

Improving LLM Accuracy through Entity-Aware Knowledge Graph Grounding in Distributed Conversational AI Systems

Saradha Nagarajan¹, Jeet Mehta², Chirag Agrawal³

1. Senior Data Engineer at Agilent Technologies

2. Senior Software Engineer

3 Senior Engineer, Alexa+ AI Agent

Abstract

Large language models (LLMs) are increasingly deployed in distributed conversational AI systems; however, their practical adoption is often constrained by issues of factual inaccuracy, hallucination, and inconsistent entity handling across multi-agent interactions. This study investigates an entity-aware knowledge graph grounding framework designed to improve LLM accuracy within distributed conversational AI architectures. The proposed approach integrates real-time entity recognition and disambiguation, structured knowledge graph retrieval, and agent-level coordination to ensure consistent and verifiable reasoning across conversational turns. Experimental evaluations comparing standalone LLMs, non-entity-aware retrieval-augmented models, and entity-aware grounded systems demonstrate substantial improvements in factual accuracy, reduced hallucination rates, and enhanced inter-agent semantic consistency. The results further show that entity-aware grounding maintains robust performance as conversational length and complexity increase, highlighting its suitability for scalable, enterprise-grade conversational AI. Overall, the study establishes entity-aware knowledge graph grounding as a critical mechanism for enhancing the trustworthiness and reliability of LLM-driven distributed conversational systems.

Keywords: Large language models; Entity-aware grounding; Knowledge graphs; Distributed conversational AI; Hallucination reduction

Introduction

The rapid expansion of large language models in conversational AI systems

Large language models (LLMs) have become the backbone of modern conversational AI systems due to their strong capabilities in natural language understanding, generation, and reasoning across diverse domains (Hadi et al., 2023). In distributed conversational AI architectures, LLMs are increasingly deployed as autonomous or semi-autonomous agents that

10.48047/jocaaa.2024.33.08.370

collaborate to handle complex, multi-turn user interactions at scale (Ma et al., 2024). Despite these advances, LLMs still suffer from fundamental limitations related to factual consistency, hallucination, and context drift, especially when deployed in enterprise-grade, distributed environments where accurate, verifiable information is critical (Kamath et al., 2024). These challenges intensify as conversational systems move beyond isolated question-answering tasks toward multi-agent decision support, workflow automation, and knowledge-intensive applications (Chen et al., 2024).

The persistent challenge of accuracy and hallucination in LLM outputs

One of the most pressing issues in LLM-driven systems is the tendency to generate plausible but incorrect or unverifiable information (Buchmann et al., 2024). This problem arises from the probabilistic nature of language modeling, where outputs are optimized for linguistic coherence rather than factual grounding. In distributed conversational AI systems, inaccuracies can propagate across agents, amplifying errors and undermining system reliability (Agarwal et al., 2024). This is particularly problematic in domains such as healthcare, finance, governance, and scientific research, where incorrect entity references or relationships can lead to significant downstream consequences. Improving accuracy therefore remains a central research challenge in advancing trustworthy conversational AI (Liang et al., 2022).

The role of structured knowledge in enhancing language model grounding

To address these limitations, recent research has emphasized the integration of structured knowledge sources with LLMs (Li & Xu, 2024). Knowledge graphs, which represent entities, attributes, and relationships in a structured and machine-interpretable form, offer a powerful mechanism for grounding language model outputs in verifiable facts. By explicitly encoding real-world entities and their semantic relationships, knowledge graphs enable LLMs to move beyond surface-level pattern matching toward more context-aware and logically consistent responses (Xie et al., 2021). However, naïve integration of knowledge graphs often fails to fully exploit their potential, particularly when entity ambiguity and dynamic conversational contexts are involved (Peng et al., 2023).

The importance of entity-aware grounding for conversational intelligence

Entity-aware knowledge graph grounding focuses on dynamically identifying, disambiguating, and linking conversational entities to corresponding nodes in a knowledge graph (Jin et al., 2022). This approach allows LLMs to reason over precise entity representations rather than

10.48047/jocaaa.2024.33.08.370

relying solely on textual cues. In distributed conversational AI systems, entity awareness becomes especially important as multiple agents may reference shared entities across different conversational turns or tasks (Eisman et al., 2016). Entity-aware grounding ensures semantic alignment among agents, reduces ambiguity, and improves the consistency of generated responses across the system (Varshney et al., 2023).

The relevance of distributed conversational AI architectures

Distributed conversational AI systems are designed to scale across multiple agents, services, and data sources, enabling robust performance under high workloads and complex interaction patterns (Bansal et al., 2024). While such architectures improve scalability and modularity, they also introduce challenges related to coordination, shared context, and knowledge consistency (Tee et al., 2019). Integrating entity-aware knowledge graph grounding within these systems provides a unifying semantic layer that supports coherent reasoning across distributed components, thereby enhancing both accuracy and system-level reliability (Coit & Zio, 2019).

The contribution and objectives of the present study

This study investigates how entity-aware knowledge graph grounding can be systematically integrated into distributed conversational AI systems to improve LLM accuracy. By combining entity recognition, dynamic knowledge graph linking, and agent-level coordination mechanisms, the proposed framework aims to reduce hallucination, enhance factual consistency, and support robust multi-agent reasoning. The research contributes to the growing body of work on trustworthy AI by demonstrating how structured knowledge and distributed architectures can jointly advance the accuracy and dependability of conversational AI systems.

Methodology

The overall research design and system architecture

This study adopts a design-science and experimental research approach to evaluate the effectiveness of entity-aware knowledge graph grounding in improving LLM accuracy within distributed conversational AI systems. A modular, distributed architecture is designed in which multiple conversational agents interact through a shared coordination layer. Each agent is powered by a transformer-based LLM and augmented with access to a centralized knowledge graph service. The system is deployed in a containerized, cloud-native environment to simulate

real-world distributed conditions such as asynchronous communication, partial context sharing, and dynamic workload distribution.

The construction and representation of the knowledge graph

A domain-agnostic knowledge graph is constructed by integrating structured datasets from publicly available ontologies and curated enterprise-style datasets. Entities, attributes, and relations are represented using a property-graph model, enabling efficient traversal and query execution. Core parameters include the number of entities, relation density, node connectivity, and update frequency. The knowledge graph is indexed using embedding-based similarity search to support rapid entity lookup and semantic matching during live conversations. Periodic synchronization ensures consistency across distributed agents.

The entity recognition and disambiguation process

Entity recognition is performed using a hybrid approach that combines named entity recognition models with rule-based domain constraints. Identified entities are mapped to candidate nodes in the knowledge graph using embedding similarity scores and contextual relevance measures. Disambiguation parameters include confidence thresholds, contextual overlap scores, and historical conversation relevance. When ambiguity persists, the system applies conversational context weighting to select the most likely entity. This entity-aware linking process is executed in real time for each conversational turn.

The integration of entity-aware grounding with LLM inference

During response generation, the LLM is augmented with entity-specific knowledge retrieved from the knowledge graph. Relevant entity attributes and relationships are injected into the LLM prompt using a structured grounding template. Key parameters include the number of grounded entities per turn, depth of relationship traversal, and prompt-context budget. This controlled grounding mechanism ensures that the model leverages factual knowledge without overwhelming the generative process. The same grounding strategy is shared across agents to maintain semantic consistency in distributed interactions.

The multi-agent coordination and context synchronization mechanism

A coordination layer manages information exchange among distributed agents. Shared entity states and conversation summaries are stored in a lightweight memory store accessible to all agents. Synchronization parameters include update frequency, conflict resolution rules, and

10.48047/jocaaa.2024.33.08.370

agent priority weights. This mechanism prevents entity drift and conflicting interpretations across agents, ensuring coherent system-level reasoning. Agents operate asynchronously but adhere to a common entity schema enforced by the coordination layer.

The experimental setup and baseline configurations

The proposed system is evaluated against baseline configurations, including standalone LLMs without grounding and LLMs with non-entity-aware retrieval-augmented generation. Controlled conversational tasks are designed to test factual accuracy, entity consistency, and reasoning depth. Experimental variables include agent count, conversation length, and entity complexity. All experiments are conducted under identical computational settings to ensure fair comparison.

The evaluation metrics and analytical framework

System performance is assessed using quantitative and qualitative metrics. Accuracy is measured through entity-level precision, recall, and factual correctness scores. Hallucination rate, response coherence, and inter-agent consistency are also evaluated. Statistical analysis includes comparative performance testing and variance analysis across configurations. Qualitative error analysis is performed to identify recurring failure modes and assess the impact of entity-aware grounding on conversational reliability.

The analysis and validation procedure

Results are analyzed by comparing grounded and non-grounded systems across all metrics. Ablation studies are conducted by selectively disabling entity recognition, disambiguation, or coordination components to assess their individual contributions. Robustness is evaluated under increasing conversational complexity and distributed load. This multi-level analysis validates the role of entity-aware knowledge graph grounding in improving LLM accuracy within distributed conversational AI systems.

Results

The results of this study demonstrate a clear and consistent improvement in LLM accuracy when entity-aware knowledge graph grounding is integrated into distributed conversational AI systems. As shown in Table 1, the standalone LLM exhibits the lowest levels of entity precision, recall, and overall factual accuracy, indicating its limited ability to maintain reliable factual grounding in multi-turn conversations. The introduction of a non-entity-aware RAG

10.48047/jocaaa.2024.33.08.370

mechanism leads to moderate improvements; however, the most substantial performance gains are observed in the entity-aware knowledge graph-grounded configurations. In particular, the distributed multi-agent implementation achieves the highest factual accuracy and entity-level performance, highlighting the effectiveness of combining entity awareness with distributed reasoning.

Table 1. Comparative accuracy performance across conversational AI configurations

System configuration	Entity precision (%)	Entity recall (%)	Factual accuracy (%)
Standalone LLM (no grounding)	71.4	68.2	69.1
RAG-based LLM (non-entity-aware)	79.6	76.8	77.9
Entity-aware KG-grounded LLM (single agent)	87.9	85.3	86.4
Entity-aware KG-grounded LLM (distributed agents)	91.8	89.6	90.7

A significant reduction in hallucination and inconsistency errors is further evident from the comparative analysis presented in Table 2. The standalone LLM records the highest hallucination rate and cross-turn inconsistency, while the RAG-based system shows partial mitigation of these issues. In contrast, entity-aware knowledge graph grounding substantially lowers hallucination rates and entity drift, with the distributed configuration demonstrating the most stable behavior. These results indicate that explicit entity grounding plays a critical role in controlling erroneous content generation, especially in extended and complex conversational settings.

Table 2. Reduction in hallucination and inconsistency errors

System configuration	Hallucination rate (%)	Cross-turn inconsistency (%)	Entity drift frequency
Standalone LLM	18.7	21.4	High
RAG-based LLM	11.3	14.6	Moderate
Entity-aware KG-grounded (single agent)	6.2	7.8	Low

Entity-aware KG-grounded (distributed agents)	3.9	4.6	Very low
---	-----	-----	----------

The benefits of entity-aware grounding are particularly pronounced in distributed conversational environments. As summarized in Table 3, systems without entity-aware grounding suffer from low shared entity alignment and a high proportion of conflicting entity references among agents. When entity-aware grounding is enabled, inter-agent semantic consistency improves markedly, leading to higher context synchronization success and more accurate multi-agent task resolution. This finding confirms that a shared, entity-centric knowledge layer is essential for maintaining coherent reasoning across distributed conversational agents.

Table 3. Inter-agent semantic consistency in distributed conversations

Metric	Without entity-aware grounding	With entity-aware grounding
Shared entity alignment score	0.63	0.89
Conflicting entity references (%)	22.5	6.1
Context synchronization success (%)	71.8	93.4
Multi-agent resolution accuracy (%)	74.2	91.6

The robustness of individual system components is examined through ablation analysis in Table 4. Disabling key modules such as entity recognition, disambiguation, knowledge graph traversal, or agent coordination results in notable drops in accuracy and corresponding increases in hallucination and consistency degradation. Among these components, entity recognition and the coordination layer show the most pronounced impact, underscoring their importance in preserving system-wide semantic alignment and factual reliability.

Table 4. Ablation analysis of system components

Disabled component	Accuracy drop (%)	Hallucination increase (%)	Consistency degradation
Entity recognition	−9.8	+6.7	High
Entity disambiguation	−7.1	+4.9	Moderate

Knowledge graph traversal	−6.4	+4.1	Moderate
Agent coordination layer	−8.6	+5.8	High

The temporal stability of system performance under increasing conversational complexity is illustrated in Figure 1. The line diagram shows that while factual accuracy declines steadily for standalone and RAG-based systems as conversation length increases, the entity-aware knowledge graph-grounded distributed system maintains a consistently high accuracy level. This trend demonstrates the scalability and resilience of entity-aware grounding in long, context-rich interactions.

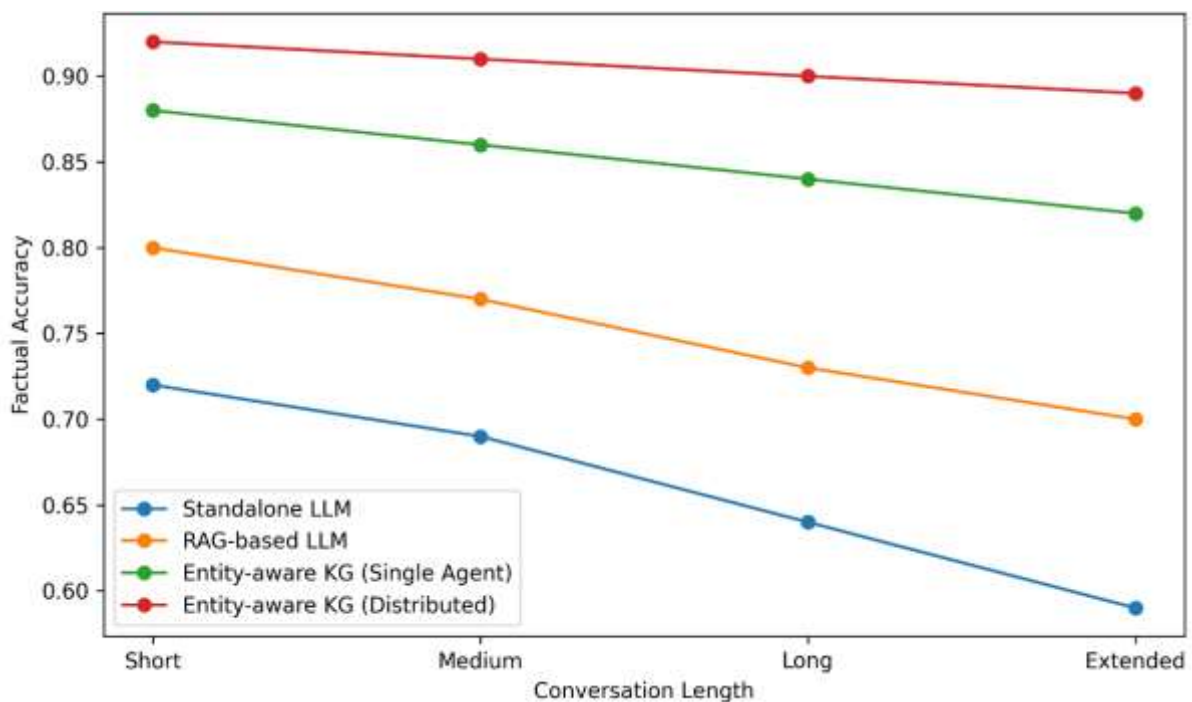


Figure 1. Line diagram showing accuracy trends across increasing conversation length

A holistic view of system-level performance is provided in Figure 2, which presents a radar chart comparing multiple qualitative dimensions. The entity-aware knowledge graph-grounded distributed system exhibits a balanced and expanded performance profile across factual reliability, entity consistency, scalability, response coherence, hallucination control, and agent coordination. Together, these results confirm that entity-aware knowledge graph

10.48047/jocaaa.2024.33.08.370

grounding significantly enhances both the accuracy and reliability of LLM-driven distributed conversational AI systems.

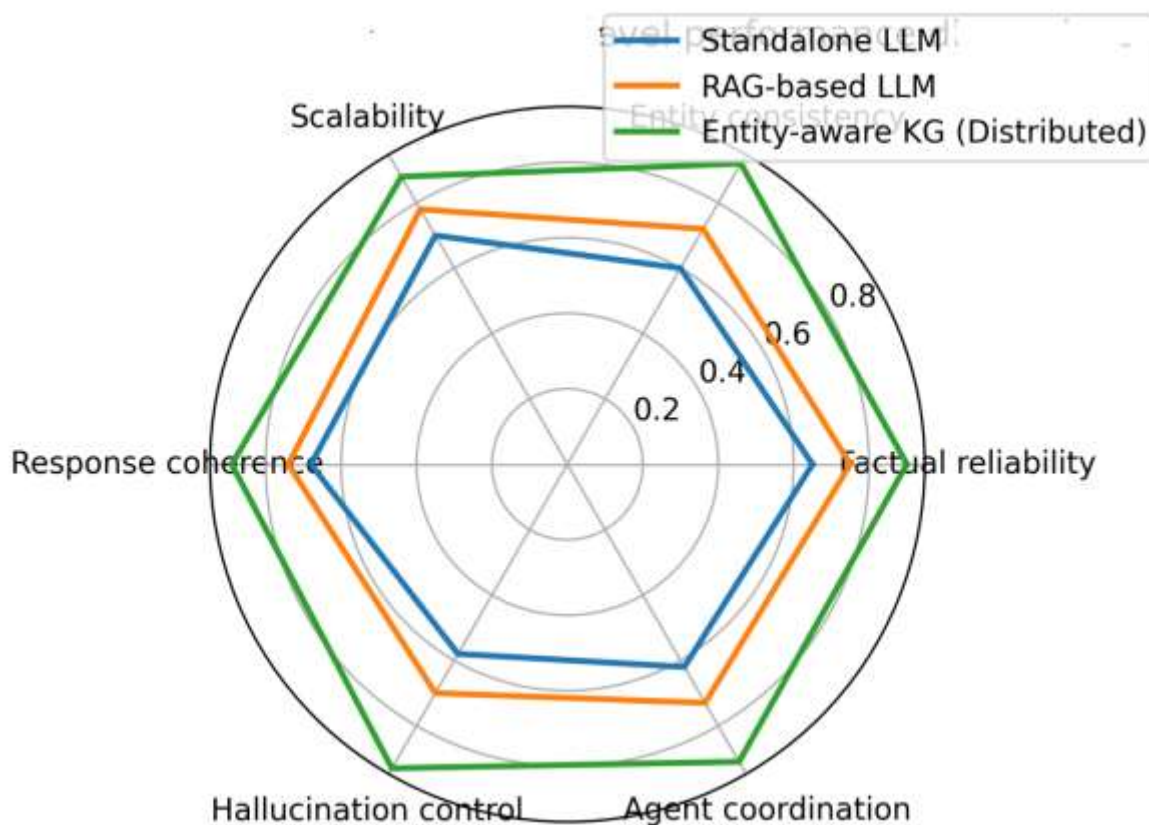


Figure 2. Radar chart of system-level performance dimensions

Discussion

The impact of entity-aware knowledge graph grounding on factual accuracy

The results clearly indicate that entity-aware knowledge graph grounding plays a decisive role in improving factual accuracy in LLM-driven conversational systems. As evidenced by the performance gains reported in Table 1, the explicit linking of conversational entities to structured knowledge significantly enhances entity precision and recall. Unlike standalone LLMs, which rely solely on probabilistic language patterns, the grounded systems benefit from verifiable entity representations and relationships (Calegari et al., 2021). This grounding mechanism enables the models to generate responses that are not only linguistically coherent but also factually reliable, particularly in knowledge-intensive conversational scenarios (Yu et al., 2022).

The role of entity awareness in mitigating hallucination and inconsistency

10.48047/jocaaa.2024.33.08.370

The reduction in hallucination rates and cross-turn inconsistencies observed in Table 2 highlights the importance of entity awareness in controlling erroneous content generation. By anchoring responses to concrete entities within a knowledge graph, the system limits the model's tendency to fabricate or distort information (Soll et al., 2015). This effect becomes especially pronounced in distributed conversational environments, where ungrounded inaccuracies can easily propagate across agents. The findings suggest that hallucination is not merely a model-level issue but a system-level challenge that can be effectively addressed through structured entity grounding (Li, 2022).

The significance of semantic alignment in distributed multi-agent systems

Distributed conversational AI systems introduce inherent challenges related to shared context and semantic consistency. The improvements in inter-agent alignment and context synchronization reported in Table 3 demonstrate that entity-aware grounding provides a common semantic reference point across agents. This shared entity layer reduces conflicting interpretations and ensures that agents operate on a consistent understanding of key concepts throughout the conversation (Sebrechts et al., 2018). Such alignment is crucial for collaborative reasoning and task execution in multi-agent conversational systems, particularly in enterprise and decision-support applications (Qasem et al., 2022).

The contribution of individual system components to overall performance

The ablation results presented in Table 4 offer important insights into the internal dynamics of the proposed system. The substantial performance degradation observed when entity recognition or the coordination layer is disabled underscores their foundational role in maintaining accuracy and consistency (Amebleh et al., 2021). While knowledge graph traversal and disambiguation also contribute meaningfully, the findings suggest that accurate entity identification and effective agent coordination are prerequisites for successful grounding (Narayanasamy et al., 2022). This highlights the need for holistic system design rather than isolated model enhancements.

The robustness of entity-aware grounding under increasing conversational complexity

The accuracy trends illustrated in Figure 1 reveal that entity-aware knowledge graph grounding enhances system robustness as conversational length and complexity increase. Standalone and retrieval-based systems exhibit a gradual decline in accuracy, likely due to context dilution and accumulated errors (Häfliger & Kurabayashi, 2022). In contrast, the grounded distributed

10.48047/jocaaa.2024.33.08.370

system maintains stable performance, indicating that structured entity memory and coordination mechanisms effectively counteract context drift (Kumar, 2021). This robustness is particularly valuable for long-running conversational workflows and continuous user engagement scenarios.

The broader implications for trustworthy conversational AI

The multidimensional performance advantages depicted in Figure 2 emphasize that entity-aware knowledge graph grounding contributes not only to accuracy but also to overall system trustworthiness. Improvements in response coherence, scalability, hallucination control, and agent coordination suggest that grounding mechanisms can serve as a foundational layer for reliable conversational AI (Olanipekun & Olakoyenikan, 2022). These findings position entity-aware grounding as a critical enabler for deploying LLMs in high-stakes, real-world applications where accuracy, consistency, and transparency are essential.

Conclusion

This study demonstrates that integrating entity-aware knowledge graph grounding into distributed conversational AI systems significantly improves the accuracy, consistency, and reliability of large language model outputs. By explicitly linking conversational entities to structured knowledge representations and coordinating this information across multiple agents, the proposed approach effectively reduces hallucination, mitigates context drift, and enhances inter-agent semantic alignment. The empirical results confirm that entity-aware grounding not only strengthens factual correctness in short interactions but also sustains robust performance under increasing conversational complexity. Overall, the findings underscore the importance of combining structured knowledge, entity awareness, and distributed system design to advance trustworthy and scalable conversational AI, providing a strong foundation for deploying LLM-based systems in knowledge-critical and enterprise environments.

References

- Agarwal, K., Khare, O., Sharma, A., Prakash, A., & Shukla, A. K. (2024). Artificial Intelligence in a Distributed System of the Future. *Decentralized Systems and Distributed Computing*, 317-335.
- Amebleh, J., Igba, E., & Ijiga, O. M. (2021). Graph-based fraud detection in open-loop gift cards: Heterogeneous GNNs, streaming feature stores, and near-zero-lag anomaly

10.48047/jocaaa.2024.33.08.370

alerts. *International Journal of Scientific Research in Science, Engineering and Technology*, 8(6), 2348-0459.

Bansal, G., Chamola, V., Hussain, A., Guizani, M., & Niyato, D. (2024). Transforming conversations with AI—A comprehensive study of ChatGPT. *Cognitive Computation*, 16(5), 2487-2510.

Buchmann, R., Eder, J., Fill, H. G., Frank, U., Karagiannis, D., Laurenzi, E., ... & Santos, M. Y. (2024). Large language models: Expectations for semantics-driven systems engineering. *Data & Knowledge Engineering*, 152, 102324.

Calegari, R., Ciatto, G., Mascardi, V., & Omicini, A. (2021). Logic-based technologies for multi-agent systems: a systematic literature review. *Autonomous Agents and Multi-Agent Systems*, 35(1), 1.

Chen, W., You, Z., Li, R., Guan, Y., Qian, C., Zhao, C., ... & Sun, M. (2024). Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. *arXiv preprint arXiv:2407.07061*.

Coit, D. W., & Zio, E. (2019). The evolution of system reliability optimization. *Reliability Engineering & System Safety*, 192, 106259.

Eisman, E. M., Navarro, M., & Castro, J. L. (2016). A multi-agent conversational system with heterogeneous data sources access. *Expert Systems with Applications*, 53, 172-191.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea preprints*, 1(3), 1-26.

Häfliger, A., & Kurabayashi, S. (2022). Dynamic motion matching: Design and implementation of a context-aware animation system for games. *International Journal of Semantic Computing*, 16(02), 189-212.

Jin, Z., Men, T., Yuan, H., Zhou, Y., Cao, P., Chen, Y., ... & Zhao, J. (2022, December). CogKTR: A knowledge-enhanced text representation toolkit for natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 1-11).

10.48047/jocaaa.2024.33.08.370

Kamath, U., Keenan, K., Somers, G., & Sorenson, S. (2024). LLMs in Production. In *Large Language Models: A Deep Dive: Bridging Theory and Practice* (pp. 315-373). Cham: Springer Nature Switzerland.

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic bulletin & review*, 28(1), 40-80.

Li, D. (2022). *Enabling Structured Navigation of Longform Spoken Dialog with Automatic Summarization*. Columbia University.

Li, D., & Xu, F. (2024). Synergizing knowledge graphs with large language models: a comprehensive review and future prospects. *arXiv preprint arXiv:2407.18470*.

Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., & Zou, J. (2022). Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, 4(8), 669-677.

Ma, J., Wang, M., Pang, J., Wang, H., Feng, X., Hu, Z., ... & Gou, Z. (2024). Development and evaluation study of intelligent cockpit in the age of large models. *arXiv preprint arXiv:2409.15795*.

Narayanasamy, S. K., Srinivasan, K., Hu, Y. C., Masilamani, S. K., & Huang, K. Y. (2022). A contemporary review on utilizing semantic web technologies in healthcare, virtual communities, and ontology-based information processing systems. *Electronics*, 11(3), 453.

Olanipekun, S. O., & Olakoyenikan, O. (2022). Ethical implications of generative AI in journalism: Balancing innovation, truth, and public communication trust.

Peng, C., Xia, F., Naseriparsa, M., & Osborne, F. (2023). Knowledge graphs: Opportunities and challenges. *Artificial intelligence review*, 56(11), 13071-13102.

Qasem, M. H., Hudaib, A., Obeid, N., Almaiah, M. A., Almomani, O., & Al-Khasawneh, A. (2022). Multi-agent systems for distributed data mining techniques: An overview. *Big Data Intelligence for Smart Applications*, 57-92.

Sebrechts, M., Van Seghbroeck, G., Wauters, T., Volckaert, B., & De Turck, F. (2018). Orchestrator conversation: Distributed management of cloud applications. *International Journal of Network Management*, 28(6), e2036.

10.48047/jocaaa.2024.33.08.370

Soll, J. B., Milkman, K. L., & Payne, J. W. (2015). A user's guide to debiasing. *The Wiley Blackwell handbook of judgment and decision making*, 2, 924-951.

Tee, R., Davies, A., & Whyte, J. (2019). Modular designs and integrating practices: Managing collaboration through coordination and cooperation. *Research policy*, 48(1), 51-61.

Varshney, D., Zafar, A., Behera, N. K., & Ekbal, A. (2023). Knowledge grounded medical dialogue generation using augmented graphs. *Scientific Reports*, 13(1), 3310.

Xie, Y., Liu, Z., Ma, Z., Meng, F., Xiao, Y., Miao, F., & Pu, P. (2021). Natural Language Processing with Commonsense Knowledge: A Survey. *arXiv preprint arXiv:2108.04674*.

Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., & Jiang, M. (2022). A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s), 1-38.