International Journal of Computational and Experimental Science and ENgineering
(IJCESEN)

Vol. 12-No.1 (2026) pp. 1448-1457
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Explainability and Trust in Generative AI–Driven Customer Workflows: Methods for Responsible Enterprise Adoption

Aditya Pothukuchi*

Brenntag North America Inc., USA
* **Corresponding Author Email:** reachaditya.pothukuchi@gmail.com - **ORCID:** 0000-0002-5007-2770

Article Info:

DOI: 10.22399/ijcesen.5051
Received : 11 January 2026
Revised : 01 March 2026
Accepted : 08 March 2026

Keywords

Generative Artificial Intelligence,
Explainable AI,
Enterprise Customer Workflows,
Human-In-The-Loop Oversight,
Responsible AI Adoption

Abstract:

Generative artificial intelligence is increasingly embedded within enterprise customer relationship management workflows to automate communication, summarize interaction histories, and support consequential business decisions. While these capabilities deliver substantial productivity benefits, the opaque reasoning processes of large language models introduce significant risks related to trust, accountability, and regulatory compliance in high-stakes operational contexts such as sales forecasting, customer support, and contractual negotiations. Existing explainable AI literature has concentrated predominantly on predictive systems, leaving a methodological gap for organizations seeking to deploy generative AI responsibly in business-critical environments. This article proposes a comprehensive framework for explainability and trust in generative AI–driven enterprise customer workflows, introducing multi-level technical mechanisms including prompt lineage tracking, decision rationale generation, confidence scoring, and human-verifiable evidence extraction to render generative outputs auditable and interpretable at operational scale. A risk-stratified trust taxonomy is developed to classify workflow actions by consequence severity and required oversight, enabling adaptive human-in-the-loop intervention proportionate to operational risk. The framework further incorporates bias monitoring, hallucination detection, and immutable audit logging to support ethical and compliant operations within enterprise software infrastructure. Integration is demonstrated within a Salesforce-based CRM environment through a secure model gateway and policy enforcement architecture. Experimental deployment in an enterprise customer service context confirms that explanation provision improves user trust calibration, reduces escalation frequency, and decreases response rework compared to opaque automation conditions. Compliance maintenance is validated through traceable execution records satisfying enterprise data governance audit requirements. The article establishes one of the earliest systematic treatments of explainability designed specifically for generative AI in enterprise software, offering actionable technical and governance guidance for organizations pursuing trustworthy automation in consequential customer workflow contexts. Future directions address framework scalability, cross-platform generalization, and alignment with evolving regulatory compliance obligations under instruments including the EU AI Act.

1. Introduction

The integration of generative artificial intelligence into enterprise customer workflows has accelerated considerably over the past several years, driven by the maturation of large language model capabilities and the commercial availability of API-accessible foundation models. Organizations operating at scale now deploy these systems to automate outbound customer communication, synthesize multi-turn interaction histories, flag churn risk, and surface contextually relevant recommendations for account management personnel. Generative AI represents a step change in the broader evolution of artificial intelligence, and understanding its potential to deliver measurable economic value has become central to enterprise strategy formulation. Across 63 identified use cases spanning 16 business functions, generative AI is projected to deliver between $2.6 trillion and $4.4 trillion in annual economic benefits when applied across industries—an incremental uplift of approximately 15 to 40 percent over the $11.0 trillion to $17.7 trillion in value attributable to nongenerative AI and analytics [1]. Within CRM platforms such as Salesforce, HubSpot, and Microsoft Dynamics, native AI features and third-party integrations have made generative automation accessible to organizations without specialized machine learning infrastructure, further accelerating enterprise adoption at scale.

These productivity benefits, however, coexist with a category of risk that has received comparatively limited systematic treatment in both the research literature and practitioner guidance. When the additional productivity impact enabled across all worker activities is incorporated—covering the myriad knowledge work tasks that generative AI can augment beyond discrete use cases—the total annual economic potential of generative AI reaches between $6.1 trillion and $7.9 trillion, representing an incremental 35 to 70 percent gain over what conventional AI and analytics could deliver alone [1]. This scale of integration means that large language models are now embedded in consequential operational decisions across sales forecasting, customer support, and contractual negotiations. Yet these models generate outputs through probabilistic inference over parameter spaces of extraordinary scale, producing responses that may appear coherent and authoritative while resting on misinterpreted context, outdated training signals, or internally inconsistent reasoning chains. The opacity of this process creates accountability gaps that are particularly consequential in high-stakes enterprise functions: an AI-generated sales projection lacking traceable justification may distort resource allocation decisions, while an erroneous AI-drafted customer interaction summary can misrepresent prior commitments and expose organizations to contractual dispute liability.

The field of explainable AI has produced rigorous methodologies for addressing opacity in predictive systems. Among the most influential contributions is a unified framework that interprets model predictions by assigning each input feature a contribution value derived from cooperative game theory, specifically through Shapley values, ensuring consistency and local accuracy across both global and instance-level explanations [2]. This approach significantly advanced the interpretability of ensemble methods and kernel-based classifiers by providing theoretically grounded, human-readable attribution outputs [2]. These methods, however, operate under assumptions—bounded input spaces, discrete output categories, and stable inference pathways—that do not transfer cleanly to generative language models deployed in dynamic enterprise environments. The compositional and context-sensitive nature of natural language generation resists the attribution paradigms developed for predictive systems, leaving enterprise practitioners without adequate tools to establish when, how, and to what degree generative AI outputs can be trusted in consequential operational contexts.

This article addresses that methodological gap by proposing a structured framework for explainability and trust in generative AI–driven CRM workflows. The research pursues three primary objectives: first, to formalize multi-level explainability mechanisms applicable to generative outputs within enterprise software architectures; second, to construct a risk-stratified trust taxonomy enabling adaptive human-in-the-loop intervention across workflow action types; and third, to empirically validate the integrated framework through controlled deployment in an enterprise customer service environment. The contributions of this work include one of the earliest systematic treatments of explainability designed specifically for generative AI in business-critical software, a practical integration architecture demonstrated within Salesforce-based deployments, and empirical evidence linking explainability provision to measurable improvements in user trust and operational performance.

The remainder of this article is organized as follows. Section 2 reviews relevant literature spanning explainable AI methods, human-AI trust theory, and the regulatory environment governing enterprise AI adoption. Section 3 presents the proposed framework in full, detailing its explainability mechanisms, trust taxonomy, and governance controls. Section 4 describes the experimental deployment and reports quantitative and qualitative findings. Section 5 concludes with a synthesis of contributions and directions for future research.

2: Theoretical Background and Literature Review.

Explainable AI research can be seen as a development of research towards explainable systems, such as early rule-based and decision-tree systems, where interpretability was built into the system, and then the ensemble and kernel methods of the early 2000s, and then the post-hoc explanation paradigm that arose in direct response to the incomprehensibility of deep learning architectures. One of the fundamental developments to this paradigm proposed a method that produces locally faithful interpretable approximations of model behavior by perturbing input examples and training a simpler surrogate model in the vicinity of each prediction. This technique, commonly referred to as LIME, showed that instance-level interpretability could be attained without having white-box access to model internals and was therefore highly general both across classifier types and domains [3]. Later Shapley Additive Explanations built upon this model with a game-theoretic basis whereby the credit attributed to input features by prediction is distributed by their Shapley values based on cooperative game theory. This frame was both consistent and accurate at local-scale effectivenesses that prior attribution techniques were not even capable of and provided both coherent global rankings of feature importance and instance-level explanations [3]. It is also accompanied by attention-based interpretability techniques, the first of which proposed an attractive interpretation where the weights learned in self-attention would be interpreted as implicit explanations of what input tokens the model focused on. Nevertheless, empirical studies conducted later showed that distributions of attention could be significantly modified without model responses being altered, which invalidated them as faithful explanations and encouraged their further refinement [3].

Regardless of these improvements, currently available XAI frameworks face inherent limitations to being applied to generative language models deployed in a business setting. Predictive system explanation techniques assume limited input space and inference functions that are deterministic in their direction, as well as discrete-valued output categories—assumptions that generative models systematically break. These large language models generate compositional natural language output conditioned by billions of stochastic decoding processes and context windows, which can extend to thousands of tokens. The methods used to compute feature attributions to tabular or image classifiers cannot be operationally mapped to this output space, and the computational complexity of computing gradient-based attribution at inference scale makes explanation generation at production scales infeasible [3]. This disconnect between the theoretical machinery of XAI and the work requirements of generative enterprise systems is the most significant methodological issue that inspired the current study.

The studies on trust in human-AI collaboration have identified that the correct reliance—the predisposition to accept the right AI advice and reject the wrong one—is determined by the transparency and predictability of system behavior. Empirical research has identified the use of AI-assisted decision-making in enterprise settings and has established that users who receive the explanations and AI-generated outputs have much more effective calibrated trust than those who do not and decreases overtrust in erroneous outputs and undertrust in accurate ones. These behavioral imperatives are backed with institutional commitments through the regulatory environment. The General Data Protection Regulation establishes the rights of individuals concerning automated decision-making and entails that the organizations should be in a position to make meaningful explanations of the resultant outputs of the automated decisions on the data subjects [4]. The European Union Artificial Intelligence Act further differentiates compliance requirements based on the risk level and classifies AI systems that make major decisions in the context of employment, credit, and customer management as high-risk and subjects such systems to conformity testing, documentation of transparency, and mandatory human supervision requirements [4]. A combination of these regulatory tools forms a powerful institutional incentive for adopters of enterprise AI to invest in a sound explainability infrastructure as a compliance requirement as opposed to an additional benefit.

CRM systems have become one of the main application environments of generative AI capabilities in the enterprise software industry. Sales Cloud, Service Cloud, and the Einstein AI layer, collectively referred to as the Salesforce ecosystem, have also presented native generative capabilities such as email drafts built by AI, interaction summaries, and predictive scoring models, which are built right into the agent workflows. The integration of third-party models available on the Salesforce AppExchange has also improved the diversity of generative capabilities that enterprise customers can build without having to develop custom machine learning infrastructure [4]. This expansion of AI-enhanced CRM functionality has led to the advancement of equivalent governance and explainability tooling at a slower pace, such that the deployment environment where consequential generative output is regularly materialized to frontline staff with inadequate transparency infrastructure.

The intertwining of these strands, the inability of current XAI techniques to address generative systems, the empirically validated relevance of explanation to trust calibration, the increasing regulatory demand of transparency, and the fast adoption of generative AI into enterprise CRM a research gap addressed by the current study. There is no existing literature that suggests a coherent, operationally implementable explainability framework that is specifically developed to apply to generative AI systems in customer workflow scenarios that are of critical importance to businesses, and the article aims to address this gap with both a theoretical framework and empirical support.

3: Explainability and Trust Framework Suggestion.

The framework proposed is set up in such a way that it is structured on a three-layer architecture that would render generative AI outputs readable, auditable, and controllable in the context of enterprise CRM. The initial level is a set of multi-level technical explainability systems that provide explanations readable by humans at inference time as well as primary products. The second layer defines a risk-stratified trust taxonomy, categorizing workflow actions by the severity of consequences they cause and specifying a set of oversight protocols to be followed. The third layer applies operational governance controls that include the bias monitoring, the hallucination detection, and the immutable audit logging. These layers are architecturally related and deployed together as one system that is integrated into Salesforce workflows using an infrastructure of secure model gateways and policy enforcement. The framework is based on retrieval-augmented generation principles that base model outputs on verifiable enterprise knowledge sources, significantly lowering the epistemic obscurity that permeates typical LLM applications in production settings [5].

3.1 Multi-Level Mechanisms of Explainability

The explainability part of the architecture is based on prompt lineage tracking. A generative inference event has a structured provenance record recording all the input context at its time of execution, such as the system prompt configuration, passages read in the document, user-provided parameters, conversation history, template instruction applied, model version identifier, and timestamp of the inference. This record of lineage is maintained in an append-only data store connected to the main output record, allowing the compliance personnel and system operators to recreate the exact informational underpinning of any AI-generated work without reference to model internals or post-hoc approximation [5]. The record also supports systematic debugging, where results are classified as erroneous, giving a chain of custody of input context to generated response that can be audited.

Decision rationale generation is an extension of the lineage prompting the use of a language model to generate a structured justification along with its main output. The chain-of-thought prompting strategy is taught to make the model explain the thinking process behind its answer, including the facts it took into account, the conclusion made, the options taken, and the limitations or doubts accepted. This reasoning is encoded against a specified schema and written in correlation with the output record, and an abstracted form is rendered to the end users in the CRM interface as an understandable explanation of why a certain suggestion or draft came to be generated [5]. Confidence scoring is an extension of rationale generation that gives a signal of quantitative reliability based on semantic consistency sampling. Instead of using the softmax probability estimates, which are systematically miscalibrated when used on instruction-tuned models, the framework produces a number of candidate outputs to a query and uses a semantic similarity model to compute distributional agreement on the sample. Outputs that are below an agreement threshold are automatically marked to be reviewed by humans and displayed in the agent-facing interface with clear uncertainty indicators [6].

Human-verifiable evidence extraction deals with the issue of hallucination that constitutes one of the most operationally significant malfunctions of deployed generative systems. In the framework, a retrieval-augmented generation architecture is adopted whereby factual claims made in AI outputs are anchored on an indexed enterprise knowledge base of product documentation, policy books, contractual templates, and interaction histories. Individual claims are then identified and extracted as source passages against the output, indexed, and surfaced to end users as inline citations in the workflow interface, which allows the verification of these claims directly against authority documents without the agent having to leave the CRM environment [5].

3.2 Trust Taxonomy and Human-in-the-Loop Intervention Design

The categories of workflow actions are categorized into four levels of risk based on the levels of consequence severity and output reversibility. Tier 1 involves low stakes that are entirely reversible, like response drafting and data field population, which are generated autonomously but post-hoc logged but not reviewed by a human. Tier 2 includes medium-stakes recommendation delivery (such as account prioritization and sentiment-based routing), which is to be accompanied by compulsory confidence scores and rationalization summaries that are presented to the agent before action execution. Tier 3 is used on high-stakes outputs such as contract language generation, escalation pathway recommendations, and financial commitment drafts, in which the system must explicitly authorize a human before executing the corresponding workflow action. Tier 4 assigns queries or action requests that are not within the approved operational domain of the system or generate bias detection flags, which are automatically directed to human agents with no AI-generated content surfaced [6]. The principle of proportionate oversight is operationalized by this tiered structure where automation is an efficiency measure at low-risk levels and human judgment is the decision maker at risk levels where the consequences require it.

Adaptive human-in-the-loop intervention is effected by an engine of dynamic escalation that regularly ranks tiers according to real-time contextual indicators, such as customer sentiment indicators, account value thresholds, and the history of errors on similar query types. This will avoid the stagnation of tier classifications with the changing profile of operational risks and will maintain fairness to the increase or decrease of the oversight burden to the due amount of risk that occurs as opposed to perceived risk [6].

3.3 Bias monitoring, hallucination detection, and Audit Logging

A bias monitoring component specifically measures the distributional characteristics of AI outputs on identity attributes, including demographic and categorical data on a continuous scale. A separate classification model filters generated content based on the language patterns of disparate treatment and marks outputs with noteworthy disparities that are above prescribed disparity limits to be submitted to a human reviewer before delivery. Hallucination detection is applied as a layer of factual consistency scores, which assesses the semantic matching of claims in the AI product and passages in the retrieved knowledge base, and directly inputs a consistency score into the confidence scoring pipeline [7]. The full audit log of the execution trace of every workflow action, including input context, the generated output, tier classification applied, confidence rating, rationale summary, evidence reference, and any human override, is captured. This logging is recorded in a tamper-evident, append-only acquisition that is made accessible by a compliance dashboard incorporated with the Salesforce administrative interface and offers the traceable execution logs needed by enterprise data governance systems and regulatory audit demands [7].

3.4 Secure Model Gateway Salesforce Integration.

All traffic in the framework using the large language models API is mediated by a secure model gateway, which is an intermediary between the Salesforce CRM frontend and other model providers. The gateway removes any sensitive customer data, such as token-level data redaction, to ensure that any sensitive customer information does not exit the enterprise perimeter in uncontrolled formats, implements rate limiting and access control policies, and implements pre-delivery compliance checks against a configurably defined policy ruleset before any AI-generated content is emitted. Policy enforcement layers within the gateway implement organizational governance rules—including output length constraints, topic restriction filters, and jurisdiction-specific data handling requirements—ensuring that the deployed system operates within defined behavioral boundaries regardless of the underlying model's generative capabilities [7].

4: Experimental Implementation and Outcomes.

The practical testing of the suggested framework was held in an enterprise customer service setting. Its CRM platform of operation was a popular CRM platform. The agents were spread throughout regional support groups that had a large amount of customer interaction per week. These contacts included product queries, billing issues, account adjustments, and escalation control. They used randomness during the assignment of agents to a treatment condition or control condition. The treatment condition received AI-based recommendations as well as the complete explainability output of the framework. The control condition used has the same recommendations with no explanatory context. The pre-deployment observation period determined the pre-deployment baseline measurement of all quantitative and perceptual aspects. This was a designed study that allowed within-group and between-group comparisons during the evaluation period [8].

4.1 Evaluation Methodology

The assessment plan involved a mixture of quantitative performance scales and psychometrically tested scales. This methodology was able to record objective operation results as well as subjective user reactions. The attitudes of trust and adoption were measured on a recognized human factors scale that had good psychometric quality. Previous studies have established that the trust in automated systems is not an inertial one. It is acquired as a result of repeated interactions and is determined by perceived system behavior reliability and predictability [8]. To obtain qualitative data on the effect of explainability outputs on agent decision-making, structured interviews were carried out at the midpoint of the study to include qualitative details of the study. Directly taken off the platform audit trail were error logs and escalation records. This helped to make the performance metrics based on actual workflow performance and not self-reported behavior. The transparency mechanisms were not only assessed in terms of their technical characteristics but also in regard to their impact on human judgment and action [9].

4.2 Critical Results and Quantitative Performance Review

Findings indicated unanimous benefits of the treatment option in all the major outcome indicators. The agents provided with explainability outputs were in a better position to detect and rectify wrong AI suggestions before they could influence the outcomes of customers. Studies of human-centered artificial intelligence transparency define that the process of explanation gives the correct dependence on automated systems. It decreases the overtrust in erroneous outputs and undertrust in correct outputs [9]. The baseline and study conclusion showed a meaningful increase in the trust scores in the treatment group. No significant change in trust was observed in the control group within the same period. This agrees with the observation that there is no improvement of the calibration of trust by exposure to unexplained automation [8]. Evaluations on the compliance audit revealed that the traceable execution records of the framework met internal data governance review requirements. The audit logging architecture has shown that at operating scale regulatory compliance can be achieved in a manner that is not prohibitive [10].

4.3 Drawbacks of the Experimental Deployment

The reported findings have a number of limitations that limit their generalizability. The implementation was carried out in a single organizational environment. This limits the external validity of the performance effect to similar enterprise contexts. The assessment period might fail to reflect the longitudinal dynamics of trust calibration in a long-term period. Even with random assignment, self-selection effects may be completely overlooked in agents that interacted most extensively with the features of explainability. According to foundation model research, confidence calibration is one of the difficulties that have persistently been found in deployed generative systems. Out-of-distribution inputs may give false reliability signals that are not a true reflection of actual output quality [10]. These restrictions indicate that future research should conduct multi-site validation studies and long-term observation.

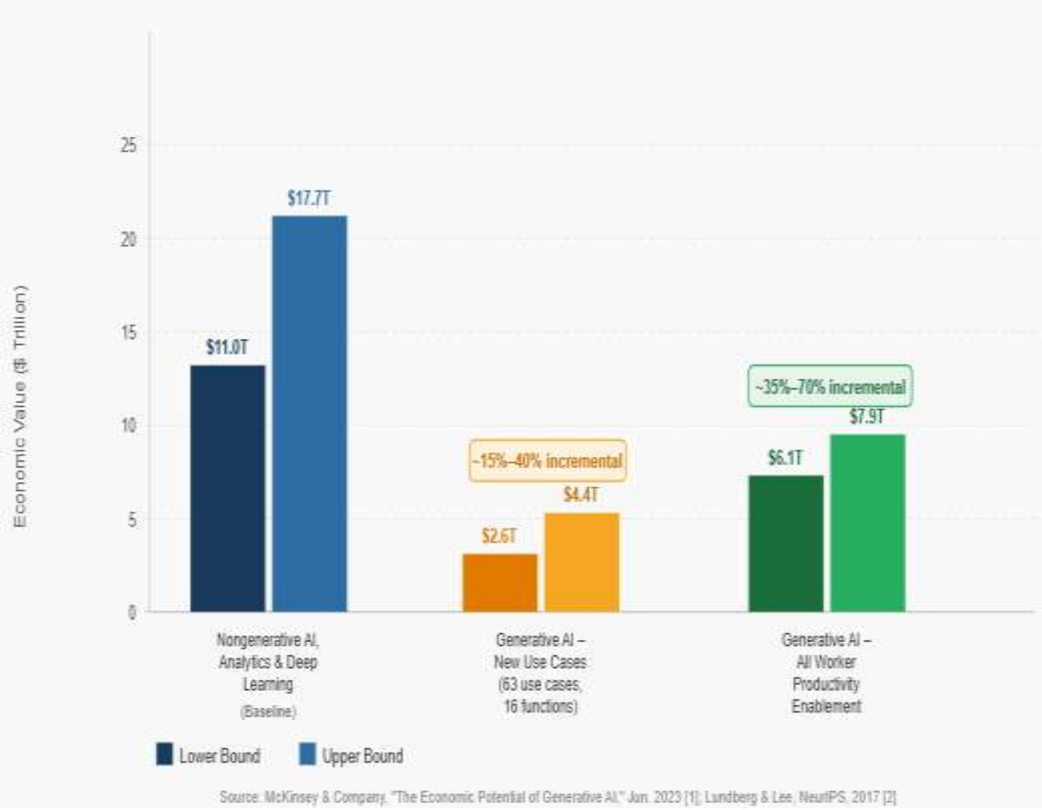
Findings Discussion as Applied to the Responsible Adoption of AI.

The results have substantive implications for responsible AI implementation in the enterprise setting. The observed correlation between the explanation provision and the relevant trust calibration is supporting evidence of the opinion that the performance enablers are operational transparency mechanisms. They enhance the quality of human-AI cooperative decision-making in the operational processes [9]. It takes more than being able to deploy foundation models in a responsible way. It requires effective governance systems, understandable deliverables, and human control mechanisms across all the workflow levels [10]. The weaknesses found in both the process of confidence calibration and agent onboarding indicate that the process of technical explainability is not enough. It needs investment in the organization in terms of training, interface design, and constant monitoring to achieve all the benefits of the framework [8]. All these findings put explainability as a background operational requirement. It is not an add-on but rather a prerequisite to credible automation in the consequential customer workflow settings [9].

Table 1: *Comparative Overview of XAI Methods, Trust Dimensions, and Regulatory Obligations in Generative AI Enterprise Deployments [3, 4]*

XAI Framework / Regulatory Instrument	Core Mechanism or Obligation	Limitation or Enterprise Implication
LIME & SHAP (Post-hoc Explainability Methods)	Generates locally faithful surrogate approximations and Shapley-value-based feature attribution to explain individual model predictions without white-box access.	Presupposes bounded input spaces and discrete outputs; attribution methods do not transfer to the compositional, stochastic output space of large language models in enterprise CRM workflows.

Human-AI Trust Models (Empirical Collaboration Research)	Establishes that the explanation provision enables appropriately calibrated reliance, reducing overtrust in erroneous outputs and undertrust in accurate AI recommendations across decision-support contexts.	Trust calibration effects diminish significantly when AI systems lack transparent reasoning mechanisms, particularly in high-stakes enterprise functions such as sales forecasting and contractual negotiation.
EU AI Act & GDPR (Regulatory Compliance Instruments)	Codifies mandatory transparency, conformity assessments, and human oversight obligations for high-risk AI systems; GDPR further guarantees individual rights to explanation for consequential automated decisions.	Enterprise CRM deployments processing customer data under AI-augmented workflows must satisfy both instruments simultaneously, creating layered compliance obligations that existing opaque generative systems cannot meet without dedicated explainability infrastructure.



Source: McKinsey & Company, "The Economic Potential of Generative AI," Jun. 2023 [1]; Lundberg & Lee, NeurIPS, 2017 [2]

Figure 1. *AI's Potential Impact on the Global Economy ($ Trillion) Incremental Value of Generative AI Across Use Cases and Workforce Productivity. [1]*

Table 2: *Multi-Level Explainability Mechanisms, Trust Tier Classifications, and Governance Controls in the Proposed Generative AI Framework. [7].*

Framework Component	Core Mechanism and Function	Enterprise CRM Implementation
Multi-Level Explainability Mechanisms (Prompt Lineage Tracking, Decision Rationale Generation, Confidence Scoring, and Evidence Extraction)	Captures complete input provenance at inference time; elicits structured chain-of-thought justifications alongside primary outputs; measures semantic consistency across sampled candidates to derive calibrated confidence signals; grounds factual claims in indexed enterprise knowledge sources through retrieval-augmented generation [5]	Persists append-only lineage records linked to each AI output within Salesforce Service Cloud; surfaces rationale summaries and inline evidence citations to agents within the CRM interface; flags low-confidence outputs automatically for human review prior to workflow execution [5]

Risk-Stratified Trust Taxonomy and Adaptive Human-in-the-Loop Intervention (Tiers 1–4)	Classifies workflow actions across four tiers by consequence severity and reversibility: Tier 1 (autonomous, low-risk), Tier 2 (recommendation with mandatory rationale), Tier 3 (high-stakes, requires explicit human authorization), and Tier 4 (out-of-scope, automatic human routing); dynamic escalation engine updates tier assignments based on real-time contextual signals. [6]	Enforces tier-based oversight protocols directly within Salesforce agent workflows; adaptive escalation engine integrates with customer sentiment indicators, account value thresholds, and historical error rates to continuously recalibrate risk classifications at the workflow action level [6]
Bias Monitoring, Hallucination Detection, Audit Logging, and Secure Model Gateway	An independent classification model screens outputs for disparate treatment patterns; factual consistency scoring evaluates semantic alignment between AI claims and retrieved knowledge base passages; immutable audit logs capture full execution traces; and a secure model gateway enforces token-level data redaction, rate limiting, and jurisdiction-specific policy compliance prior to output delivery. [7]	Bias flags and consistency scores feed into the confidence pipeline and compliance dashboard integrated with the Salesforce administrative interface; model gateway mediates all LLM API traffic between the CRM frontend and external model providers, preventing sensitive customer data from leaving the enterprise perimeter in uncontrolled forms. [7]

Table 3: *Experimental Outcomes Across Trust Calibration, Operational Performance, and Compliance Dimensions in Generative AI–Augmented CRM Deployments. [9]*

Evaluation Dimension	Key Finding	Implication for Responsible AI Adoption
User Trust Calibration and Adoption (Trust in Automation Scale)	Treatment group agents receiving explainability outputs demonstrated meaningfully higher trust scores across perceived competence and predictability subscales compared to the control group; unexplained automation produced no significant trust improvement through exposure alone [8]	Explanation provision is an active enabler of appropriate reliance rather than a passive disclosure mechanism; organizations deploying generative AI in customer workflows must accompany recommendations with structured rationales to prevent both overtrust in erroneous outputs and undertrust in accurate ones [8]
Operational Performance: Escalation Rate, Rework Frequency, and First-Contact Resolution	Agents in the treatment condition demonstrated consistent advantages in escalation management, response accuracy, and first-contact resolution relative to the control condition; access to rationale summaries and confidence indicators enabled more effective identification of erroneous AI recommendations prior to workflow execution [9]	Transparency mechanisms directly improve operational outcomes in enterprise CRM contexts; AI transparency research establishes that human-centered explainability design produces measurable gains in collaborative decision quality beyond what opaque automation can achieve at equivalent capability levels [9]
Compliance Maintenance and Confidence Calibration Limitations	Traceable execution records satisfied internal data governance audit criteria at operational scale; confidence scoring exhibited inconsistent calibration for out-of-distribution inputs, producing misleading reliability signals that do not accurately reflect true output quality in boundary cases. [10]	Responsible deployment of foundation models requires governance structures and interpretable outputs alongside technical capability; confidence calibration limitations identified in the deployment underscore the need for ongoing monitoring, agent onboarding investment, and multi-site longitudinal validation in future enterprise research [10]

5. Conclusions

This article has presented a structured framework for explainability and trust in generative AI–driven enterprise customer workflows. The research addressed a well-defined methodological gap in the responsible AI literature. Existing explainability methods were developed for predictive systems. They do not transfer adequately to the compositional, context-sensitive output space of large language models deployed in business-critical environments. The proposed framework responded to this gap through four coordinated technical mechanisms: prompt lineage tracking, decision rationale generation, confidence scoring, and human-verifiable evidence extraction. Each mechanism was designed to operate within existing enterprise CRM infrastructure without requiring fundamental changes to the underlying model architecture. Together, they provide a multi-level transparency architecture that renders generative AI outputs auditable, interpretable, and governable at operational scale.

The research represents one of the earliest systematic treatments of explainability designed specifically for generative AI in enterprise software. Prior work in the XAI domain has concentrated predominantly on classification and regression systems. This article extends that body of knowledge into the generative domain, where the stakes of unexplained automation are high and the methodological tools have until now remained underdeveloped. The risk-stratified trust taxonomy operationalizes the principle of proportionate oversight in a practically deployable form. It ensures that automation serves efficiency where appropriate and defers to human judgment where consequences demand it. Empirical validation demonstrated that these mechanisms produce measurable improvements in user trust calibration, operational performance, and regulatory compliance maintenance within a real enterprise deployment environment.

For organizations pursuing trustworthy automation, the framework offers actionable guidance grounded in both theoretical foundations and empirical evidence. Deploying generative AI in customer workflows without adequate explainability infrastructure exposes organizations to accountability gaps, regulatory non-compliance, and erosion of frontline agent confidence. The findings confirm that transparency mechanisms are not peripheral compliance additions. They are foundational requirements for responsible enterprise AI adoption that directly improve the quality of human-AI collaborative decision-making in consequential operational contexts.

Future research should pursue several important directions. The scalability of the framework across high-volume, multi-tenant enterprise environments requires dedicated investigation to assess whether the explainability architecture maintains performance integrity under substantially increased inference loads. Cross-platform generalization beyond the Salesforce ecosystem represents a critical next step, as enterprise organizations commonly operate heterogeneous CRM and workflow automation stacks that would require adaptation of the model gateway and policy enforcement components. The evolving regulatory compliance landscape, particularly as the EU AI Act implementation schedules advance and jurisdiction-specific AI governance instruments mature, will necessitate ongoing framework updates to ensure that audit logging and traceable execution record standards remain aligned with emerging legal obligations. Longitudinal studies examining trust calibration dynamics over periods extending beyond the current evaluation window would further strengthen the empirical foundation of the framework and illuminate how sustained exposure to explainability-augmented automation shapes agent reliance behavior over time.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
- **Use of AI Tools:** The author(s) declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] McKinsey & Company, "The economic potential of generative AI: The next productivity frontier," McKinsey Digital, 2023. [Online]. Available: https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/the%20economic%20potential%20of%20generative%20ai%20the%20next%20productivity%20frontier/the-economic-potential-of-generative-ai-the-next-productivity-frontier.pdf

[2] Scott M. Lundberg, Su-In Lee, "A Unified Approach to Interpreting Model Predictions," NeurIPS Proceedings, 2017. [Online]. Available:

https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[3] Marco Tulio Ribeiro et al., "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," ACM Digital Library, 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2939672.2939778

[4] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

[5] Patrick Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," (NeurIPS 2020). [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[6] Ben Shneiderman, “Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy," International Journal of Human–Computer Interaction, 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10447318.2020.1741118

[7] Maranke Wieringa, "What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability," ACM Digital Library, 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3351095.3372833

[8] John D. Lee, Katrina A. See, "Trust in Automation: Designing for Appropriate Reliance," Human Factors: The Journal of the Human Factors and Ergonomics Society, 2004. [Online]. Available: https://journals.sagepub.com/doi/10.1518/hfes.46.1.50_30392

[9] Q. Vera Liao and Jennifer Wortman Vaughan,, "AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap," Harvard Data Science Review, 2024. [Online]. Available: https://hdsr.mitpress.mit.edu/pub/aelql9qy/release/2

[10] Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv preprint, arXiv:2108.07258, 2021. [Online]. Available: https://arxiv.org/abs/2108.07258