Research on bias and fairness extends the concept of reliability beyond factual correctness. Unfair outcomes can creep into a system through data representations, labels, model selection, evaluation sets and deployment environments [15], as demonstrated by bias surveys. These concerns extend to enterprise LLM platforms via recommendations, summarization, automated support, and knowledge retrieval. Trust-aware evaluation must therefore include demographic, linguistic, role-based, and domain-specific stress tests, in addition to aggregate answer quality, in the evaluation of trust-aware systems.

Finally, research in the field of trust in automation offers a conceptual framework. The critical factors in trust include system performance, process transparency, characteristics of users, risk of the task, and interaction history [16]. Additionally, meta-analytic evidence suggests reliability, validity, process and design characteristics are all important contributors to trust in automated systems [17]. The scholarship of human-centred AI therefore focuses on the operational goal of reliable, safe and trustworthy design as a value, not as an abstract concept [18]. For enterprise LLM platforms, these findings point to the need for the use of a trust-aware evaluation metric that considers the appropriateness of user reliance on the model trust when appropriate, distrust when warranted, and escalation through appropriate governance channels when uncertainty is high.

3. Methodology

Based on the reviewed literature, four broad families of evaluation methods can be identified: benchmark-oriented evaluation, error- and risk-oriented evaluation, transparency and documentation, and human reliance and governance assessment. Its advantage is comparability across models and releases; its weakness is that enterprise scenarios depend on specific domain expertise, retrieval system, workflow limitations, and user roles which are not easily captured in public benchmarks [6]. Strong general capability is not always a reliable indicator of acceptable performance in a contract review, customer support, cyber triage, financial analysis, or other high-risk enterprise workflows.

The second family is error-oriented and risk-oriented evaluation. The output failures, unsupported claims, representational harms, and domain-specific risk can be identified via hallucination taxonomies and fairness reviews, which provide structured categories [5], [15]. These techniques are useful since enterprise trust is compromised by particular failure modes, not because of low average performance. This family is, however, metrically unstable. Hallucinations can be annotated differently across annotators, evidence sources, and task definitions. Bias evaluation depends on access to relevant attributes, benchmark quality, and locally appropriate definitions of harm. Risk-oriented methods thus need to be backed up by severity scales, escalation thresholds and contextual documentation.

Transparency and documentation constitute the third family. Evaluation evidence is translated into artefacts for stakeholders to review, including FactSheets, explainability taxonomies and responsible trust design studies [10]–[12]. Documentation is essential to enterprise platforms since model selection, vendor evaluation, audits, and internal approvals should be based on documented evidence that will last beyond the end of any given experiment. But keeping records of declarations, without independent testing, monitoring logs, incident review and update control, can be performative. Its value lies in requiring evaluation evidence to be traceable, comparable and accountable.

Human reliance and governance assessment is the fourth family. These three aspects (perceived reliability, perceived transparency of automation and perceived user experience) are reported to affect reliance [16], [17] in the field of trust in automation. Some studies in information systems research [8, 9] find that enterprise value is a function of organizational capability and management routines, and platform integration. These are broader evaluations of organizational capability. Roles based access, boundaries of tasks, fall-back, quality review, user training, incident reporting, and change management are all features of a good LLM platform.

Table 2 Method comparison

Ref	Method	Strengths	Limitations
[5]	Hallucination taxonomy and factuality analysis	Separates grounding errors, internal inconsistency, and external factual conflict.	Does not by itself define enterprise severity, escalation, or business impact.
[6]	Benchmark survey across LLM capabilities	Maps evaluation dimensions including reasoning, safety, robustness, and alignment.	Public benchmark performance may not transfer to proprietary workflows.
[10]	FactSheet-style service documentation	Creates audit-ready declarations about data, metrics, intended use, and safeguards.	Quality depends on the completeness, verifiability, and independence of the supporting evidence.
[11]	Responsible trust communication model	Links system cues, uncertainty, competence, and user reliance.	Requires empirical validation in specific enterprise tasks and user groups.
[12]	Explainability taxonomy	Differentiates explanation goals, audiences, and technical methods.	Explanations can be persuasive without being faithful or decision-relevant.
[16]	Trust-in-automation review model	Identifies performance, process, purpose, and user factors affecting trust.	Developed mainly for automation broadly, not LLM-specific interaction.

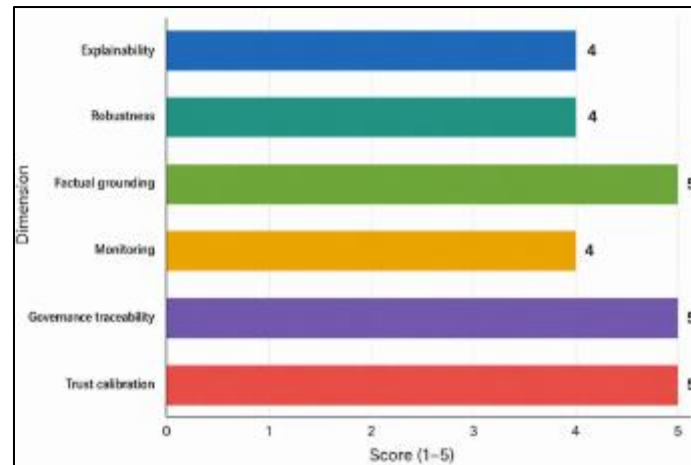


Figure 1 Trust-aware reliability evaluation dimensions for enterprise LLM platforms

Figure 1 shows a radar chart of six trust-aware reliability dimensions for enterprise LLM platforms. This figure is important because enterprise reliability depends on balancing multiple forms of evidence rather than optimizing factuality or benchmark accuracy alone. The dimensions were coded from selected peer-reviewed literature [5], [6], [10], [11], [16], [17].

4. Discussion

The reviewed studies indicate that LLM reliability in enterprise AI platforms cannot be assessed using a single performance measure. The strongest technical finding is that reliability evidence must be disaggregated. Research in hallucination [5] indicates that there are different ways in which factual error can develop: source conflict, training-data absence, generation behaviour and lack of grounding. According to a survey of LLM evaluation, these specific tasks of reasoning, knowledge recall, safety, robustness and instruction following are typically assessed by distinct benchmark suites [6]. This fragmentation is scientifically helpful in preventing distinct capabilities from being treated as equivalent. Simultaneously, it poses an enterprise challenge: a reliable judgment about a model needs to be made across a diversity of evidence. A trust-aware framework should therefore be able to associate each metric with an enterprise control question.

The second trend is that reliability is platform-dependent. The focus of generative AI research has shifted from treating LLMs as standalone models to treating them as integral components of larger information systems [7]. Research in enterprise management further demonstrates that the outcomes of an AI system are also influenced by the routines, governance and maturity of capabilities of the organization [8, 9]. This finding changes how model evaluation should be understood. A model with moderate stand-alone factuality may be more trustworthy in a well-designed retrieval-augmented platform with proper validation, logging and escalation than a more factual model that is attached to uncontrolled data and weak governance. On the other hand, if the prompts are vague, the retrieval is outdated, tool outputs is unvetted, or users are encouraged to use generated text without critical engagement, a model with strong benchmark performance may still produce unreliable enterprise outcomes. Reliability therefore emerges from interactions among the model, data, interface, workflow, and oversight mechanisms.

A third discovery relates to transparency. It has been shown in documentation frameworks like FactSheets that AI services can be documented through specific statements of purpose, metrics, training context, robustness, and operational constraints [10]. Responsible-trust research emphasizes communicating competence, uncertainty, and system boundaries [11]. Further, from a perspective of explainable AI, it has been stated that explanations should be chosen based on the task and audience. Overall, these findings suggest that enterprise LLM evaluation must generate evidence that is useful for three audiences: technical evaluators to assess behaviour, governance actors to approve deployments, and end users to make reliance decisions. A single explanation object cannot satisfy all three audiences with a single explanation object. For instance, a retrieval trace might be needed to check the sources, a risk officer might need incident-rate dashboards, and a platform engineer might need prompt-level regression results. Trust-aware evaluation therefore requires layered evidence artifacts.

Table 3 Results comparison

Ref	System / Context	Metric / Basis of Comparison	Outcome
[5]	Natural language generation and LLM-style output	Intrinsic versus extrinsic hallucination categories	Factuality failures require evidence-grounding tests rather than fluency assessment.
[6]	LLM benchmark ecosystem	Capability coverage across reasoning, knowledge, safety, and robustness	No single benchmark family provides sufficient enterprise reliability evidence.
[7]	Generative AI in information systems	Platform value versus risk in organizational use	Reliability must be evaluated through task fit, governance, and workflow control.
[8]	Organizational AI management	Governance of data, models, users, and routines	Enterprise AI reliability depends on management structures as much as model design.
[9]	Enterprise AI capability	Resource, skill, and coordination capability	Model performance creates value only when integrated with organizational capability.
[10]	AI service assurance documentation	Structured supplier declarations	Audit-ready transparency improves comparability but requires verification.
[11]	User-facing AI trust design	Communication of competence and uncertainty	Trust should be calibrated to actual system capability and task risk.
[15]	Fairness and bias evaluation	Bias sources across data, modelling, and deployment	Reliability must include group-sensitive and context-sensitive harm assessment.

Table 3 compares trust-aware evaluation with conventional LLM benchmarking. Benchmark studies provide evidence of model behaviour, enterprise studies explain organizational context, and deployment assurance determines whether the evidence is actionable. A model could perform well on general reasoning, but it could be weak on enterprise reliability, such as enforcing citation discipline and access controls, or distinguishing between accepted policies and outdated knowledge, or its failure to escalate in regulated workflows. Conversely, a platform may be well documented and well governed yet remain unreliable if monitoring does not cover prompt drift, retrieval decay, and user overreliance.

Information about inconsistencies in the literature from multiple sources is also informative. Explainability scholarship argues that explanation can be a tool of trust and accountability [12] while interpretable-model scholarship warns that explanation can become a justification for opaque systems in high-stakes situations [13]. Surveys of black-box explanation methods show substantial methodological variation, with different methods addressing different evaluation targets [14]. This tension is evident in rationale-generating enterprise LLM systems. A model can produce plausible but causally unfaithful rationales. Citations may be missing or incomplete, lack strong connections to claims, or selectively used. Trust-aware evaluation must therefore distinguish factual support from persuasive explanation. A reliable framework should measure whether explanations, citations and uncertainty markers will improve users' correct decisions, rather than merely increasing users' confidence.

Fairness research provides another important extension of reliability assessment by highlighting how bias can emerge across data, modelling, evaluation, and deployment contexts [15]. Bias is not merely a standalone responsible-AI problem, but enterprise LLM platforms can produce biased customer messages, knowledge answers, job descriptions, customer-support outputs, among others [15]. The reviewed literature supports a multidimensional interpretation: reliability is dependable performance across relevant demographic and contextual variation. The interpretation, based on literature reviewed that suggests a multi-dimensional interpretation is: Reliability is the dependable performance on the relevant variation, which includes a demographic and contextual variation. This view aligns with trustworthy AI scholarship and indicates the need for more specific enterprise-level metrics (e.g., difference in refusal rates, difference in sentiment in identity terms, differences in refusal rates, sentiment associated with identity terms, retrieval quality, and factual accuracy across user groups).

There's also a change in the meaning of reliability when considering human trust evidence. The research on trust in automation shows that users' responses are not only related to the actual performance of the automation but also to task cues, to the interface design, and to their previous experience with the automation and to the risk level of the task [16, 17]. This is crucial for large language models because there might be an implicit authority signal in the generated text. Users may accept results they see on the interface as a "black box" without any uncertainty, provenance or escalation paths. Excessive warnings may also undermine appropriate reliance when the available evidence is strong. Calibration, not acceptance should form the basis of trust-aware evaluation. Useful metrics should assess whether users accept well-grounded answers, reject ungrounded answers, escalate high-risk outputs, and incorporate uncertainty labels.

There are also very few empirical bases in the literature. A large number of LLM evaluation techniques are based on static datasets, short-term tests, and benchmark leaderboards [6]. Enterprise platforms are dynamic and continuously undergoing changes in model providers, platform knowledge bases, prompts, and users' understanding of the platform, and adversarial prompting strategies. Static evaluation is thus not enough. The following should be considered as part of reliability assurance: Regression testing following model changes; retrieval-quality audits; Logging of discrepancies between model output and available verified data; sampling of user-modified production outputs; Incident analysis. There are indications of such practices in the governance and documentation literature [8] and [10] but no peer-reviewed LLM-specific enterprise studies have validated such practices at scale.

Figure 2 presents thematic coding of the reviewed journal studies across six reliability dimensions. The evidence base is uneven: model behaviour and governance are well documented, whereas deployment monitoring and user-reliance calibration are less developed. The articles were coded manually, and a study could contribute to more than one dimension [5]–[18].

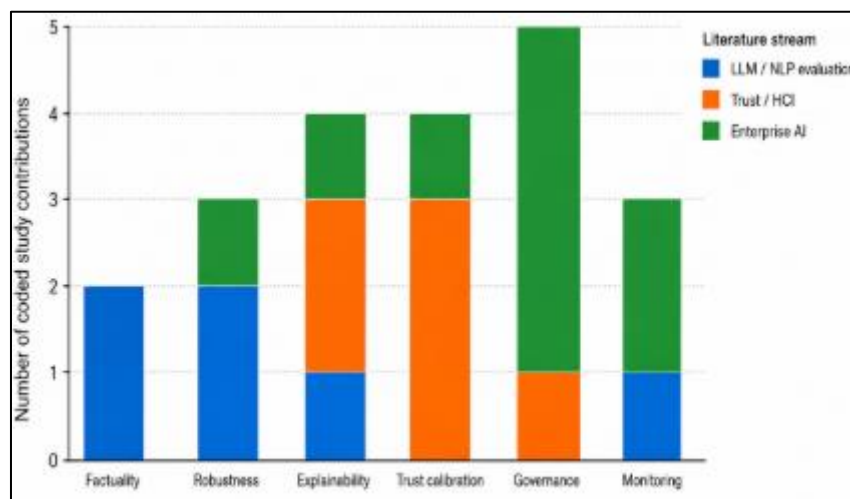


Figure 2 Evidence distribution across trust-aware enterprise LLM evaluation dimensions

Another methodological implication is the design of the platform. Enterprise LLM systems are increasingly layered, such as prompting, retrieval, tools, memory, moderation, access controls, and human review. In traditional evaluation, the model is evaluated; in trust-aware evaluation, the deployed configuration is evaluated. Retrieval-Augmented Generation is an example of this. While this can reduce hallucination, it can also introduce fresh failure modes such as missing documents, out-of-date indexes, poor chunking, retrieval of irrelevant passages, and incorrect citation alignment. A platform level framework should therefore consider the entire end-to-end process: query understanding, relevance of retrieved passages, answer generation, citation fidelity, expressing uncertainty, user action and audit record. In the literature, there is conceptual support for such integration, although few empirical studies are available in the journal literature that take all the components into account in a single study.

The biggest research gap is the absence of validated enterprise-reliability thresholds. Researchers have identified dimensions of trustworthy AI, benchmark categories, elements of documentation and elements of trust, as well as organizational capabilities [1, 3, 6, 10, 16]. But enterprise decision makers seek thresholds: acceptable hallucination rates for some tasks; acceptable citation-fidelity rates for knowledge work; thresholds for prompt perturbations; fairness-disparity thresholds; thresholds for escalation on uncertain outputs; thresholds for incidents after deployment. Without thresholds, operational limits cannot be measured, and evaluation frameworks remain descriptive. It should

be emphasized that it is not the role of a trust-aware framework to claim that a model is trustworthy in and by itself. It should contain information on the task, context, user group, risk level, evidence basis, reliability threshold, monitoring plan and decision authority.

The figure below depicts an overview of the enterprise LLM risk domains based on their business criticality level, their evaluation maturity level and the amount of literature coverage. The figure highlights domains where high business criticality coincides with relatively limited evaluation maturity, indicating areas that may require more rigorous empirical testing. The mapping is based on thematic interpretation of the reviewed literature on hallucination, benchmarking, governance, trust calibration, and fairness [5]–[18].

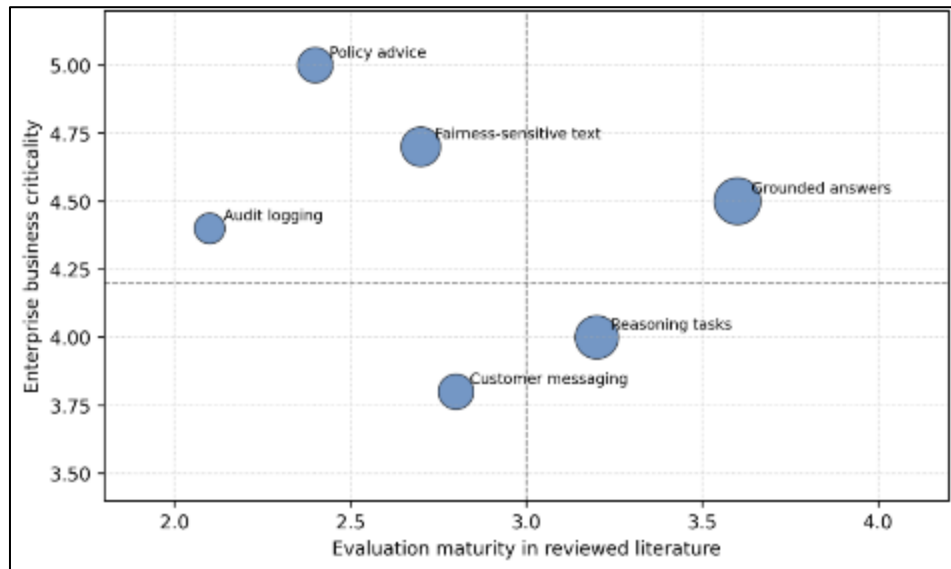


Figure 3 Integrative risk-assurance map for enterprise LLM reliability evaluation

The results of the reviewed studies confirm the layering approach to evaluation. The first layer evaluates reliability of the model output in terms of factuality, robustness, safety, fairness and task performance. The second layer assesses the reliability of the platform by analyzing retrieval quality, access control, reliability of tools, logging, monitoring and management of updates. The third layer concerns human reliance, such as trust calibration, interpretation of uncertainty, user escalation behaviour and editing patterns. The fourth layer assesses governance reliability by means of documentation, ownership, auditability, incident response and lifecycle control. These layers need to come together to give rise to trust. When enterprise risk is high, a high score in one layer does not make up for the lack of evidence in another.

Future Directions

More research is required to set task- and risk-calibrated reliability thresholds for enterprise LLM platforms. Existing literature provides dimensions of trustworthy AI, hallucination taxonomies, and benchmark categories, but it does not yet provide widely accepted operational cutoffs for enterprise deployment decisions. A more comprehensive research program would set a minimum required level of factuality, citation fidelity, robustness, and escalation for each level of enterprise risk and assess whether, in deployed platforms, these thresholds correspond to a reduction in incidents [5], [6], [10]. This would shift evaluation from broad model comparison toward release criteria for specific enterprise contexts.

The second priority is to conduct longitudinal post-deployment assessment. The reliability of an enterprise LLM system may shift during updates to the model, when new knowledge bases are added or previously available knowledge becomes inaccessible, as prompt distributions change or repeated prompts reveal new failure modes, and when users adjust their behaviour. Monitoring designs that use a combination of automated regression suites and expert review, user correction logs, retrieval audits, and incident reports should be investigated in the journal literature. Such designs would connect AI-governance and management practices with measurable platform reliability [8, 9].

The third priority is to empirically measure calibrated trust. The findings of responsible trust research indicate that communicating competence and uncertainty impacts reliance [11]; findings from trust-in-automation evidence indicate

that performance, process and task context influence reliance [16, 17]. Future research should examine how combinations of interface features such as confidence bands, evidence panels, refusal explanations, source ranking, and escalation prompts affect appropriate reliance rather than merely perceived trust.

The fourth priority is the evaluation of composite LLM systems. Typically, enterprise platforms integrate with foundation models, retrieval, tools, policies, filters, and workflow automation. Thus, reliability evaluations should assess full pipelines, including retrieval relevance, tool-call correctness, access control, citation alignment, and audit-log completeness, completeness of audit logs), full pipelines should be taken into account. This would help to align LLM-capability evaluation with explainability, fairness and human-centered AI scholarship [12], [15], [18].

Last, but not least, there is a need for interdisciplinary collaboration around measurable assurance artifacts. The demands of accountability can be clarified by legal and compliance scholarship, reliance can be measured by human-computer interaction, and governance fit and organizational capability can be assessed by information systems; adversarial tests and benchmark diagnostics can be defined by computer science. A validated evaluation framework would be most useful if it produced evidence which could be used by engineers, risk committees, auditors and enterprise users.

5. Conclusion

LLM failures are not entirely technical; therefore, trust-aware assessment frameworks remain necessary for determining whether large language models are reliable in enterprise AI platforms. These failures affect knowledge integrity, reliance, organizational accountability, and business risk. The literature shows that there is no single benchmark, explanation method, governance document, or trust measure that alone can define enterprise reliability. Alignment among evidence on model behaviour, platform architecture, human interaction, and governance controls is crucial to reliable deployment.

This review contributes to an emerging area of scholarship by connecting two lines of investigation that are often treated separately. Hallucination and LLM benchmark studies identify and explain failure modes. Trustworthy-AI and explainability studies identify assurance requirements. In the enterprise AI research context, organizational capabilities and governance shape operational reliability. Trust-in-automation studies show why greater trust is not, by itself, the appropriate goal.

Persistent gaps remain, including the absence of validated enterprise thresholds, limited evidence on longitudinal monitoring, a lack of standardized trust-calibration metrics, and insufficient integration of retrieval-augmented and tool-using LLM platforms. Compact, auditable and task-specific frameworks to turn reliability evidence into deployment decisions will be needed for future progress. Trust-aware assessment should therefore be considered a continuous process rather than a one-time enterprise-platform selection exercise.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

- [1] Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, 31(2), 447–464.
- [2] Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660.
- [3] Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Computing Surveys*, 55(2), Article 39.
- [4] Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 66(1), 111–126.
- [5] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248.

- [6] 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., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), Article 39.
- [7] Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, 33(1), Article 63.
- [8] Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3), 1433–1450.
- [9] Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management*, 58(3), Article 103434.
- [10] Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Natesan Ramamurthy, K., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier’s declarations of conformity. *IBM Journal of Research and Development*, 63(4/5), 6:1–6:13.
- [11] Liao, Q. V., & Sundar, S. S. (2022). Designing for responsible trust in AI systems: A communication perspective. *ACM Transactions on Computer-Human Interaction*, 29(4), Article 31.
- [12] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence: Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
- [13] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- [14] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), Article 93.
- [15] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), Article 115.
- [16] Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434.
- [17] Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, 58(3), 377–400.
- [18] Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504.