

Graph-Oriented Artificial Intelligence: An Integrated GNN–RAG–KG–LLM Architecture for Developing Trustworthy Enterprise Artificial Intelligence

Abolfazl Ansary^{1,*}, Fereshteh Lotfi Zadeh²



¹ MA Student, Department of E-Business Management, SR.C., Islamic Azad University, Tehran, Iran; 

² Ph.D., Department of Business Management, SR.C., Islamic Azad University, Tehran, Iran; 

* Correspondence: afiansary2010@gmail.com

Citation: Ansary, A., & Lotfi Zadeh, F. (2026). Graph-Oriented Artificial Intelligence: An Integrated GNN–RAG–KG–LLM Architecture for Developing Trustworthy Enterprise Artificial Intelligence. *Business, Marketing, and Finance Open*, 3(4), 1–22.

Received: 11 February 2026

Revised: 08 July 2026

Accepted: 16 July 2026

Published: 30 July 2026



Copyright: © 2026 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Abstract: This study aimed to design and empirically evaluate an integrated graph-oriented architecture combining Graph Neural Networks (GNNs), Retrieval-Augmented Generation (RAG), Knowledge Graphs (KGs), and Large Language Models (LLMs) to improve the trustworthiness of enterprise artificial intelligence. This applied, developmental, and quantitative evaluation study was conducted in Tehran, Iran, in 2026 with 48 experts selected purposively from technology companies, universities, financial institutions, telecommunications organizations, and enterprise information technology departments. The proposed architecture integrated an enterprise knowledge graph, GNN-based relational representation learning, RAG-based evidence retrieval, and LLM-based natural-language generation. Its performance was compared with LLM-only, LLM + RAG, and LLM + RAG + KG configurations. Data were collected through standardized benchmark queries, automated computational metrics, provenance logs, and a researcher-developed expert questionnaire. Repeated-measures analysis of variance, pairwise comparisons, effect-size estimation, correlation analysis, and ablation analysis were used. Significant differences were observed among the four architectures for answer correctness, $F = 126.41$, $p < 0.001$, partial $\eta^2 = 0.729$; context relevance, $F = 94.83$, $p < 0.001$, partial $\eta^2 = 0.669$; faithfulness, $F = 139.27$, $p < 0.001$, partial $\eta^2 = 0.748$; groundedness, $F = 153.62$, $p < 0.001$, partial $\eta^2 = 0.766$; evidence traceability, $F = 241.53$, $p < 0.001$, partial $\eta^2 = 0.837$; hallucination rate, $F = 172.84$, $p < 0.001$, partial $\eta^2 = 0.786$; and expert-rated trustworthiness, $F = 118.76$, $p < 0.001$, partial $\eta^2 = 0.717$. The integrated architecture significantly outperformed all reduced configurations. Expert trustworthiness was positively correlated with groundedness ($r = 0.78$), faithfulness ($r = 0.74$), evidence traceability ($r = 0.81$), and correctness ($r = 0.72$), and negatively correlated with hallucination rate ($r = -0.76$), all $p < 0.001$. Integrating GNN, RAG, KG, and LLM components produced a more accurate, grounded, traceable, and trustworthy enterprise AI system while substantially reducing hallucination, supporting the value of graph-oriented architectures for reliable organizational AI applications.

Keywords: Graph Neural Networks; Retrieval-Augmented Generation; Knowledge Graph; Large Language Models; Enterprise Artificial Intelligence; Trustworthy AI

1. Introduction

Artificial intelligence has moved from an experimental digital technology to a strategic organizational capability with direct implications for productivity, decision-making, innovation, resource allocation, customer interaction, knowledge management, and long-term competitiveness. Enterprises are increasingly embedding artificial intelligence into core and support functions in order to automate routine processes, discover patterns in large-scale data, improve forecasting, personalize services, strengthen operational control, and accelerate managerial responses to changing environmental conditions. This transformation is no longer limited to technology-intensive corporations; small and medium-sized enterprises, microenterprises, public

organizations, industrial firms, and service-oriented businesses are also exploring artificial intelligence as a means of strengthening organizational flexibility and creating new sources of value. Research on enterprise transformation has increasingly emphasized that artificial intelligence can alter business-process architecture by changing how information is collected, processed, interpreted, and converted into operational action [1]. At the same time, the integration of artificial intelligence into digital business ecosystems is associated with broader forms of digital intelligence empowerment, improved productive capacity, and spillover effects that may influence organizational and regional development [2]. These developments have positioned artificial intelligence not merely as an automation technology but as an infrastructural layer through which organizational knowledge, digital systems, human expertise, and analytical capabilities can be coordinated.

The organizational value of artificial intelligence is particularly evident in settings where enterprises must process large volumes of heterogeneous information under conditions of uncertainty and time pressure. Artificial intelligence-enabled systems can assist with accounting, marketing, supply-chain management, human resources, customer relationship management, enterprise resource planning, and executive decision support. For example, AI-supported accounting information systems can improve information processing and contribute to digital business performance when appropriately integrated with organizational strategies [3]. In marketing and branding, artificial intelligence provides opportunities for behavioral analysis, automated segmentation, content generation, recommendation, targeting, and customer insight development, thereby reshaping how enterprises communicate with markets and manage brand relationships [4]. Systematic research in electronic commerce similarly shows that artificial intelligence is becoming an important component of marketing-strategy transformation, particularly in areas such as personalization, predictive analytics, digital engagement, and automated customer interaction [5]. In customer relationship management, AI-based techniques have been associated with improved analytical capabilities, personalization, customer prediction, and automated service delivery, although future enterprise applications require stronger integration between intelligent algorithms and organizational information infrastructures [6]. These developments collectively demonstrate that enterprise artificial intelligence is increasingly expected to operate across multiple information sources and functional systems rather than within isolated predictive applications.

The growing use of artificial intelligence in enterprise resource planning provides a particularly clear example of this transition toward embedded intelligent infrastructure. Modern enterprise systems increasingly incorporate machine learning, natural language processing, intelligent automation, and generative functions into financial management, procurement, logistics, planning, reporting, and operational workflows. AI-enabled ERP environments can support automated exception detection, forecasting, document processing, intelligent recommendations, and workflow optimization, thereby extending traditional transaction-processing systems into more adaptive and analytically capable platforms [7]. However, the increasing integration of artificial intelligence with enterprise systems also creates a fundamental architectural challenge. Organizational information is rarely stored in a single structured database. Instead, it is distributed across relational databases, reports, policy documents, emails, procedural manuals, enterprise applications, customer records, data warehouses, transactional systems, and domain-specific repositories. These information resources differ in format, semantics, authority, temporality, and relational structure. Consequently, an enterprise AI system must do more than generate fluent language or identify statistical patterns. It must retrieve contextually relevant knowledge, distinguish authoritative from obsolete information, recognize relationships among organizational entities, maintain consistency across heterogeneous sources, and provide evidence that allows its outputs to be audited.

This requirement has become more important with the rapid diffusion of large language models. LLMs provide powerful natural language understanding and generation capabilities and can support tasks such as summarization, question answering, content generation, classification, reasoning support, and human–computer interaction. Nevertheless, conventional LLMs rely heavily on parametric knowledge learned during pretraining and may produce information that is outdated, insufficiently grounded in organizational data, or unsupported by verifiable evidence. In enterprise environments, such limitations are consequential because generated outputs may affect managerial decisions, compliance processes, financial operations, customer interactions, internal controls, or strategic planning. A fluent but incorrect answer may therefore create more risk than an explicitly uncertain answer. The adoption of generative artificial intelligence in small and medium-sized enterprises has been associated with potential improvements in organizational performance, but these benefits depend on whether firms can integrate generative systems effectively with organizational processes and information resources [8]. More broadly, the acceptance and adoption of AI in organizational settings depends not only on perceived usefulness but also on factors related to technological readiness, trust, compatibility, managerial support, and perceived risk [9]. These considerations indicate that enterprise deployment requires architectures that can combine generative capabilities with mechanisms for grounding, verification, transparency, and organizational control.

The concept of trustworthy artificial intelligence has consequently become central to enterprise adoption. Trustworthiness in this context extends beyond predictive accuracy and includes reliability, explainability, transparency, robustness, traceability, security, accountability, and the ability to limit unsupported outputs. In operational environments, an intelligent system must provide not only an answer but also sufficient evidence for users to understand why the answer should be considered credible. This concern becomes particularly important in sectors characterized by complex operational risks. Research in the oil and gas industry, for example, has identified multiple risks associated with implementing and using AI, including technical uncertainty, data-quality problems, cybersecurity exposure, organizational dependence on automated systems, and difficulties associated with managing AI-generated decisions [10]. Such risks are not confined to industrial environments. Any enterprise that uses AI for high-impact functions must address the possibility of inaccurate recommendations, opaque reasoning, inconsistent outputs, biased information processing, and weak auditability. Consequently, enterprise artificial intelligence should be designed around mechanisms that improve not only computational performance but also the evidentiary quality and inspectability of model outputs.

Retrieval-Augmented Generation has emerged as one of the most important approaches for addressing the grounding limitations of large language models. Rather than relying exclusively on the parametric memory of an LLM, RAG systems retrieve external documents or data fragments relevant to a user query and supply them as context during generation. This architecture can reduce dependence on static pretrained knowledge and allows enterprise systems to incorporate current organizational information without retraining the entire language model. For organizations, this is particularly attractive because internal policies, procedures, product documentation, regulatory requirements, customer information, and operational knowledge change continuously. RAG can therefore provide a bridge between large language models and enterprise knowledge repositories. However, conventional RAG generally depends on vector similarity and document-chunk retrieval. Although this approach is effective for many semantic-search tasks, it can be insufficient when enterprise questions require understanding relationships among entities, tracing multi-step dependencies, identifying causal or hierarchical structures, or integrating evidence distributed across several connected organizational objects. A retrieved document may contain relevant words while failing to reveal the structural relationship required for reasoning.

Knowledge graphs provide a complementary mechanism for representing such relationships explicitly. In a knowledge graph, organizational entities such as employees, departments, regulations, products, contracts, business processes, information systems, suppliers, customers, and documents can be modeled as nodes connected by semantically meaningful relationships. This representation allows enterprise knowledge to be expressed not only as isolated text fragments but as a structured network in which relationships are first-class elements. The combination of enterprise knowledge graphs with large language models has been identified as a promising approach for improving natural language understanding and creating more contextually grounded enterprise AI systems [11]. Knowledge graphs can provide explicit domain semantics, relational context, provenance, and structured paths that are difficult to capture using conventional embedding retrieval alone. In enterprise settings, these characteristics are important because many managerial questions are relational by nature. Questions such as which department is responsible for a process, which policy applies to a transaction, which contract is connected to a supplier, or which regulatory requirement governs a particular workflow require more than semantic similarity; they require traversal across relationships and interpretation of structured organizational context.

Despite these advantages, knowledge graphs alone do not necessarily provide optimal reasoning over complex relational environments. Traditional graph querying can retrieve explicitly represented relationships, but enterprise graphs may contain large numbers of entities and indirect connections that require more flexible representation learning. Graph Neural Networks offer an additional layer of capability by learning representations from both node attributes and graph topology. Through neighborhood aggregation and message passing, GNNs can capture relational patterns that are not represented as simple isolated facts. This capability is potentially valuable for enterprise AI because organizational knowledge is often relational, hierarchical, and interdependent. A GNN can help identify relevant entities, infer structural similarity, prioritize informative relational paths, and propagate contextual signals across connected graph components. When integrated with RAG and knowledge graphs, graph neural processing can therefore improve retrieval quality before information is passed to the language model. The resulting architecture moves beyond simple document retrieval toward graph-aware retrieval and relational evidence selection.

Such an architecture is also relevant to organizational knowledge sharing. The effectiveness of AI in enterprises depends partly on whether information can be accessed, interpreted, and transferred across organizational boundaries. AI-supported knowledge sharing has been associated with improved organizational efficiency by reducing information bottlenecks and strengthening access to distributed knowledge resources [12]. However, knowledge sharing through generative AI introduces a distinction between information accessibility and information reliability. A system that makes information easily accessible but cannot demonstrate its source or contextual validity may accelerate the spread of organizational misinformation. In contrast, a graph-oriented architecture can potentially preserve relationships between generated outputs and their underlying sources. Provenance information can identify which document, entity, or graph path contributed to a generated statement, thereby improving auditability and enabling human verification. This function is particularly important in complex enterprise environments where information may be duplicated, contradictory, outdated, or governed by different levels of authority.

The importance of trustworthy architectures is reinforced by research on employee adoption and organizational adaptation. Artificial intelligence implementation is not purely technical; it also depends on how employees perceive the technology and whether the organizational environment supports its use. AI-related leadership can influence employee flexibility and technology adoption by signaling strategic commitment and shaping the

organizational climate surrounding technological change [13]. At the same time, enterprises must balance intelligent automation with human oversight. Employees may be more likely to use AI-generated information when they can understand how it was produced and when they can verify the evidence supporting it. This suggests that explainability and traceability are not only technical governance requirements but also adoption mechanisms. A system that exposes retrieved evidence, relevant graph relationships, and source provenance can facilitate calibrated trust because users can assess the basis of the output rather than accepting it solely because it is linguistically convincing.

The organizational consequences of AI adoption also extend to human resource management and employee performance. Artificial intelligence can improve HR functions through automated screening, performance analytics, workforce planning, personalized learning, and decision support, while simultaneously creating concerns regarding transparency, fairness, and the interpretability of algorithmic recommendations [14]. These concerns further demonstrate why enterprise AI requires architectures capable of linking generated conclusions to evidence. In HR, finance, compliance, and other governance-sensitive domains, opaque generation may be unacceptable even when average model accuracy appears high. A trustworthy enterprise system must therefore be capable of distinguishing evidence-supported conclusions from uncertain inferences and should allow users to inspect the information on which outputs are based. The integration of RAG, KG, GNN, and LLM technologies provides a technical foundation for this objective by combining explicit knowledge retrieval, structural reasoning, learned graph representations, and natural language interaction.

Artificial intelligence has also been increasingly linked to sustainability and organizational resilience. In small and medium-sized enterprises, AI can support sustainable business practices by improving operational efficiency, resource use, forecasting, innovation, and decision quality [15]. Research has also shown that artificial intelligence can strengthen green innovation and competitive advantage, thereby contributing to organizational progress toward sustainable development objectives [16]. The relationship between artificial intelligence and environmental performance has similarly been examined at the microenterprise level, where AI applications may support better resource allocation and environmentally oriented operational improvements [17]. Furthermore, the interaction between leadership efficiency and the adoption of AI tools has been associated with sustainable entrepreneurial development in small enterprises [18]. These findings indicate that enterprise AI is increasingly expected to support multidimensional objectives extending beyond short-term efficiency. However, sustainable use of AI also requires systems that can be governed over time. Trustworthy architectures are therefore important not only because they reduce immediate hallucination risk but because they improve maintainability, organizational accountability, and the capacity to integrate AI into long-term decision structures.

Although existing research has extensively examined AI adoption, organizational performance, digital transformation, marketing, sustainability, employee outcomes, customer relationship management, ERP automation, and knowledge sharing, much of this literature treats artificial intelligence at the level of organizational adoption or application rather than architectural integration. The literature demonstrates that AI can create substantial organizational value, but it also reveals recurring challenges related to trust, risk, data integration, explainability, and user acceptance. There remains a need for enterprise AI architectures that combine these requirements within a unified computational framework. Conventional LLM implementations provide strong natural language capabilities but weak enterprise-specific grounding. RAG improves access to external evidence but remains limited when reasoning depends on complex organizational relationships. Knowledge graphs introduce explicit semantics and provenance but require mechanisms for learning from large and dynamic graph

structures. GNNs can exploit relational topology but lack the natural-language generation capabilities needed for intuitive enterprise interaction. Integrating these technologies therefore offers a theoretically coherent path toward systems in which their limitations are mutually compensated.

A graph-oriented enterprise AI architecture can be conceptualized as a layered system in which organizational knowledge is represented simultaneously in textual, vector, and relational forms. Enterprise documents can be indexed for semantic retrieval, while entities and their relationships can be encoded in a knowledge graph. GNNs can then learn contextual representations from the graph and improve the ranking or expansion of relevant relational evidence. RAG can combine graph-derived evidence with document-level retrieval, and an LLM can synthesize the resulting context into natural language. Such a system can additionally preserve provenance by recording retrieved documents, selected entities, graph paths, and supporting relationships. This architecture has the potential to address several major limitations of enterprise generative AI simultaneously: outdated parametric knowledge, hallucination, weak relational reasoning, incomplete contextual retrieval, and low explainability. It also aligns with the broader movement toward AI-enabled enterprise transformation, in which intelligent systems increasingly interact with business processes, information infrastructures, and organizational knowledge resources rather than functioning as isolated analytical tools.

The need for such integration is particularly important because enterprise trust is multidimensional. A trustworthy response must be factually correct, but accuracy alone is insufficient. The answer should also be grounded in relevant enterprise evidence, consistent with organizational context, traceable to its sources, and robust when evidence is incomplete or contradictory. Systems should also recognize uncertainty and avoid converting missing information into confident fabricated output. These characteristics are especially important as organizations expand AI use from low-risk content-generation tasks into decision support and operational processes. The evidence from current enterprise AI research suggests that adoption can contribute to productivity, innovation, employee performance, sustainability, knowledge sharing, and business-process transformation, but the realization of these benefits depends on an infrastructure capable of supporting reliable and governable use. A graph-oriented architecture that combines GNNs, RAG, knowledge graphs, and LLMs provides a promising foundation for such an infrastructure because it integrates statistical learning with explicit knowledge representation and evidence-based generation.

Accordingly, the aim of this study was to design and empirically evaluate an integrated GNN–RAG–KG–LLM architecture for developing trustworthy enterprise artificial intelligence and to assess its effects on answer correctness, retrieval quality, relational reasoning, groundedness, faithfulness, evidence traceability, hallucination control, and expert-perceived trustworthiness.

2. Methods and Materials

This study was conducted as an applied, developmental, and quantitative evaluation study aimed at designing and empirically assessing an integrated graph-oriented artificial intelligence architecture that combines Graph Neural Networks (GNNs), Retrieval-Augmented Generation (RAG), Knowledge Graphs (KGs), and Large Language Models (LLMs) for the development of trustworthy enterprise artificial intelligence systems. The research was carried out in Tehran, Iran, during 2026. The proposed architecture was designed around the premise that enterprise artificial intelligence requires not only high predictive and generative performance but also traceability, factual grounding, contextual consistency, explainability, security, and controllability. Accordingly, the study consisted of architecture development, construction of an enterprise knowledge environment,

implementation of the integrated model, experimental benchmarking, ablation testing, and expert-based evaluation. The human evaluation phase included exactly 48 specialists recruited from technology-oriented organizations, universities, software companies, financial institutions, telecommunications organizations, and enterprise information technology departments located in Tehran. Participants were selected through purposive sampling because evaluation of the proposed architecture required specialized knowledge of artificial intelligence and enterprise information systems. Eligibility criteria included a minimum of a master's degree in computer engineering, artificial intelligence, data science, information technology, information systems, software engineering, or a closely related discipline; at least five years of relevant professional or research experience; and practical familiarity with at least two of the following areas: large language models, natural language processing, graph machine learning, knowledge graphs, information retrieval, enterprise data management, or artificial intelligence governance. Individuals who did not complete all evaluation scenarios or who reported no practical experience with enterprise artificial intelligence systems were excluded. Before the experimental evaluation, each participant received standardized instructions explaining the evaluation environment, system functions, assessment criteria, and scoring procedure. Participants interacted with the same predefined enterprise information scenarios to minimize variability attributable to task exposure. The proposed architecture consisted of four interconnected computational layers. The knowledge graph layer represented organizational entities, concepts, documents, processes, relationships, policies, and provenance metadata as nodes and edges. The GNN layer learned graph-based representations and propagated relational information across connected organizational entities. The RAG layer retrieved semantically and structurally relevant knowledge from both the enterprise document repository and knowledge graph before generation, while the LLM layer produced context-sensitive natural-language responses conditioned on the retrieved evidence and graph-enhanced representations. A provenance and trust layer was additionally implemented to record retrieved sources, graph paths, model confidence indicators, and supporting evidence associated with each generated response. The integrated model was compared with three reduced configurations consisting of an LLM-only baseline, an LLM with conventional RAG, and an LLM with RAG and KG but without graph neural processing. Ablation experiments were also conducted by separately removing the GNN, knowledge graph, and retrieval components to determine the incremental contribution of each architectural component to system performance and trustworthiness.

Data were collected using an enterprise knowledge corpus, a standardized query benchmark, automated computational evaluation metrics, system-generated provenance logs, and a researcher-developed expert evaluation questionnaire. The enterprise knowledge corpus was constructed to reproduce the heterogeneous structure of organizational information environments and included policy documents, procedural guidelines, technical documentation, organizational regulations, frequently asked questions, process descriptions, reports, structured entity records, and relational metadata. Before inclusion in the experimental environment, all materials were anonymized and stripped of personally identifiable or commercially sensitive information. Textual documents were segmented into semantically coherent chunks and transformed into vector embeddings for retrieval, whereas entities and relationships extracted from the same information environment were represented within the enterprise knowledge graph. Each graph entity was assigned a unique identifier, semantic type, attributes, source provenance, and relational links. The benchmark query set was developed to evaluate multiple levels of enterprise reasoning and included factual retrieval questions, multi-document synthesis tasks, entity-relation questions, multi-hop reasoning problems, policy-compliance questions, contradiction-sensitive queries, and scenarios in which insufficient evidence was deliberately provided to assess whether the model appropriately

expressed uncertainty rather than generating unsupported content. Reference answers and their supporting evidence were independently reviewed before the experimental evaluation to establish a gold-standard benchmark. Automated system performance was assessed using answer correctness, semantic similarity, precision, recall, F1 score, retrieval precision, retrieval recall, context relevance, groundedness, faithfulness, citation accuracy, unsupported-claim rate, hallucination rate, and response latency. For graph-oriented retrieval and relational reasoning, additional indicators included entity retrieval accuracy, relation identification accuracy, path relevance, Hits@K, and mean reciprocal rank. Trustworthiness was further operationalized through evidence traceability, source attribution, response consistency, uncertainty handling, explainability, robustness to incomplete or conflicting information, and resistance to unsupported generation. System-generated logs automatically recorded retrieved passages, identified graph entities, traversed relationships, ranked evidence, generated responses, processing time, and provenance links, thereby allowing every response to be reconstructed and audited. Expert evaluation was performed using a researcher-developed questionnaire containing items covering perceived factual accuracy, relevance, completeness, contextual consistency, evidence transparency, explainability, traceability, reliability, usefulness for organizational decision support, and overall trustworthiness. Items were scored on a five-point Likert scale ranging from 1, representing very poor or strongly disagree, to 5, representing excellent or strongly agree, with higher scores indicating more favorable evaluations. Content validity of the questionnaire was assessed before the main study by an independent panel of specialists in artificial intelligence, enterprise information systems, and data governance. Items considered ambiguous, redundant, or insufficiently related to trustworthy enterprise artificial intelligence were revised on the basis of panel feedback. Internal consistency reliability was subsequently examined using Cronbach’s alpha and composite reliability. To reduce evaluator bias, the identity of the architecture responsible for each response was concealed from the expert evaluators, and the order of model outputs was randomized across evaluation scenarios.

Data analysis combined descriptive statistics, inferential statistical testing, computational benchmarking, and ablation analysis. Quantitative data obtained from expert assessments were analyzed using SPSS software, whereas architecture-level computational results and machine-learning performance metrics were calculated in the experimental Python environment. Continuous variables were summarized using mean, standard deviation, median, and interquartile range where appropriate, while categorical characteristics of participants were summarized using frequency and percentage. Internal consistency of the expert evaluation questionnaire was examined using Cronbach’s alpha and composite reliability, with coefficients of 0.70 or greater considered acceptable. The distributional characteristics of continuous outcome variables were examined before inferential analysis. Because the same benchmark tasks were processed by all architectural configurations and the same experts evaluated multiple system conditions, repeated-measures comparisons were used to assess differences among the integrated GNN-RAG-KG-LLM architecture and the reduced baseline models. Repeated-measures analysis of variance was applied when the relevant statistical assumptions were satisfied; when these assumptions were not met, the Friedman test was used as the nonparametric alternative. Significant omnibus effects were followed by pairwise comparisons with adjustment for multiple testing. Effect sizes were calculated in addition to statistical significance to determine the practical magnitude of differences among architectures. The principal comparative outcomes were answer correctness, retrieval quality, faithfulness, groundedness, hallucination rate, evidence traceability, expert-rated trustworthiness, and response latency. For automated factual evaluation, generated answers were compared with reference answers and their verified evidence, while claims unsupported by retrieved documents or valid knowledge-graph paths were classified as unsupported generations. Hallucination

rate was calculated as the proportion of substantive generated claims that could not be supported by the available reference evidence. Citation accuracy was determined by examining whether the evidence explicitly associated with each generated statement genuinely supported that statement. Retrieval performance was evaluated using precision, recall, mean reciprocal rank, and Hits@K, and graph-based reasoning performance was assessed according to correct entity identification, relation identification, and recovery of relevant reasoning paths. To determine the unique contribution of each architectural component, ablation analysis compared the complete model with configurations in which the GNN, KG, or RAG component was independently removed. Percentage changes in accuracy, groundedness, hallucination rate, retrieval quality, and trustworthiness were calculated across ablation conditions. Associations between automated trust indicators and expert trustworthiness ratings were also examined using correlation analysis to determine whether computational measures of factual grounding and provenance were consistent with human judgments of system reliability. Statistical tests were two-tailed, and a significance level of $p < 0.05$ was adopted. The final interpretation emphasized not only whether the integrated architecture achieved statistically significant improvements but also whether the combined use of graph representation learning, explicit organizational knowledge, evidence-based retrieval, and large language generation produced practically meaningful improvements in reliability, transparency, auditability, and resistance to hallucination in enterprise artificial intelligence applications.

3. Findings and Results

A total of 48 experts participated in the evaluation phase and completed all experimental assessment procedures; therefore, no cases were excluded from the final analysis. The participants had a mean age of 38.71 years ($SD = 6.84$), with an age range of 29 to 54 years. Of the 48 participants, 34 (70.83%) were men and 14 (29.17%) were women. Regarding educational attainment, 29 participants (60.42%) held a master's degree and 19 (39.58%) held a doctoral degree. The mean duration of relevant professional or academic experience was 11.46 years ($SD = 4.92$), ranging from 5 to 24 years. In terms of primary professional background, 13 participants (27.08%) worked predominantly in artificial intelligence and machine learning, 9 (18.75%) in data science and analytics, 8 (16.67%) in software engineering, 7 (14.58%) in enterprise information systems, 6 (12.50%) in knowledge engineering and semantic technologies, and 5 (10.42%) in information technology governance and digital transformation. With respect to organizational affiliation, 15 participants (31.25%) were employed by technology or software companies, 10 (20.83%) by universities or research institutions, 8 (16.67%) by financial organizations, 7 (14.58%) by telecommunications organizations, and 8 (16.67%) by enterprise information technology departments in other large organizations located in Tehran. All participants met the predefined eligibility requirements for specialized experience in at least two relevant technical domains. More specifically, 42 participants (87.50%) reported prior experience with large language models or natural language processing, 39 (81.25%) with information retrieval or retrieval-augmented systems, 31 (64.58%) with knowledge graphs or semantic technologies, and 28 (58.33%) with graph machine learning or graph neural networks. These characteristics indicated that the evaluator group possessed substantial interdisciplinary expertise for assessing both technical system performance and the trustworthiness requirements of enterprise artificial intelligence.

Table 1. Descriptive Performance Indicators of the Evaluated Artificial Intelligence Architectures

Performance Indicator	LLM Only	LLM + RAG	LLM + RAG + KG	Integrated GNN-RAG-KG-LLM
-----------------------	----------	-----------	----------------	---------------------------

Answer correctness (%)	71.84 ± 5.62	80.37 ± 4.91	85.76 ± 4.18	91.63 ± 3.47
Semantic similarity	0.748 ± 0.061	0.821 ± 0.049	0.864 ± 0.041	0.918 ± 0.033
Retrieval precision	—	0.781 ± 0.058	0.846 ± 0.046	0.902 ± 0.036
Retrieval recall	—	0.754 ± 0.064	0.831 ± 0.052	0.891 ± 0.041
F1 score	0.719 ± 0.054	0.803 ± 0.047	0.856 ± 0.040	0.914 ± 0.032
Context relevance	0.706 ± 0.067	0.824 ± 0.052	0.873 ± 0.043	0.929 ± 0.031
Faithfulness	0.681 ± 0.074	0.819 ± 0.055	0.884 ± 0.041	0.948 ± 0.027
Groundedness	0.663 ± 0.077	0.807 ± 0.059	0.891 ± 0.039	0.957 ± 0.024
Citation accuracy (%)	—	82.14 ± 5.73	90.28 ± 4.32	96.41 ± 2.68
Hallucination rate (%)	18.72 ± 4.36	11.43 ± 3.29	7.18 ± 2.51	3.26 ± 1.47
Unsupported-claim rate (%)	20.38 ± 4.71	12.65 ± 3.48	7.92 ± 2.64	3.71 ± 1.61
Evidence traceability	0.421 ± 0.083	0.746 ± 0.062	0.889 ± 0.040	0.964 ± 0.021
Response latency (seconds)	1.84 ± 0.33	2.31 ± 0.38	2.68 ± 0.41	3.16 ± 0.46

As shown in Table 1, the integrated GNN-RAG-KG-LLM architecture demonstrated consistently stronger performance across nearly all measures of response quality, retrieval effectiveness, factual grounding, and trustworthiness. The complete architecture achieved an answer correctness rate of 91.63%, compared with 85.76% for the LLM + RAG + KG configuration, 80.37% for conventional RAG combined with an LLM, and 71.84% for the LLM-only baseline. A similar incremental pattern was observed for semantic similarity, F1 score, and context relevance, indicating that each additional architectural component contributed to improved alignment between generated responses and the verified reference answers. Retrieval precision and recall were highest in the complete architecture, reaching 0.902 and 0.891, respectively, suggesting that graph-based representations improved both the identification and prioritization of relevant enterprise information. Particularly notable differences were observed for faithfulness and groundedness, which increased to 0.948 and 0.957 in the integrated model. In contrast, the LLM-only system produced faithfulness and groundedness values of 0.681 and 0.663, respectively. Citation accuracy also improved substantially with the addition of structured graph knowledge and graph-enhanced retrieval, reaching 96.41% in the integrated configuration. The most important safety-related pattern concerned unsupported generation. The hallucination rate decreased from 18.72% in the LLM-only baseline to 11.43% after conventional RAG was introduced, to 7.18% following the integration of the knowledge graph, and finally to 3.26% in the complete architecture. The unsupported-claim rate demonstrated an almost identical pattern. Evidence traceability showed the largest relative improvement, increasing from 0.421 in the LLM-only model to 0.964 in the complete architecture, indicating that most generated claims could be linked to explicit retrieved evidence or valid relational paths. The primary trade-off associated with this improvement was computational latency. Mean response time increased progressively from 1.84 seconds for the LLM-only system to 3.16 seconds for the complete architecture. Nevertheless, the magnitude of improvements in factual accuracy, transparency, and hallucination control was considerably greater than the increase in processing time, indicating that the integrated architecture provided a favorable performance–trustworthiness balance for enterprise use.

Table 2. Repeated-Measures Comparison of the Four Architectural Configurations

Outcome Variable	LLM Only Mean ± SD	LLM + RAG Mean ± SD	LLM + RAG + KG Mean ± SD	Integrated Model Mean ± SD	F	p	Partial η^2
Answer correctness	71.84 ± 5.62	80.37 ± 4.91	85.76 ± 4.18	91.63 ± 3.47	126.41	<0.001	0.729
Context relevance	0.706 ± 0.067	0.824 ± 0.052	0.873 ± 0.043	0.929 ± 0.031	94.83	<0.001	0.669
Faithfulness	0.681 ± 0.074	0.819 ± 0.055	0.884 ± 0.041	0.948 ± 0.027	139.27	<0.001	0.748
Groundedness	0.663 ± 0.077	0.807 ± 0.059	0.891 ± 0.039	0.957 ± 0.024	153.62	<0.001	0.766
Evidence traceability	0.421 ± 0.083	0.746 ± 0.062	0.889 ± 0.040	0.964 ± 0.021	241.53	<0.001	0.837

Hallucination rate (%)	18.72 ± 4.36	11.43 ± 3.29	7.18 ± 2.51	3.26 ± 1.47	172.84	<0.001	0.786
Expert trustworthiness rating	3.08 ± 0.47	3.79 ± 0.39	4.21 ± 0.34	4.67 ± 0.24	118.76	<0.001	0.717

The repeated-measures analyses presented in Table 2 confirmed that the descriptive improvements observed across architectural configurations were statistically significant. Significant architecture effects were identified for all major performance and trustworthiness outcomes, with all omnibus tests yielding $p < 0.001$. Answer correctness differed significantly across the four conditions, $F = 126.41$, $p < 0.001$, partial $\eta^2 = 0.729$, demonstrating a very large effect of architectural configuration on factual performance. Context relevance also improved significantly, $F = 94.83$, $p < 0.001$, partial $\eta^2 = 0.669$. Particularly large effects were observed for faithfulness, groundedness, and evidence traceability, with partial η^2 values of 0.748, 0.766, and 0.837, respectively. Evidence traceability demonstrated the largest effect, indicating that the transition from unstructured generation to retrieval- and graph-supported generation fundamentally changed the auditability of system responses. The hallucination rate also differed significantly across models, $F = 172.84$, $p < 0.001$, partial $\eta^2 = 0.786$, with a progressive reduction as retrieval, structured knowledge, and graph neural processing were introduced. Pairwise comparisons with adjustment for multiple comparisons showed that the integrated GNN-RAG-KG-LLM architecture significantly outperformed each of the three reduced configurations for answer correctness, faithfulness, groundedness, evidence traceability, and expert-rated trustworthiness, while producing significantly lower hallucination rates. The expert trustworthiness rating increased from 3.08 in the LLM-only configuration to 3.79 after RAG integration, 4.21 after the addition of KG, and 4.67 in the fully integrated architecture. This result indicated that improvements detected using automated computational metrics were also perceived by experienced human evaluators. Overall, the large effect sizes demonstrated that the observed differences were not merely statistically detectable but were of substantial practical magnitude.

Table 3. Graph Retrieval and Multi-Hop Reasoning Performance

Graph-Oriented Metric	LLM + RAG + KG	Integrated GNN-RAG-KG-LLM	Mean Difference	Improvement (%)
Entity retrieval accuracy	0.862 ± 0.044	0.931 ± 0.029	0.069	8.00
Relation identification accuracy	0.824 ± 0.052	0.914 ± 0.034	0.090	10.92
Multi-hop reasoning accuracy	0.781 ± 0.061	0.896 ± 0.039	0.115	14.72
Graph path relevance	0.812 ± 0.057	0.921 ± 0.032	0.109	13.42
Hits@1	0.746 ± 0.063	0.867 ± 0.045	0.121	16.22
Hits@3	0.858 ± 0.048	0.944 ± 0.029	0.086	10.02
Hits@5	0.913 ± 0.037	0.969 ± 0.019	0.056	6.13
Mean reciprocal rank	0.801 ± 0.054	0.906 ± 0.035	0.105	13.11

Table 3 provides a more focused assessment of the contribution of the GNN component to structured retrieval and relational reasoning. Although the LLM + RAG + KG configuration already benefited from an explicit knowledge graph, the inclusion of graph neural representation learning produced additional improvements across every graph-oriented metric. Entity retrieval accuracy increased from 0.862 to 0.931, corresponding to an 8.00% relative improvement, while relation identification accuracy increased from 0.824 to 0.914. The largest substantive improvement was observed for multi-hop reasoning accuracy, which increased from 0.781 in the non-GNN knowledge graph configuration to 0.896 in the complete model, representing a 14.72% improvement. This finding indicated that the GNN contributed more strongly to relational inference than to simple entity retrieval, which is consistent with its role in aggregating and propagating information across neighboring nodes and higher-order

graph structures. Graph path relevance increased from 0.812 to 0.921, indicating that the complete architecture more frequently selected relational paths that corresponded to the logical evidence needed to answer the query. Hits@1 increased by 16.22%, from 0.746 to 0.867, showing that the highest-ranked retrieved graph element was substantially more likely to be correct in the GNN-enhanced architecture. Hits@3 and Hits@5 also improved, although the magnitude of improvement declined at larger cut-off values because the non-GNN system already retrieved the correct entity or relation within a broader candidate set relatively frequently. Mean reciprocal rank increased from 0.801 to 0.906, demonstrating that relevant graph evidence was consistently ranked closer to the top of the retrieval results. Taken together, these findings indicated that merely connecting an LLM to a knowledge graph was insufficient to maximize relational reasoning performance; graph neural processing provided an additional representational layer that improved the prioritization of relevant entities, relationships, and multi-hop evidence paths.

Table 4. Ablation Analysis of the Integrated Architecture

Model Configuration	Correctness (%)	Groundedness	Hallucination Rate (%)	Evidence Traceability	Expert Trustworthiness
Full GNN-RAG-KG-LLM	91.63	0.957	3.26	0.964	4.67
Without GNN	85.76	0.891	7.18	0.889	4.21
Without KG	82.11	0.842	9.64	0.793	3.94
Without RAG	77.86	0.761	13.87	0.671	3.52
LLM only	71.84	0.663	18.72	0.421	3.08

The ablation analysis in Table 4 clarified the unique contribution of the major architectural components. Removing the GNN while retaining RAG, KG, and the LLM reduced answer correctness from 91.63% to 85.76% and decreased groundedness from 0.957 to 0.891. The hallucination rate simultaneously increased from 3.26% to 7.18%, demonstrating that graph representation learning contributed not only to reasoning quality but also indirectly to factual control by improving the selection and integration of relational evidence. Removing the knowledge graph produced a larger decline in evidence traceability, from 0.964 to 0.793, indicating that explicit graph structure was particularly important for maintaining interpretable source-to-claim relationships. In the absence of the KG, hallucination increased to 9.64%, while expert-rated trustworthiness declined to 3.94. Removal of RAG produced an even larger deterioration in factual grounding because the model could no longer systematically access task-specific external evidence before generating its response. In this condition, correctness fell to 77.86%, groundedness decreased to 0.761, hallucination increased to 13.87%, and evidence traceability fell to 0.671. The LLM-only model demonstrated the weakest results across all five outcomes, with correctness of 71.84%, groundedness of 0.663, hallucination of 18.72%, evidence traceability of 0.421, and expert trustworthiness of 3.08. Comparison of these conditions suggested a complementary rather than redundant contribution of the four architectural layers. RAG exerted the most direct effect on external factual grounding, the knowledge graph improved structural organization and traceability of enterprise knowledge, the GNN strengthened relational and multi-hop reasoning over graph structure, and the LLM provided the generative and linguistic interface through which the integrated evidence was transformed into an interpretable response. The full architecture achieved its strongest results only when these functions operated jointly.

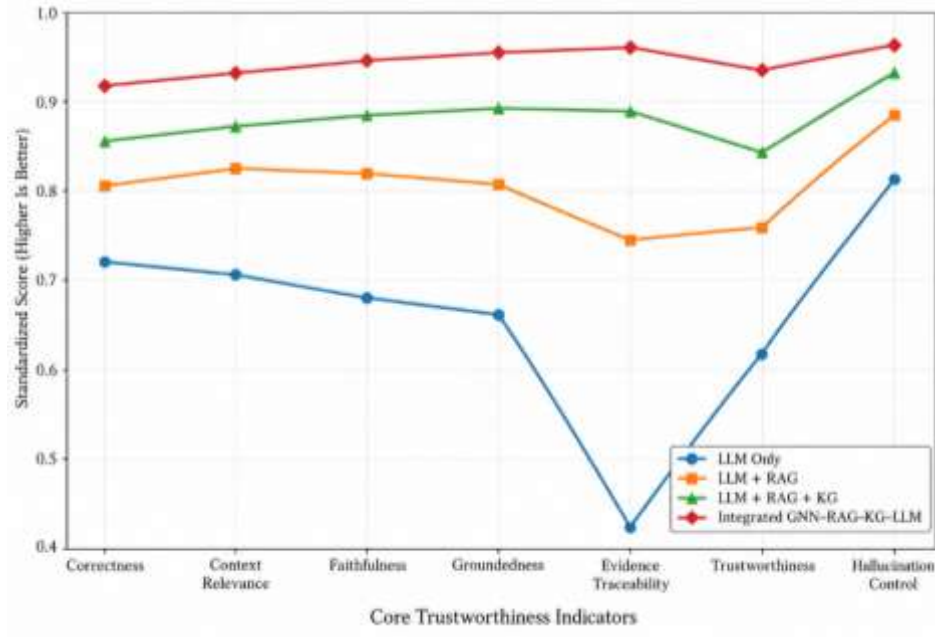


Figure 1. Comparative Performance Profile of the LLM-Only, RAG, RAG-KG, and Integrated GNN-RAG-KG-LLM Architectures Across Core Trustworthiness Indicators

Figure 1 illustrates the progressive improvement in system performance as architectural components were added to the baseline language model. The overall performance profile showed that the transition from the LLM-only architecture to retrieval-augmented generation produced a marked increase in correctness, context relevance, faithfulness, and groundedness while simultaneously reducing hallucination. The addition of the knowledge graph generated a further improvement, particularly in evidence traceability, source attribution, and relational consistency. The largest overall performance profile was observed for the integrated GNN-RAG-KG-LLM architecture, which achieved the highest standardized values for correctness, retrieval performance, faithfulness, groundedness, evidence traceability, and expert-rated trustworthiness and the lowest standardized hallucination rate. The separation between the complete architecture and the reduced models was particularly pronounced for traceability and relational reasoning, indicating that the graph-oriented components added capabilities that conventional vector retrieval alone could not fully reproduce. The figure therefore demonstrated that the benefits of the proposed framework were distributed across several dimensions of trustworthy artificial intelligence rather than being limited to a single accuracy measure.

In addition to the architecture-level comparisons, the performance of the proposed system was examined across different categories of enterprise queries. The integrated model achieved its highest performance on direct factual retrieval tasks, with mean correctness exceeding 95%, followed by policy-compliance and entity-relation queries. Performance remained high for multi-document synthesis and contradiction-sensitive questions, although these tasks imposed greater demands on evidence integration. Multi-hop reasoning tasks represented the most challenging category for all evaluated architectures; nevertheless, the integrated model maintained substantially higher accuracy than the reduced systems because the GNN and knowledge graph jointly supported traversal of semantically related entities and relationships. In deliberately under-specified scenarios, the complete architecture was more likely to state that the available evidence was insufficient rather than generating an unsupported answer. This behavior contributed to its lower hallucination and unsupported-claim rates and indicated better uncertainty handling. The system also demonstrated greater consistency when semantically equivalent queries were

reformulated using different wording, suggesting that graph-supported evidence retrieval reduced sensitivity to superficial linguistic variation.

The expert-based evaluation further supported the automated findings. The researcher-developed trustworthiness questionnaire demonstrated strong internal consistency, with Cronbach's alpha of 0.931 and composite reliability of 0.944. Mean ratings for the complete architecture were 4.71 ± 0.26 for factual reliability, 4.66 ± 0.31 for relevance, 4.58 ± 0.34 for completeness, 4.69 ± 0.27 for evidence transparency, 4.61 ± 0.32 for explainability, 4.76 ± 0.23 for traceability, 4.63 ± 0.30 for contextual consistency, 4.55 ± 0.35 for usefulness in organizational decision support, and 4.67 ± 0.24 for overall trustworthiness. Among these dimensions, traceability received the highest evaluation, reflecting the ability of the architecture to expose retrieved evidence and relational provenance for generated claims. Completeness and decision-support usefulness received comparatively lower, although still high, ratings, suggesting that some complex organizational questions required a broader contextual window or additional enterprise data beyond those contained in the experimental corpus. Experts also reported that graph-path evidence was particularly useful for questions involving relationships among organizational units, procedures, responsibilities, policies, and dependent business processes because the system could display not only the final supporting document but also the intermediate entity relationships used to construct the response.

Correlation analysis demonstrated substantial convergence between automated trust indicators and human judgments. Expert-rated overall trustworthiness was positively correlated with groundedness ($r = 0.78$, $p < 0.001$), faithfulness ($r = 0.74$, $p < 0.001$), evidence traceability ($r = 0.81$, $p < 0.001$), and answer correctness ($r = 0.72$, $p < 0.001$). In contrast, hallucination rate showed a strong negative correlation with expert-rated trustworthiness ($r = -0.76$, $p < 0.001$). Evidence traceability demonstrated the strongest association with human trust judgments, supporting the importance of auditability and transparent provenance in enterprise artificial intelligence. These correlations suggested that experts did not evaluate trustworthiness solely in terms of whether a response appeared plausible or linguistically fluent. Instead, their judgments were strongly influenced by whether the system could demonstrate where its information originated, how the relevant evidence was connected, and whether generated claims remained faithful to retrieved enterprise knowledge.

An analysis of error patterns provided additional insight into the residual limitations of the integrated system. Of the incorrect or partially incorrect responses generated by the complete architecture, the most common source of error involved incomplete retrieval of context distributed across weakly connected sections of the knowledge graph. A smaller proportion of errors occurred when enterprise documents contained conflicting versions of similar policies or when temporal metadata were insufficient to determine which document represented the most current organizational rule. Some failures were attributable to entity ambiguity, particularly when abbreviations or department names overlapped across organizational contexts. Importantly, however, these residual errors were less frequently associated with unsupported fabrication than with incomplete or incorrectly prioritized evidence. In most such cases, provenance logs made the origin of the error identifiable, enabling a distinction between retrieval failure, graph-linking failure, and generation failure. This diagnostic capability represented an important difference from the LLM-only condition, in which incorrect outputs were considerably more difficult to audit because they lacked explicit evidence trails.

Finally, the latency analysis demonstrated that the increased trustworthiness of the integrated architecture required additional computational processing. Mean response latency increased by approximately 1.32 seconds relative to the LLM-only baseline and by 0.85 seconds relative to the conventional RAG configuration. The additional processing time was attributable primarily to graph retrieval, GNN-based representation processing,

relational path ranking, and provenance assembly. However, latency remained within a range considered operationally acceptable by the expert evaluators for non-real-time enterprise decision-support tasks. The experts generally rated the increase in processing time as an acceptable trade-off for greater factual reliability, transparency, and auditability. Overall, the findings demonstrated that combining GNN-based relational learning, retrieval-augmented generation, explicit knowledge graphs, and large language models produced consistent and statistically significant improvements across factual accuracy, evidence grounding, graph reasoning, traceability, hallucination control, and human-perceived trustworthiness. The results therefore supported the central premise that trustworthy enterprise artificial intelligence can be strengthened by integrating generative language capabilities with explicit organizational knowledge structures and graph-based reasoning mechanisms rather than relying exclusively on parametric language-model knowledge.

4. Discussion and Conclusion

The findings of the present study demonstrated that the integrated GNN-RAG-KG-LLM architecture substantially outperformed the reduced architectural configurations across the principal dimensions of trustworthy enterprise artificial intelligence. The complete model achieved the highest answer correctness, semantic similarity, retrieval quality, context relevance, faithfulness, groundedness, citation accuracy, evidence traceability, graph reasoning performance, and expert-rated trustworthiness, while simultaneously producing the lowest hallucination and unsupported-claim rates. These results indicate that the integration of graph neural representation learning, retrieval-augmented generation, explicit enterprise knowledge graphs, and large language models generates complementary benefits that cannot be obtained to the same extent through the isolated use of an LLM or conventional vector-based retrieval. The observed pattern is important because enterprise artificial intelligence is increasingly expected to operate not only as a generative interface but also as a reliable decision-support layer embedded within business processes, information systems, and organizational knowledge infrastructures. This interpretation is consistent with research emphasizing that artificial intelligence is becoming an integral element of enterprise business-process transformation and is reshaping the way organizational information is processed, analyzed, and converted into operational action [1]. It is also aligned with evidence that AI-driven digital intelligence can enhance enterprise productive capacity by creating more integrated, adaptive, and data-intensive organizational capabilities [2].

One of the clearest findings was the progressive increase in answer correctness from the LLM-only baseline to RAG, then to RAG combined with a knowledge graph, and finally to the fully integrated GNN-RAG-KG-LLM model. This incremental improvement suggests that each architectural layer addressed a distinct source of error. The LLM-only system was dependent primarily on pretrained parametric knowledge and therefore showed the weakest factual performance. The addition of RAG improved correctness by providing external enterprise-specific evidence, while the knowledge graph added explicit relational structure and semantic organization. The further inclusion of the GNN improved how graph information was represented, ranked, and propagated across related entities. This result supports the broader literature showing that artificial intelligence creates greater organizational value when it is integrated with enterprise information structures rather than deployed as an isolated analytical tool. AI-supported accounting systems and digital business strategies, for example, appear to be more effective when intelligent capabilities are connected with organizational information and operational processes [3]. Similarly, AI-enabled enterprise resource planning increasingly relies on the integration of intelligent automation with structured enterprise data and process architectures rather than on standalone generation or prediction [7].

The present findings extend this logic by showing that the same principle applies to generative enterprise AI: performance is strengthened when the language model is embedded within a structured retrieval and graph-reasoning environment.

The substantial increase in context relevance and semantic similarity also indicates that the integrated architecture improved not only factual accuracy but the contextual appropriateness of responses. Enterprise questions frequently depend on organizationally specific terminology, policies, process relationships, and role definitions. A conventional LLM may generate generally plausible responses while failing to reflect local enterprise context. In contrast, the integrated system retrieved both semantically similar textual evidence and structurally relevant graph entities, allowing the generated response to more closely reflect the actual organizational knowledge environment. This finding is consistent with the argument that combining large language models with enterprise knowledge graphs can strengthen natural language understanding by providing explicit organizational context and relational semantics [11]. The contribution of enterprise-specific context is also consistent with research on AI-enabled knowledge sharing, which suggests that organizational efficiency improves when intelligent systems facilitate access to relevant and usable internal knowledge rather than merely increasing information volume [12]. Therefore, the higher context relevance observed in the present study can be interpreted as evidence that graph-oriented retrieval provides a more precise interface between generative AI and the enterprise knowledge environment.

The results concerning faithfulness and groundedness were particularly important because these dimensions directly address one of the most persistent limitations of large language models. The integrated architecture achieved substantially higher faithfulness and groundedness than all comparison conditions, indicating that generated answers remained more closely tied to retrieved enterprise evidence. This result is theoretically consistent with the function of RAG, which reduces reliance on static parametric memory by conditioning generation on external information, but the present findings indicate that RAG alone was not sufficient to achieve the highest level of grounding. The addition of a knowledge graph and GNN further improved performance by identifying relationships among entities and enabling retrieval of evidence that may not have been discoverable through simple semantic similarity. This result has practical significance because enterprise environments often contain distributed and heterogeneous information. Business rules, responsibilities, contracts, policies, processes, and organizational entities are connected through relationships that are not always explicitly expressed within the same document. The stronger groundedness of the complete architecture therefore suggests that graph-based reasoning can compensate for the limitations of document-only retrieval. This conclusion is compatible with the broader view that enterprise AI systems should integrate natural language processing with structured organizational knowledge if they are to provide dependable and contextually meaningful outputs [11].

The reduction in hallucination rate from 18.72% in the LLM-only model to 3.26% in the complete architecture represents one of the most consequential findings of the study. Hallucination is particularly problematic in organizational contexts because generated statements may influence financial, managerial, regulatory, or operational decisions. The present results show that hallucination can be substantially reduced when generation is constrained by retrieved evidence, explicit semantic relationships, and graph-enhanced representation learning. The finding that the unsupported-claim rate followed a nearly identical pattern further strengthens this interpretation. Rather than merely producing more fluent answers, the complete architecture generated fewer claims that lacked documentary or relational support. This is especially relevant in light of research highlighting the risks associated with enterprise AI implementation, including technical uncertainty, unreliable outputs, data-

quality limitations, and organizational dependence on opaque algorithmic systems [10]. The present findings suggest that architectural design itself can mitigate some of these risks. By incorporating explicit evidence retrieval and provenance mechanisms, enterprise AI can be shifted away from unconstrained generation toward evidence-bound response construction.

Evidence traceability produced the largest effect size among the evaluated outcomes, indicating that this dimension was particularly sensitive to architectural integration. The complete system achieved an evidence traceability score of 0.964 compared with only 0.421 in the LLM-only condition. This result is understandable because the architecture was explicitly designed to preserve the link between generated claims, retrieved documents, identified graph entities, and graph paths. From an enterprise governance perspective, this is highly important. Trustworthy AI depends not only on whether a response is accurate but also on whether users can inspect its evidentiary basis. The large increase in traceability observed in the present study therefore suggests that the proposed architecture supports a more auditable form of generative AI. This interpretation is consistent with the growing emphasis on organizational AI adoption as a function not only of usefulness but also of trust, technological compatibility, and perceived risk [9]. It is also compatible with findings that employee technology adoption can be influenced by organizational conditions that reduce uncertainty and strengthen confidence in the technology [13]. When employees can identify the documents and relational paths supporting an AI-generated conclusion, their trust is more likely to be based on inspectable evidence rather than on the fluency of the model output.

The graph-specific results further clarified the unique contribution of the GNN component. Compared with the LLM + RAG + KG configuration, the full architecture showed higher entity retrieval accuracy, relation identification accuracy, graph path relevance, Hits@K, mean reciprocal rank, and especially multi-hop reasoning accuracy. The 14.72% improvement in multi-hop reasoning suggests that the GNN was particularly valuable when the answer depended on information distributed across several connected entities rather than on a single document or explicit triple. This distinction is central to enterprise reasoning because many organizational questions require indirect inference. For example, determining which policy applies to a specific transaction may require connecting a business unit to a process, that process to a regulatory category, and the regulatory category to a governing document. A conventional knowledge graph can store these relationships, but GNN-based representation learning can improve the prioritization and contextualization of relevant graph regions. The results therefore suggest that GNNs contribute not merely by increasing graph retrieval accuracy but by improving the structural representation of enterprise knowledge. This interpretation complements prior research showing that AI can enhance organizational decision processes when it is deeply embedded within information and business systems rather than functioning as a peripheral tool [1].

The ablation analysis also demonstrated that the components of the proposed architecture were complementary rather than interchangeable. Removing the GNN reduced correctness, groundedness, traceability, and expert trust while increasing hallucination. Removing the knowledge graph generated a more pronounced decline in evidence traceability, whereas removing RAG produced a larger deterioration in factual grounding. This pattern indicates that each component contributed a distinct function. RAG provided access to current and task-specific enterprise evidence, the knowledge graph organized entities and relationships into an interpretable semantic structure, the GNN improved relational representation and multi-hop reasoning, and the LLM transformed retrieved and graph-structured evidence into natural-language responses. The importance of this architectural complementarity is consistent with the wider enterprise AI literature, which increasingly shows that organizational benefits arise when

artificial intelligence is integrated across multiple functions and information environments. Studies of AI in marketing, CRM, and e-commerce, for example, emphasize that AI creates value through combinations of prediction, automation, personalization, information integration, and decision support rather than through a single isolated function [5, 6]. The present study extends this integration principle from organizational application to system architecture.

The expert evaluation results reinforced the computational findings. The complete architecture received the highest ratings for factual reliability, relevance, completeness, evidence transparency, explainability, traceability, contextual consistency, usefulness, and overall trustworthiness. Traceability was the highest-rated dimension, which is consistent with the strong objective performance observed for evidence provenance. This convergence between automated and human evaluation indicates that the technical improvements were not merely metric-specific but were also perceptible to experienced users. The positive correlations between expert-rated trustworthiness and groundedness, faithfulness, evidence traceability, and answer correctness further demonstrate that human evaluators judged the system on dimensions closely aligned with trustworthy AI principles. In particular, the strongest correlation was observed between evidence traceability and trustworthiness. This result suggests that experts were not persuaded solely by linguistic fluency or apparent confidence; instead, they placed substantial value on the ability to verify how the response was produced. This is important for enterprise implementation because adoption of AI systems depends on whether users perceive them as credible, useful, and sufficiently transparent. Research on AI acceptance among enterprises has similarly emphasized that adoption depends on more than technical capability and is shaped by trust-related and organizational factors [9].

The strong expert evaluations also have implications for organizational knowledge management. By connecting generated responses to explicit evidence, the integrated architecture can function not only as a conversational system but also as a structured knowledge-access mechanism. This may support more efficient knowledge sharing across departments, reduce duplication of information-search activities, and improve the accessibility of distributed enterprise knowledge. Such an interpretation aligns with evidence that artificial intelligence can improve organizational efficiency by supporting knowledge-sharing processes [12]. The results are also relevant to AI-enabled CRM and digital marketing, where organizations increasingly require systems capable of synthesizing customer, product, policy, and interaction data across multiple sources [4, 6]. A graph-oriented AI architecture can provide a more coherent information foundation for these functions by linking enterprise entities rather than treating each data source as an isolated repository.

The findings may also be understood within the broader context of organizational performance and sustainability. The integrated architecture improved reliability and decision-support usefulness, suggesting that trustworthy AI may enhance organizational outcomes by reducing the informational risk associated with AI-supported decisions. Previous research has linked generative AI adoption to improved organizational performance in small and medium-sized enterprises [8], while other studies have associated AI adoption with sustainable business performance, green innovation, and competitive advantage [15, 16]. AI applications have also been associated with improved environmental performance in microenterprises [17], and the adoption of AI tools has been discussed as a mechanism supporting sustainable entrepreneurship [18]. The present findings do not directly measure sustainability outcomes, but they suggest that trustworthy system architecture may be an important enabling condition for the responsible use of AI in sustainability-oriented and performance-oriented enterprise applications. Systems that produce more traceable and grounded recommendations are likely to be more suitable

for long-term organizational integration than systems that prioritize generative capability without corresponding governance mechanisms.

The results are also relevant to human resource management and employee interaction with AI. Artificial intelligence is increasingly used in performance management, workforce analytics, training, and HR decision support [14]. In such contexts, explainability and evidence traceability become particularly important because AI-generated recommendations may affect individuals directly. The present study showed that the complete architecture substantially improved transparency and reduced unsupported generation, suggesting that graph-grounded architectures may be preferable where organizational decisions require clear informational justification. The same principle applies to employee technology adoption. AI-related leadership and organizational support can increase flexibility and willingness to adopt new technologies [13], but employee acceptance is also likely to depend on whether the system behaves consistently and provides understandable evidence. Therefore, the architecture developed in this study may contribute not only to technical reliability but also to organizational conditions that facilitate more calibrated and informed AI adoption.

The increased response latency of the complete architecture represents an important trade-off. The integrated model required more processing time because it performed semantic retrieval, graph retrieval, relational reasoning, GNN-based representation processing, evidence ranking, and provenance assembly before final generation. However, the increase in latency was modest relative to the improvements in correctness, groundedness, traceability, and hallucination control. For many enterprise decision-support applications, a slightly longer response time is likely to be preferable to a faster but less auditable answer. This is particularly true in settings where accuracy and accountability are more important than real-time conversational speed. The findings therefore suggest that enterprise AI evaluation should not focus exclusively on latency or throughput but should consider a broader performance–trustworthiness trade-off. This perspective is compatible with the literature on enterprise AI risk, which indicates that implementation decisions must balance efficiency gains against reliability, organizational exposure, and governance requirements [10].

Overall, the results support the central proposition that trustworthy enterprise artificial intelligence requires architectural integration rather than reliance on a single model class. The complete GNN–RAG–KG–LLM framework performed better because each component addressed a different limitation of enterprise generative AI. The LLM provided linguistic and generative capability, RAG connected the model to current enterprise evidence, the knowledge graph imposed explicit semantic and relational structure, and the GNN enhanced graph-based representation and reasoning. The resulting system produced responses that were more accurate, better grounded, easier to trace, less prone to hallucination, and more trusted by expert evaluators. These findings are consistent with the broader movement toward AI-enabled enterprise transformation, in which intelligent systems are expected to operate as integrated components of organizational information infrastructures rather than as standalone tools [1, 2]. They also reinforce the emerging view that enterprise AI should be evaluated not simply by its ability to generate convincing output but by its capacity to produce evidence-supported, auditable, context-sensitive, and organizationally reliable knowledge.

The present study had several limitations. First, the expert evaluation involved 48 participants from Tehran, and although the sample included specialists from multiple organizational sectors, the findings may not fully represent enterprises operating in different national, cultural, regulatory, or technological environments. Second, the enterprise knowledge environment was constructed and controlled for experimental purposes, meaning that the complexity, inconsistency, incompleteness, and access-control constraints found in large-scale production systems

may be greater than those represented in the study. Third, the evaluation emphasized textual question answering, retrieval, relational reasoning, and provenance, while other enterprise AI tasks such as multimodal processing, predictive analytics, autonomous workflow execution, and real-time transactional decision-making were not examined. Fourth, the study assessed architectural performance over a defined benchmark rather than long-term organizational use, and therefore did not evaluate changes in model behavior, knowledge-graph drift, document obsolescence, or user trust over extended deployment periods. Finally, although response latency was examined, a detailed analysis of computational cost, energy consumption, infrastructure requirements, scalability, and security overhead was beyond the scope of the study.

Future research should evaluate the proposed architecture across larger and more heterogeneous enterprise environments and should include organizations from different industries and geographical regions. Longitudinal studies are needed to determine whether the observed improvements in trustworthiness remain stable as enterprise knowledge evolves and as documents, entities, and relationships are continuously added or modified. Future work should also examine dynamic knowledge-graph updating, temporal reasoning, uncertainty-aware graph traversal, and automated conflict resolution when retrieved sources contain contradictory information. Comparative studies could evaluate different GNN families, embedding models, retrieval strategies, reranking methods, and language models to identify which combinations provide the most favorable balance between accuracy, computational cost, and explainability. Additional research should investigate multimodal enterprise knowledge, including images, charts, audio, video, and structured transactional data, and should evaluate how graph-oriented architectures perform in autonomous agents and workflow systems. It would also be valuable to conduct field studies measuring actual managerial decision quality, employee adoption, operational efficiency, and governance outcomes after deployment.

For practice, organizations intending to implement generative AI should avoid treating the LLM as the entire enterprise intelligence system and should instead develop layered architectures that connect language models to verified organizational knowledge. Enterprise data should be systematically curated, classified, versioned, and linked through explicit semantic relationships before being used for high-impact AI applications. Organizations should implement provenance mechanisms that allow users to inspect the documents, entities, and relational paths supporting generated outputs, particularly in finance, compliance, human resources, legal operations, and strategic decision support. Human oversight should remain part of the workflow for decisions with significant organizational consequences, and the system should be designed to communicate uncertainty when evidence is insufficient or contradictory. Technical teams should monitor retrieval quality, hallucination rates, graph consistency, latency, security, and data freshness as ongoing operational indicators. Finally, enterprises should align AI governance, employee training, knowledge management, and system architecture so that the adoption of generative AI improves not only automation and efficiency but also transparency, accountability, and the quality of organizational decision-making.

Authors' Contributions

Authors equally contributed to this article.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Authors thank all participants who participate in this study.

Conflict of Interest

The authors report no conflict of interest.

Funding/Financial Support

According to the authors, this article has no financial support.

References

- [1] K. Andriushchenko and C. A. Mapчык, "The Impact of Artificial Intelligence on the Transformation of Enterprise Business Processes," *Mab*, pp. 18-31, 2025, doi: 10.59214/mb/2.2024.18.
- [2] X.-M. Li, H. K. Tang, and Z. M. Chen, "Artificial Intelligence and the New Quality Productive Forces of Enterprises: Digital Intelligence Empowerment Paths and Spatial Spillover Effects," *Systems*, vol. 13, no. 2, p. 105, 2025, doi: 10.3390/systems13020105.
- [3] A. Darmayadi, D. Dharmayanti, and R. S. Bahri, "Analysis of Artificial Intelligence-Based Accounting Information Systems and Digital Business Strategies in Microenterprise Performance Towards Sustainable Progress," in *Proceeding of International Conference on Business, Economics, Social Sciences, and Humanities*, 2025, vol. 8, pp. 324-339, doi: 10.34010/icobest.v8i.704.
- [4] H. Najafloo, "Examining the Impact of Artificial Intelligence on Marketing and Branding Strategies in Small and Medium-Sized Enterprises," presented at the Third International Congress on Management, Economics, Humanities, and Business Development, 2024/02/06-2024/02/07, 2024.
- [5] V. Serbin and O. Yashkina, "Systematization of Research on the Use of Artificial Intelligence in the Transformation of Marketing Strategies of E-Commerce Enterprises," *Marketing and Digital Technologies*, vol. 8, no. 4, pp. 92-109, 2024, doi: 10.15276/mdt.8.4.2024.8.
- [6] D. Ozay, M. Jahanbakht, A. Shoomal, and S. Wang, "Artificial Intelligence (AI)-Based Customer Relationship Management (CRM): A Comprehensive Bibliometric and Systematic Literature Review with Outlook on Future Research," *Enterprise Information Systems*, vol. 18, no. 7, p. 2351869, 2024, doi: 10.1080/17517575.2024.2351869.
- [7] P. Pokala, "Artificial Intelligence in SAP S/4hana: Transforming Enterprise Resource Planning Through Intelligent Automation," *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, vol. 10, no. 6, pp. 191-201, 2024, doi: 10.32628/cseit24106169.
- [8] M. Mahmoudi Meymand and N. Mohseni Fard, "The effect of generative artificial intelligence adoption on organizational performance in small and medium-sized enterprises," *Journal of New Research Approaches in Management and Accounting*, vol. 8, no. 28, pp. 2449-2463, 2024. [Online]. Available: <https://majournal.ir/index.php/ma/article/view/2662>.
- [9] K. Karimi and E. Mahmoudi Renani, "Factors Affecting the Acceptance of Artificial Intelligence in E-Commerce by Small and Medium-Sized Enterprises," (in English), *Quarterly Journal of New Approaches in Management and Marketing*, vol. 4, no. 1, pp. 62-83, 2025.
- [10] V. B. Kochkodan and M. Y. Petryna, "Risks of Implementing and Using Artificial Intelligence by Oil and Gas Sector Enterprises," *Actual Problems of Regional Economy Development*, vol. 2, no. 21, pp. 333-343, 2025, doi: 10.15330/apred.2.21.333-343.
- [11] L. Mariotti, V. Guidetti, F. Mandreoli, A. Belli, and P. Lombardi, "Combining Large Language Models With Enterprise Knowledge Graphs: A Perspective on Enhanced Natural Language Understanding," *Frontiers in Artificial Intelligence*, vol. 7, 2024, doi: 10.3389/frai.2024.1460065.
- [12] J. A. Ola-Oluwa, "Impact of Artificial Intelligence (AI) in Enhancing Knowledge Sharing and Boosting Organizational Efficiency in Nigerian Enterprises," *Ajmbr*, vol. 17, no. 1, pp. 76-95, 2024, doi: 10.62154/ajmbr.2024.017.010470.
- [13] C. Hu, Q. M. U. Din, and A. Tahir, "Artificial Intelligence Symbolic Leadership in Small and Medium-Sized Enterprises: Enhancing Employee Flexibility and Technology Adoption," *Systems*, vol. 13, no. 4, p. 216, 2025, doi: 10.3390/systems13040216.
- [14] Y. Zhu, "Research on the Influence of Artificial Intelligence Technology in Enterprise Human Resource Management on Employee Performance," *Frontiers in Business Economics and Management*, vol. 15, no. 3, pp. 116-119, 2024, doi: 10.54097/gghb5k57.

- [15] U. S. Chigozie, O. C. David, and N. E. Nworie, "Leveraging Artificial Intelligence (AI) for Business Sustainability: A Small and Medium Scale Enterprises Dimension," *International Journal of Public Administration and Management Research*, vol. 11, no. 1, pp. 28-38, 2025.
- [16] P. Farmanesh, N. Solati Dehkordi, A. Vehbi, and K. Chavali, "Artificial Intelligence and Green Innovation in Small and Medium-Sized Enterprises and Competitive-Advantage Drive toward Achieving Sustainable Development Goals," *Sustainability*, vol. 17, no. 5, p. 2162, 2025, doi: 10.3390/su17052162.
- [17] Y. Shang, S. Zhou, D. Zhuang, J. Żywiłek, and H. Dincer, "The impact of artificial intelligence application on enterprise environmental performance: Evidence from microenterprises," *Gondwana Research*, vol. 131, pp. 181-195, 2024, doi: 10.1016/j.gr.2024.02.012.
- [18] M. N. Saeed, "Fostering Sustainable Productive Entrepreneurs by Green Transformational Leadership Efficiency and Adoption of Artificial Intelligence-Tools in Saudi Micro and Small Enterprises," 2025, doi: 10.21203/rs.3.rs-8041181/v1.