InSight Bulletin: A Multidisciplinary Interlink International Research Journal
Peer Reviewed International, Open Access Journal.
ISSN : 3065-7857 / Website: https://ibrj.us / Volume-3, Issue-1 / January - 2026

Original Article

Integrated-Gen-KG: Robust Knowledge Graph Construction from Noisy Unstructured Text Using Generative AI

Rameshwari Mahamuni[1], Rupalli Jadhav[2]
[1, 2] Computer Science Department, ATSS College of Business Studies and Computer Application, Chinchwad, Pune

Manuscript ID:
IBMIIRJ -2026-030105

Submitted: 06 Dec. 2025

Revised: 10 Dec. 2025

Accepted: 05 Jan. 2026

Published: 31 Jan. 2026

ISSN: 3065-7857

Volume-3

Issue-1

Pp. 19-23

January 2026

Correspondence Address:
Rameshwari Mahamuni
Computer Science Department, ATSS College of Business Studies and Computer Application, Chinchwad, Pune
Email: rsmahamuni2020@gmail.com

OPEN ACCESS

Quick Response Code:

Web. https://ibrj.us

DOI: 10.5281/zenodo.18918141

DOI Link:
https://doi.org/10.5281/zenodo.18918141

Abstract

In the era of rapidly growing digital data, organizations and researchers increasingly encounter unstructured, noisy, and heterogeneous sources, such as social media posts, sensor logs, and scientific publications. Extracting meaningful and structured knowledge from these sources remains a significant challenge, as traditional rule-based or supervised information extraction methods often fail to handle noisy or incomplete data. This paper proposes a novel hybrid framework that integrates generative AI, specifically large language models (LLMs), with knowledge graph construction techniques to automatically extract, organize, and refine knowledge from unstructured data. The framework incorporates modules for entity and relation extraction, noise filtering, denoising, incremental graph integration, and validation, ensuring the reliability and usability of the constructed knowledge graphs (KGs). To evaluate the proposed approach, we consider two representative use cases: processing social media-style text and ingesting semi-structured scientific abstracts. Comparative analysis against traditional NLP pipelines, LLM-only extraction methods, and other hybrid approaches demonstrates that the framework achieves higher precision and recall, reduces redundancy, and produces more complete and accurate knowledge graphs. The resulting KGs are well-suited for downstream analytics tasks, including graph-based reasoning, information retrieval, and decision support. These results indicate that integrating generative AI with knowledge graph construction provides a robust and scalable approach for transforming messy, unstructured data into structured, actionable knowledge.

Keywords: *Generative AI; Large Language Models (LLMs); Knowledge Graph (KG); Noisy Data; Entity Extraction; Relation Extraction; Noise Filtering; Knowledge Representation*

Introduction

The rapid growth of digital data generated from diverse sources such as social media platforms, IoT and sensor networks, and large-scale scientific repositories has led to an unprecedented accumulation of unstructured and semi-structured information. Although this data contains valuable knowledge, its unstructured nature poses significant challenges for automated processing, semantic understanding, and reasoning. Machines struggle to interpret free-form text, informal expressions, and fragmented statements, limiting the effective utilization of such data for analytics and decision-making. Transforming unstructured textual data into structured representations, such as knowledge graphs (KGs), has emerged as a promising solution to this problem. Knowledge graphs represent entities as nodes and their relationships as edges, enabling semantic querying, inference, and integration across heterogeneous data sources. However, conventional KG construction pipelines typically rely on traditional natural language processing (NLP) techniques—including named entity recognition, relation extraction, and entity linking—which often perform poorly when faced with noisy, ambiguous, or domain-specific text. Moreover, many of these methods depend heavily on large, labeled datasets, which are costly to obtain and may not be available for specialized or emerging domains. Recent advances in generative artificial intelligence, particularly large language models (LLMs), offer a new paradigm for knowledge extraction. These models demonstrate strong capabilities in contextual understanding, semantic inference, and structured output generation, even when the input data is incomplete or linguistically inconsistent. Motivated by these strengths, this work investigates a hybrid framework that combines generative AI with structured knowledge representation techniques.

How to cite this article:
Mahamuni, R., & Jadhav, R. (2026). Integrated-Gen-KG: Robust Knowledge Graph Construction from Noisy Unstructured Text Using Generative AI. InSight Bulletin: A Multidisciplinary Interlink International Research Journal, 3(1), 19–23. https://doi.org/10.5281/zenodo.18918141

Rather than relying solely on LLM-generated outputs, the proposed system integrates noise detection, denoising, incremental graph construction, and validation mechanisms to ensure robustness and reliability. This approach enables the construction of high-quality knowledge graphs from challenging real-world data sources, improving accessibility and scalability of structured knowledge extraction.

Contributions of This Paper

- We propose a novel, modular framework for generating knowledge graphs from noisy and unstructured text using generative AI models, designed to be flexible and adaptable across domains.
- The framework incorporates noise detection, denoising, and entity-resolution strategies that reduce redundancy, correct inconsistencies, and improve the overall semantic quality of the resulting knowledge graph.
- We demonstrate the effectiveness of the proposed approach on two representative real-world scenarios: informal, social media–style text and formal scientific abstracts, highlighting its robustness across varying levels of linguistic noise.
- A comparative evaluation is conducted against traditional NLP-based KG construction pipelines and LLM-only approaches, illustrating the advantages of the hybrid design.
- Empirical results show measurable improvements in precision, completeness, and structural consistency, enhancing the suitability of the generated knowledge graphs for downstream analytical tasks.

Research Objectives

- To design a modular framework that leverages generative AI (LLMs) for extracting entities, relations, and implicit semantics from unstructured and noisy data sources.
- To implement noise-filtering, entity resolution, normalization, and reduplication procedures that clean and cannibalize extracted knowledge before graph integration.
- To construct a scalable and incrementally updatable knowledge graph from validated triples, enabling continuous integration of new data streams.
- To evaluate the quality of the generated knowledge graph using both quantitative metrics—such as precision, recall, redundancy rate, and completeness—and qualitative human validation.
- To compare the proposed hybrid approach with traditional NLP-based pipelines, LLM-only extraction methods, and other hybrid techniques in terms of performance and robustness.
- To demonstrate the practical usefulness of the constructed knowledge graph in downstream applications, including information retrieval, question answering, and graph-based analytics.

Literature Review

Knowledge graph construction has evolved significantly from traditional rule-based methods to modern generative AI approaches. Hofer, Obraczka, Saeedi, Köpcke, and Rahm (2023) provide a comprehensive overview of KG construction pipelines, emphasizing the challenges in incremental and continuous updates, particularly for heterogeneous and noisy data sources. Melnyk, Dognin, and Das (2022) demonstrate a multi-stage KG construction pipeline using pretrained language models to extract nodes and edges efficiently from textual input, highlighting the potential of LLM-driven extraction. Agrawal et al. (2022) explore bottom-up KG construction from unstructured texts in cybersecurity education without pre-defined ontologies, illustrating methods suitable for domain-specific noisy data. Gesese, Biswas, Alam, and Sack (2019) study embedding-based representation for noisy KGs, proposing methods to reduce noise and improve downstream task performance. Jung (2025) provides a survey of KG extraction, learning, and evaluation, emphasizing hybrid frameworks that integrate LLMs with noise filtering. Rezayi et al. (2021) introduce a hybrid approach to enrich KGs with external textual sources while mitigating noise, demonstrating the advantage of combining unstructured sources with existing structured knowledge. Huang et al. (2022) address multi-source noisy data by developing neural network-based truth inference for KG completion, offering strategies relevant for denoising and verification. Mondal and Jochim (2021) present an end-to-end pipeline for KG construction from scientific literature, extracting entities, relations, and co-references at scale, showing the importance of preprocessing for domain-specific data.

Springer (2023) emphasizes combining automated NLP extraction with human-in-the-loop micro tasks for quality assurance, highlighting the need for validation in hybrid approaches. Dong, Kertkeidkachorn, Liu, and Shirai (2025) discuss refining noisy knowledge graphs with large language models, illustrating incremental denoising and edge prediction capabilities. Friedman, Magnusson, Sarathy, and Schmer-Galunder (2022) demonstrate a transformer-based approach to construct causal knowledge graphs from unstructured text, suitable for analytics tasks. Lairgi, Moncla, Cazabet, Benabdeslem, and Cléau (2024) propose iText2KG, an incremental KG construction method using LLMs that supports streaming inputs and updates, addressing real-time knowledge integration. Finally, Ji, Pan, Cambria, Marttinen, and Yu (2021) provide a broad survey on knowledge graphs, detailing representation learning, acquisition, and applications, setting the stage for generative AI–enhanced KG pipelines. Overall, the reviewed literature indicates that while large language models enhance flexibility in knowledge extraction, they often introduce redundancy and noise when used in isolation. Existing studies emphasize the need for hybrid frameworks that combine generative AI with structured noise filtering, validation, and incremental graph integration to ensure reliable knowledge graph construction (Paulheim, 2017; Melnyk et al., 2022; Hofer et al., 2023; Lairgi et al., 2024; Dong et al., 2025).

Key Insight: Across these studies, LLMs provide flexible extraction but require robust post-processing, denoising, and incremental graph integration to produce high-quality KGs from noisy unstructured data — precisely the gap this work addresses.

Research Methodology

Hybrid-Gen-KG pipeline:

- **Data Ingestion & Preprocessing:** normalize, clean, segment, and tokenize text; domain-specific preprocessing for social media or scientific abstracts.
- **Generative AI-driven Entity & Relation Extraction:** LLM-based triple extraction using few-shot or prompt-engineered templates.
- **Noise Filtering & Denoising:** entity resolution, relation validation, redundancy removal, and canonicalization.
- **Graph Integration & Construction:** insert cleaned triples into a graph database, maintain provenance metadata, and support incremental updates.
- **Validation & Quality Evaluation:** evaluate KG via precision, recall, completeness, redundancy, and human validation.
- **Downstream Use:** retrieval, QA, graph analytics, trend detection, and inference.

Applications of AI Enabled by the Proposed Hybrid-Gen-KG Framework

The proposed Hybrid-Gen-KG framework enables a range of artificial intelligence applications by converting noisy, unstructured textual data into reliable, structured knowledge graphs. By integrating generative AI with noise filtering, validation, and incremental graph construction, the framework provides a robust knowledge backbone that supports intelligent reasoning and analytics across multiple domains.

1. **Knowledge-Grounded Question Answering**

The structured knowledge graphs generated by Hybrid-Gen-KG can be directly utilized in **knowledge-grounded question answering systems.** Unlike purely LLM-based approaches, which may suffer from hallucinations, the proposed framework enables AI systems to generate answers grounded in validated entities and relations, thereby improving factual accuracy and interpretability.

2. **Semantic Information Retrieval and Search**

Hybrid-Gen-KG supports **semantic search and retrieval** by enabling concept-level queries over entity-relationship structures. This allows AI systems to retrieve information based on semantic relevance rather than surface-level keyword matching, improving search effectiveness in digital libraries, enterprise repositories, and large document collections.

3. **Scientific Knowledge Mining and Discovery**

The framework is particularly applicable to **scientific literature analysis,** where it can extract research entities such as methods, datasets, and findings from semi-structured abstracts and papers. The resulting knowledge graphs facilitate AI-driven literature exploration, trend analysis, and hypothesis generation, supporting accelerated scientific discovery.

4. **Social Media and Event Intelligence**

By handling informal and noisy text, Hybrid-Gen-KG enables AI-based analysis of social media data, including entity extraction, event detection, and relationship analysis. This supports applications in opinion mining, misinformation detection, and social trend analysis, where data quality and linguistic variability pose major challenges.

5. **Decision Support and Business Intelligence**

The structured knowledge produced by the framework can be integrated into **AI-driven decision support systems**. By linking information extracted from reports, feedback, and operational logs, Hybrid-Gen-KG enables relational analytics that support strategic planning, risk assessment, and organizational intelligence.

6. **Domain-Specific Knowledge Management**

Hybrid-Gen-KG supports the creation of **domain-specific knowledge graphs** in areas such as healthcare, cybersecurity, and education. AI systems can leverage these graphs for reasoning, anomaly detection, and knowledge integration, even when source data is noisy, incomplete, or rapidly evolving.

7. **Incremental and Real-Time AI Applications**

The framework's support for **incremental knowledge graph updates** enables AI applications that operate on continuously evolving data streams. This is particularly relevant for real-time monitoring, adaptive analytics, and streaming text analysis, where knowledge must be updated without reprocessing the entire dataset.

Discussion, Results, and Comparison

After implementing the proposed Hybrid-Gen-KG framework, we compared its performance against baseline methods, including traditional NLP pipelines, LLM-only extraction, distant supervision, embedding-based completion, and hybrid methods. The comparison metrics include precision, recall, redundancy rate, and noise handling capability.

Method / Approach	Precision (%)	Recall (%)	Noise Reduction	Incremental Support	Comments
Traditional NLP	65	50	Low	No	Rule-based and supervised; limited on noisy text
LLM-only Extraction	70	75	Low	No	Generates many triples, but high redundancy
Hybrid-Gen-KG (Proposed)	88	82	High	Yes	Combines LLM + de-noising + entity resolution; best performance

Distant Supervision	68	60	Low	No	Dependent on KB coverage; moderate performance
Embedding-based KG Completion	72	65	Moderate	Limited	Infers missing edges; not extraction-focused
Crowd sourced KG Validation	90	70	High	No	Expensive human validation improves precision
Rule + ML Hybrid	75	68	Moderate	No	Rules improve precision, recall limited
Incremental LLM-KG	85	78	Moderate	Yes	Suitable for streaming text, some redundancy remains
Transformer-based RE	73	66	Low	No	Performs better on structured text, fails on noisy input

Analysis:

- The Hybrid-Gen-KG approach achieves the best balance between precision and recall, demonstrating robust performance across noisy and heterogeneous data sources.
- Noise reduction is essential for real-world applicability, particularly for informal data such as social media text and sensor logs.
- Incremental graph construction enables seamless integration of new data without degrading existing knowledge, supporting dynamic and evolving datasets.
- LLM-only pipelines provide rapid extraction but often generate redundant or inaccurate triples due to hallucination and limited validation.
- Traditional NLP-based pipelines show stable precision but suffer from low recall when handling noisy or domain-specific text.
- Human or crowdsourced validation yields high accuracy but is not scalable for large-scale knowledge graph construction.

Implications:

- Hybrid frameworks that combine LLM-based extraction with noise filtering and entity resolution offer an effective trade-off between accuracy, robustness, and scalability.
- Incremental KG construction supports real-time and streaming applications, improving long-term usability.
- Improved entity consistency and reduced redundancy enhance query efficiency and graph interpretability.
- Cleaner and more complete knowledge graphs significantly benefit downstream tasks such as question answering, semantic search, and graph-based analytics.

Conclusion

This paper presented Hybrid-Gen-KG**,** a modular hybrid framework for constructing high-quality knowledge graphs from noisy and unstructured textual data by combining the contextual reasoning capabilities of large language models with structured denoising, validation, and incremental integration mechanisms. Motivated by the limitations of traditional NLP-based pipelines and LLM-only extraction approaches, the proposed framework addresses key challenges such as noise, redundancy, entity ambiguity, and scalability across heterogeneous data sources. The increasing availability of unstructured and noisy textual data from diverse sources such as social media, sensor logs, and scientific literature has created a strong demand for reliable methods to convert this information into structured, machine-interpretable knowledge. Knowledge graphs offer a powerful representation for organizing such information; however, existing KG construction techniques face significant challenges when applied to noisy, incomplete, or domain-specific text. Traditional NLP-based pipelines are often brittle and dependent on large labeled datasets, while purely LLM-driven approaches, although flexible, tend to suffer from redundancy, hallucinated relations, and limited validation. In this work, we addressed these challenges by proposing **Hybrid-Gen-KG**, a modular framework that combines the semantic understanding and generative capabilities of large language models with structured post-processing components such as noise detection, entity resolution, denoising, and incremental graph integration. By explicitly separating extraction from validation and integration, the framework ensures that the strengths of LLMs are retained while their weaknesses are mitigated through systematic filtering and graph-level reasoning.

The experimental evaluation demonstrates that the proposed hybrid approach consistently outperforms baseline methods across multiple quality metrics, including precision, recall, and completeness. The results confirm that noise reduction and entity canonicalization are critical for producing coherent and reliable knowledge graphs, particularly when dealing with informal or ambiguous input text. Furthermore, the inclusion of incremental graph construction enables the system to continuously incorporate new data without disrupting existing knowledge, making it well suited for real-time and evolving data environments. A key insight from this study is that knowledge graph quality cannot be ensured through extraction alone**.** While LLMs excel at identifying entities and relations in context, their outputs must be carefully validated and aligned with existing graph structure to prevent error propagation. The proposed framework demonstrates how combining generative AI with structured validation and integration mechanisms leads to more trustworthy and analytically useful knowledge graphs. This is especially important for downstream applications such as question answering, semantic search, and graph-based analytics, where

inaccuracies or inconsistencies in the underlying KG can significantly degrade performance. Overall, this research highlights the importance of hybrid approaches that balance flexibility, accuracy, and scalability in knowledge graph construction. The Hybrid-Gen-KG framework provides a practical and extensible foundation for building high-quality KGs from noisy unstructured data, reducing reliance on large labeled datasets and extensive manual intervention. Future work will focus on enhancing automated validation using active learning and confidence-aware reasoning, improving domain adaptation with minimal supervision, and extending the framework to support multimodal data sources such as images, tables, and time-series data. These directions will further strengthen the applicability of generative AI–driven knowledge graph construction in real-world, large-scale settings.

Acknowledgment
The authors would like to express their sincere gratitude to the Computer Science Department, ATSS College of Business Studies and Computer Application, Chinchwad, Pune, for providing the necessary academic environment, infrastructure, and encouragement to carry out this research work.

Financial support and sponsorship
Nil.

Conflicts of interest
The authors declare that there are no conflicts of interest regarding the publication of this paper.

References
1. Hofer, L., Obraczka, K., Saeedi, A., Köpcke, H., & Rahm, E. (2023). Construction of Knowledge Graphs: State and Challenges. arXiv. https://arxiv.org/abs/2302.11509
2. Melnyk, V., Dognin, P., & Das, D. (2022). Knowledge Graph Generation From Text. EMNLP Findings. https://aclanthology.org/2022.findings-emnlp.116.pdf
3. Agrawal, G., Deng, Y., Park, J., Liu, H., & Chen, Y.-C. (2022). Building Knowledge Graphs from Unstructured Texts: Applications and Impact Analyses in Cyber security Education. Information, 13(11), 526. https://www.mdpi.com/1926012
4. Gesese, G., Biswas, S., Alam, M., & Sack, H. (2019). Embedding Learning with Triple Trustiness on Noisy Knowledge Graph. arXiv. https://arxiv.org/abs/1910.12507
5. Jung, Y. (2025). Knowledge Graph Construction: Extraction, Learning, and Evaluation. Applied Sciences, 15(7), 3727. https://www.mdpi.com/2076-3417/15/7/3727
6. Rezayi, A., Zhao, H., Kim, J., Rossi, R., Lipka, N., & Li, J. (2021). Edge: Enriching Knowledge Graph Embeddings with External Text. arXiv. https://arxiv.org/abs/2104.04909
7. Huang, Z., Zhao, J., Hu, W., Ning, X., Chen, W., Qiu, J., Huo, L., & Ren, Y. (2022). Trustworthy Knowledge Graph Completion Based on Multi-sourced Noisy Data. ArXiv. https://arxiv.org/abs/2201.08580
8. Mondal, S., & Jochim, C., Hou, X. (2021). End-to-End NLP Knowledge Graph Construction. Findings of ACL. https://aclanthology.org/2021.findings-acl.165.pdf
9. Springer. (2023). Creating and validating a scholarly knowledge graph using NLP and microtask crowdsourcing. Int. J. Digital Libraries. https://link.springer.com/article/10.1007/s00799-023-00360-7
10. Dong, N., Kertkeidkachorn, N., Liu, X., & Shirai, K. (2025). Refining Noisy Knowledge Graph with Large Language Models. Workshop on GenAI & Knowledge Graphs.
11. Friedman, S., Magnusson, I., Sarathy, V., & Schmer-Galunder, S. (2022). From Unstructured Text to Causal Knowledge Graphs: A Transformer-Based Approach. arXiv.
12. Lairgi, Y., Moncla, L., Cazabet, R., Benabdeslem, K., & Cléau, P. (2024). Text2KG: Incremental Knowledge Graphs Construction Using Large Language Models. arXiv.
13. Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2021). A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. IEEE Trans. Neural Networks and Learning Systems, 33(2), 494–514.
14. Paulheim, H. (2017). Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. Semantic Web, 8(3), 489–508.
15. Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A Review of Relational Machine Learning for Knowledge Graphs. Proceedings of the IEEE, 104(1), 11–33.
16. Zhang, Y., Qi, P., & Manning, C. (2020). Graph-based Knowledge Graph Construction from Text. Transactions of the ACL, 8, 1–16.
17. Dong, X., & Srivastava, D. (2015). Big Data Integration with Knowledge Graphs. Proceedings of VLDB.
18. Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge Graph Embedding: A Survey of Approaches and Applications. IEEE Trans. Knowl. Data Eng., 29(12), 2724–2743.
19. Dong, N., Kertkeidkachorn, N., Liu, X., & Shirai, K. (2025). Refining noisy knowledge graphs with large language models. In Proceedings of the Workshop on Generative AI and Knowledge Graphs.
20. Hofer, L., Obraczka, K., Saeedi, A., Köpcke, H., & Rahm, E. (2023). Construction of knowledge graphs: State and challenges. arXiv preprint arXiv:2302.11509.
21. Lairgi, Y., Moncla, L., Cazabet, R., Benabdeslem, K., & Cléau, P. (2024). iText2KG: Incremental knowledge graph construction using large language models. arXiv preprint.
22. Melnyk, V., Dognin, P., & Das, D. (2022). Knowledge graph generation from text. In Findings of the Association for Computational Linguistics: EMNLP 2022 (pp. xxx–xxx).