AI-Driven Knowledge Externalisation: From Unstructured Documents to Structured Data Models

Dilyan Georgiev and Elissaveta Gourova
Sofia University St. Kliment Ohridski, Bulgaria

diljang@uni-sofia.bg
elis@fmi.uni-sofia.bg

Abstract: The transformation of unstructured textual information into structured, database-ready knowledge represents a critical challenge in contemporary information systems. Organisations accumulate vast amounts of documentation-technical specifications, project reports, procedural guidelines – yet much of this knowledge remains inaccessible to computational processing because it is non-structured. This paper proposes and analyses a method for artificial intelligence (AI)-driven knowledge mapping, whereby generative models are used to extract, interpret, and populate predefined data schemas from textual documents. The approach involves three key stages: (1) ingestion of a source document; (2) prompt-based semantic extraction aligned with a target database schema; and (3) structured object generation ready for persistence in relational or NoSQL environments. Unlike traditional information extraction techniques that rely on rigid rule-based systems or supervised training pipelines, the proposed method leverages prompt-engineered large language models (LLMs) to interpret context and map semantic content to structured attributes dynamically. This enables automated form completion, metadata generation, and structured representation of domain knowledge. The paper critically examines the reliability, consistency, and traceability of AI-generated mappings. While generative AI significantly reduces manual effort and accelerates data formalisation, challenges remain in ensuring validation, schema compliance, and epistemic transparency. The research, therefore, discusses validation mechanisms, confidence scoring, and human-in-the-loop verification strategies. By framing AI-driven extraction as a process of knowledge externalisation, the study positions this method within broader Knowledge Management (KM) theory. The implications extend beyond data automation: such mapping mechanisms enable scalable knowledge codification, improved interoperability between systems, and enhanced analytical capabilities. The findings suggest that AI-based structured extraction may redefine how organisations formalise expertise, shifting from document-centric storage toward schema-driven knowledge architectures.

Keywords: Knowledge management, Knowledge externalisation, AI-driven method

1. Introduction

The exponential growth of digital information within organisations has not been matched by equivalent advances in how that information is operationalised (Pai et al., 2024). A substantial proportion of organisational knowledge continues to reside in unstructured formats-documents, emails, reports, technical descriptions, and informal communications. While these artefacts are rich in context and meaning, they remain largely inaccessible to automated systems that depend on structured data representations (Castro et al., 2024; Xu et al., 2024). As a result, organisations face a persistent and costly bottleneck: the manual transformation of unstructured textual content into structured, database-ready formats.

This transformation process is inherently time-consuming and error-prone. Employees must interpret context, extract relevant fields, and match them to schemas, which creates a high cognitive load (Dunn et al., 2022). Even in relatively simple scenarios, such as transferring information from a CV into a standardised form, the process demands attention, repetition, and verification. In more complex, domain-specific contexts-such as financial documentation, regulatory filings, or technical specifications-the cognitive load increases significantly, introducing higher risks of inconsistency, omission, or misinterpretation (Han et al., 2024).

Beyond inefficiency, this reliance on manual processes constrains scalability. As organisations accumulate larger volumes of textual data, the gap between stored knowledge and actionable data widens. Critical insights remain locked within documents, limiting their integration into analytics pipelines, decision-support systems (DSS), and automated workflows. Consequently, organisations are unable to fully leverage their informational assets, undermining both operational efficiency and strategic responsiveness.

Traditional data transformation approaches offer limited relief in this context. Established methods, such as Extract, Transform, Load (ETL) pipelines, are designed to operate on already structured or semi-structured data (Kimball and Ross, 2013). They presuppose clearly defined schemas and predictable input formats, conditions that do not hold for free-form textual content. Attempts to apply rule-based extraction techniques to unstructured data often result in brittle systems that fail to generalise across document types or linguistic variations (Pai et al., 2024).

Recent advances in generative AI, particularly LLM, present a fundamentally different approach to this problem. Modern models demonstrate the ability to semantically extract information and generate structured outputs, even from highly unstructured texts (Xu et al., 2024). This shift opens new possibilities for automating the conversion of unstructured knowledge into structured representations.

However, the promise of AI-driven extraction must be approached critically. While generative models can significantly reduce the time and effort required for data transformation, they are not immune to errors. Issues such as hallucination, contextual misinterpretation, and domain inconsistency pose risks to data quality and reliability (Han et al., 2024). Full automation, therefore, is neither realistic nor desirable in high-stakes environments.

This paper addresses these challenges by proposing a method for AI-driven knowledge externalisation that combines semantic extraction with schema alignment and human validation. By integrating generative models into a structured workflow, the approach aims to reduce manual effort while preserving epistemic integrity. The objective is not merely to automate data entry, but to redefine how organisations transform and operationalise their unstructured knowledge, bridging the gap between narrative information and structured digital systems.

2. State-of-the-art in Research

The transformation of unstructured textual data into structured, machine-readable representations is a long-standing research problem in the fields of natural language processing (NLP), information extraction (IE), and KM. In recent years, this problem has gained additional importance due to the exponential growth of digitally generated data and the need for its efficient operationalisation in information systems (Yang et al., 2022). Organisations store significant volumes of knowledge in unstructured formats such as documents, emails, reports, and technical descriptions. Despite their rich semantic content, these sources remain difficult to use for automated systems that require structured input data. This creates a significant gap between the available knowledge and its operational usability, leading to high costs for manual processing and transformation (Hogan et al., 2021; Castro et al., 2024).

Early approaches to information extraction were based on rules and symbolic methods, including patterns and linguistic grammars. These methods formed the basis of tasks such as named entity recognition and relation extraction, but were limited by low adaptability and heavy reliance on manual engineering (Maynard et al., 2017; Nasar et al., 2021). With the advent of statistical models and machine learning, greater flexibility was achieved, but these methods remained dependent on large annotated corpora. Deep learning significantly improved performance by modelling context and semantic dependencies but required significant resources and suffered from domain shift (Kumar et al., 2017; Devlin et al., 2019; Zhao et al., 2024).

Knowledge graphs represent an established approach for explicitly modelling knowledge and semantic relationships (Paulheim et al., 2016). Systems such as Google Knowledge Graph and NELL demonstrate large-scale efforts to automatically extract and organise information from text (Hogan et al., 2021). Traditional pipelines include multiple sequential steps – entity recognition, entity linking, and relation extraction – which makes them complex and difficult to maintain. They are often considered fragile in dynamic and heterogeneous data. Modern approaches explore the use of LLM to directly generate graph structures, which significantly simplifies the process of knowledge externalisation (Yang et al., 2022; Zhao et al., 2023).

The introduction of the transformer architecture marks a fundamental breakthrough in NLP by introducing a self-attention mechanism that allows for efficient modelling of long-term dependencies (Vaswani et al., 2017). Modern foundation models are built on this architecture, demonstrating emergent capabilities, including zero-shot and few-shot learning (Brown et al., 2020). These models significantly expand the possibilities for automated information retrieval without domain-specific training (Wei et al., 2022). A particularly important development is the ability to directly generate structured outputs from unstructured text, which transforms the task from classical retrieval to generative semantic mapping (Lewis et al., 2020).

Current research focuses on schema-guided extraction, in which models generate output that conforms to predefined structures (e.g. JSON schemas or ontologies). This significantly reduces variability and improves integration with databases and analytical systems. The combination of structured prompting and constrained decoding allows for more reliable extraction and better consistency of results (Yang et al., 2022). In this context, LLMs function as universal semantic transformers between unstructured text and formal data models (Zhao et al, 2023). In addition, Retrieval-Augmented Generation combines generative models with external knowledge sources. This allows for context expansion and reduced hallucinations, while improving factual accuracy (Lewis

et al., 2020; Izacard and Grave, 2021). Modern systems extend this concept to retrieval & reasoning and retrieval & structuring, where the retrieved information is not only generated, but also validated and formalised.

From the point of view of KM, the transformation of unstructured data can be considered through the SECI model of Nonaka, where the externalisation phase, linked to the transformation of tacit knowledge into explicit knowledge, plays a key role (Nonaka and Takeuchi, 1995). Modern AI systems automate this phase through semantic extraction and structuring, which allows for significant acceleration of organisational knowledge formalisation (Castro et al., 2024). This creates the opportunity for scalable transformation of narrative information into operationally usable structures.

Despite significant progress, LLM-based systems remain limited by issues related to reliability, stability and consistency. Empirical research shows that models can generate incorrect or incomplete structured outputs, especially in complex domain-specific scenarios (Han et al., 2024). Additionally, there are limitations in strictly structured extraction, where accuracy and repeatability requirements are high (Maynard et al., 2017). This necessitates the use of additional mechanisms such as human-in-the-loop systems, schema validation and retrieval-based verification. Last, but not least, it could be noted that the general trend in the literature shows a clear shift from classical pipeline-based systems towards unified generative models. In this context, LLMs are asserting themselves as universal semantic intermediaries that allow direct transformation of unstructured knowledge into structured digital representations, thus realising a scalable form of knowledge externalisation (Lewis et al., 2020; Xu et al., 2024).

3. Model Representation, Experiment, and Results Analysis

3.1 AI-driven Knowledge Externalisation Method

The proposed method conceptualises AI-driven knowledge externalisation as a structured pipeline that transforms unstructured textual input into schema-aligned, machine-processable data. At its core, the method relies on the interaction between three key components: (1) unstructured data sources, (2) a predefined model schema, and (3) a generative AI agent capable of semantic interpretation and structured output generation.

The process begins with the ingestion of unstructured documents: in this experimental setup, CV files containing heterogeneous and narrative-rich information (Figure 1). These documents serve as the input corpus from which relevant data must be extracted. A critical design element of the method is the construction of a comprehensive prompt that encapsulates not only the raw document content, but also the target model definition. This includes explicit schema constraints such as field names, data types, expected formats (e.g. date representations), and controlled vocabularies (e.g. enumerations for categorisation). By embedding these constraints into the prompt, the system guides the generative model toward producing outputs that are structurally compatible with downstream persistence layers.

The AI client transmits this prompt to an AI agent, which may operate either as a general-purpose language model or as a domain-enhanced system using retrieval-augmented generation. The agent processes the input and returns a structured response in JSON format. This output is then programmatically parsed and mapped onto the predefined model. A validation layer ensures that extracted values conform to schema requirements before being stored in a database. This step is essential for mitigating inconsistencies and preserving data integrity, particularly given the probabilistic nature of generative models.

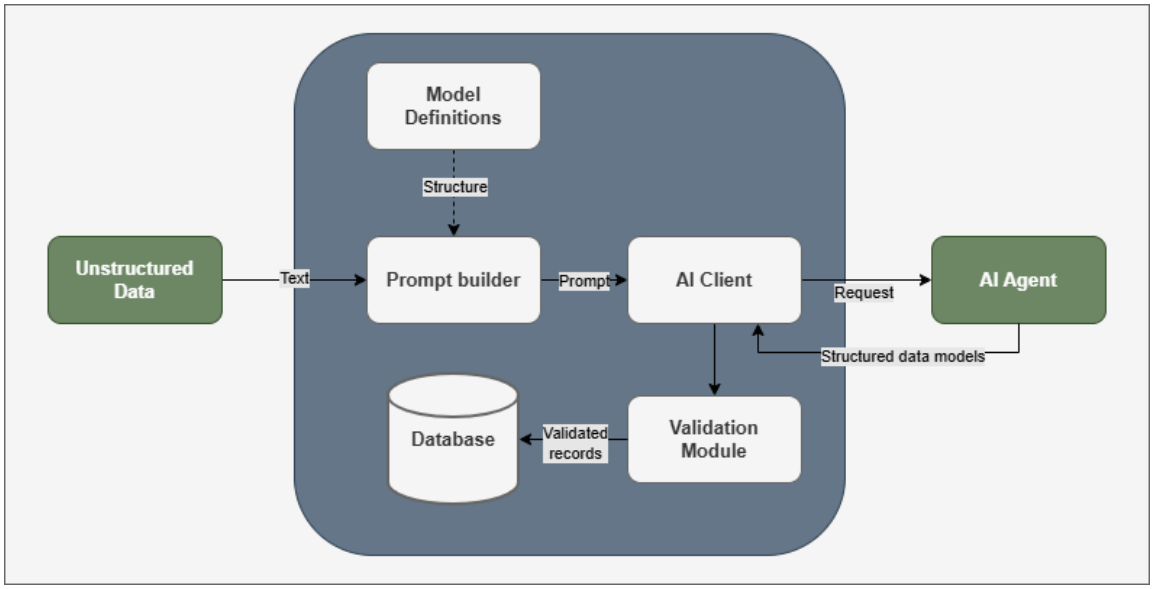

Figure 1: AI-Driven Knowledge Externalisation

3.2 Experimental Design

To evaluate the proposed AI-driven knowledge extraction method, an experimental prototype was developed in the form of a web-based application focused on CV data processing. The experiment was intentionally designed around a simple, but representative use case: transforming unstructured CV documents into structured data suitable for database storage.

The system interface consisted of a standardised form divided into four primary segments: (1) personal data, (2) education, (3) employment history, and (4) skills. Participants were asked to upload a CV in .docx format, containing unstructured textual information. Upon upload, the system automatically extracted the raw text and constructed a prompt that included three essential components: the unstructured document content, the predefined model schema corresponding to the four form sections, and additional contextual constraints such as enumerations, categorical values, and expected data formats.

This prompt was sent to an AI agent via the OpenAI API, with explicit instructions to return the output in JSON format. The returned JSON object was then programmatically parsed and mapped onto the internal model representing the form. Successfully extracted fields were used to automatically populate the corresponding sections in the user interface.

Participants were then asked to review the populated form and compare it with the original CV document. They were encouraged to identify inaccuracies, correct any errors, and ensure completeness. After validation, users proceeded to finalise the process by creating a profile based on the extracted data. Finally, participants completed a feedback form evaluating usability, accuracy, trust, efficiency, and potential application areas.

3.3 Results Analysis

The experimental results, based on 21 participants from diverse professional backgrounds, provide strong evidence supporting the effectiveness of the proposed method, while also revealing important limitations.

From a usability perspective, the system was very well received. Approximately 66.7% of participants rated the system as “very easy” to use, while an additional 23.8% found it “somewhat easy” (Figure 2). Only 4.8% reported difficulty, and 4.8% remained neutral. Instruction clarity was rated highly, with 81% assigning the maximum score and the remaining 19% indicating minor ambiguity.

Accuracy perceptions were similarly positive. Around 76% of participants rated extraction accuracy at the highest level, while 24% assigned slightly lower scores. In addition, 76% of users reported no incorrect or missing information, while 24% identified issues. This indicates that while the system performs reliably in most cases, errors are not negligible and must be addressed.

Trust levels were generally high, with approximately 86% of participants expressing strong or moderate trust (ratings 4–5). However, willingness to use the system without manual verification was significantly lower. Only

about 38% indicated full confidence in automated use, while the majority preferred either partial trust or continued validation. This gap highlights a critical insight – perceived usefulness does not provide full autonomy acceptance.

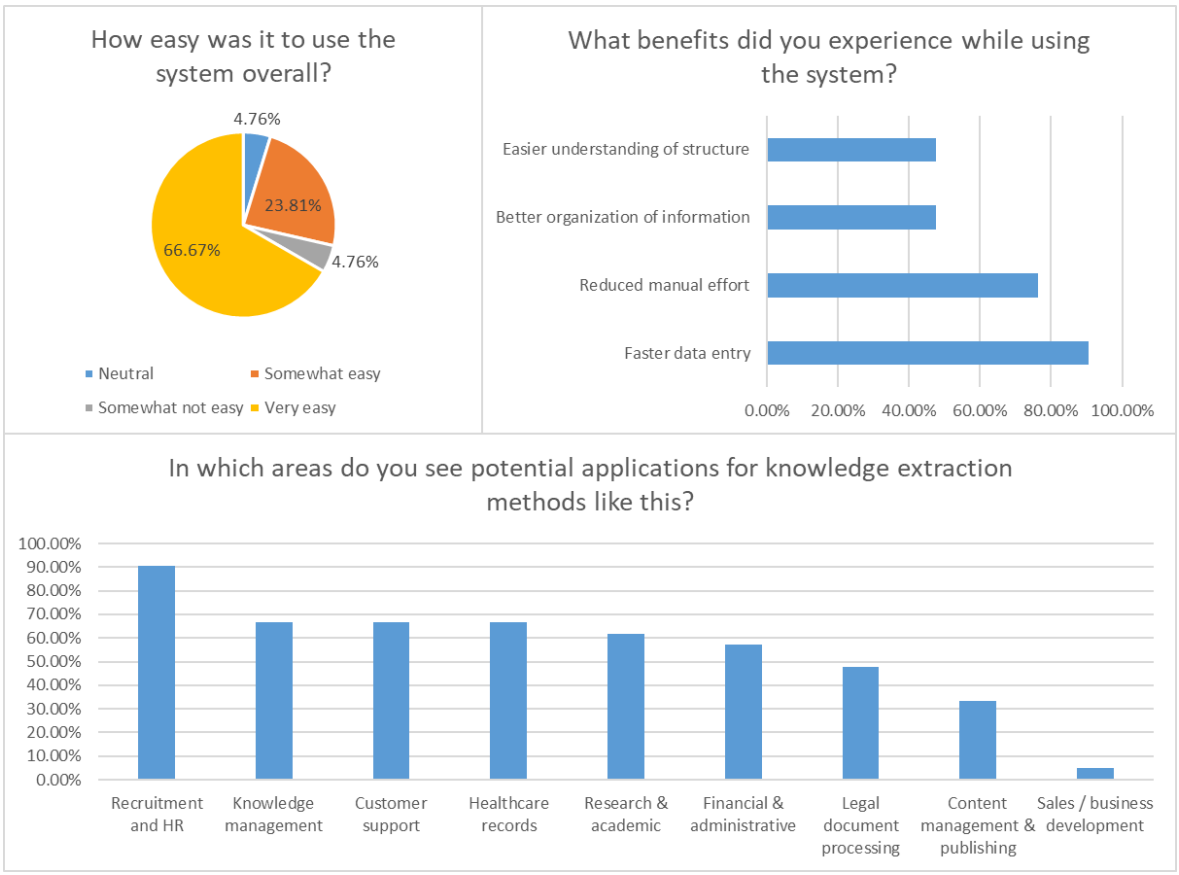

Figure 2: Results representation

Efficiency gains were the most consistently positive outcome. Approximately 76% of participants reported "significant time savings," with the remainder indicating moderate improvements. Reported benefits clustered around faster data entry (90%), reduced manual effort (76%), and improved data organisation (47%), confirming the central hypothesis of efficiency enhancement.

Future adoption intent was also strong, with 52% of participants highly likely to use such a system and the remaining 48% indicating moderate likelihood. Notably, no respondents expressed outright rejection, suggesting broad acceptance potential.

A particularly important finding emerges from the analysis of suggested application domains. The most frequently identified use case was recruitment and human resources (90%), especially CV processing and candidate screening, appearing in nearly all responses. This confirms the immediate relevance of the experimental scenario. Other highly recurring domains include KM (66%), customer support and ticket classification (66%), healthcare records processing (66%), legal (47%) and financial document handling (57%). Additionally, participants identified research data extraction and content management as valuable extensions (33%). This diversity demonstrates the generalisability of the method across industries, while also indicating that its value increases in document-intensive environments.

In conclusion, the results validate the method's effectiveness in reducing time-consuming manual work and improving structured data generation. However, they also clearly demonstrate that human validation remains essential, particularly in scenarios where accuracy is critical.

4. Discussion

The proposed method for AI-driven knowledge externalisation demonstrates applicability across a range of practical scenarios, each characterised by the need to transform unstructured input into structured outputs efficiently and reliably. The identified use cases highlight both the strengths and limitations of the approach,

particularly when considering varying levels of domain complexity and the necessity of contextual understanding.

The first use case, simple form population, represents the most immediate and low-risk application. In scenarios such as CV processing, survey responses, or generic administrative forms, the required data fields are typically well-defined and do not demand deep domain expertise. Here, AI-driven extraction can significantly reduce repetitive manual work by mapping textual inputs directly to structured fields such as names, dates, qualifications, and contact details. The primary advantage lies in efficiency gains, as well as consistency in formatting. However, even in these seemingly straightforward contexts, edge cases-such as ambiguous phrasing or incomplete information-necessitate validation mechanisms to ensure accuracy.

The second use case, domain-specific form structuring, introduces greater complexity. In contexts such as banking, insurance, or regulatory compliance, forms often require not only extraction, but also interpretation and enrichment of data. Users may lack the expertise to provide all required information explicitly, leaving gaps that must be inferred or supplemented. In such cases, integrating retrieval-augmented generation techniques becomes critical. By grounding model outputs in domain-specific knowledge bases, the system can enhance both completeness and relevance of the structured data. Nevertheless, this increased capability comes with a higher risk: incorrect enrichment or misinterpretation in regulated domains can have significant consequences. This reinforces the necessity of human-in-the-loop validation, particularly for high-impact decisions.

The third use case, automatically captured task management, extends the approach beyond static documents to dynamic data streams, such as email communication. Here, the objective is not only to extract data but to identify intent and trigger actionable workflows. For instance, an email requesting a follow-up, approval, or information retrieval can be transformed into a structured task with defined parameters, deadlines, and categories. This application illustrates the potential of AI to bridge communication and execution, reducing cognitive overhead for users. However, accurately interpreting intent remains a non-trivial challenge, especially in informal or context-dependent communication. Misclassification of tasks or incorrect prioritisation could introduce inefficiencies rather than eliminate them.

An additional and particularly impactful use case is knowledge base construction and maintenance. Organisations frequently accumulate large repositories of documents (technical manuals, project reports, meeting notes) that remain underutilised due to their unstructured nature. Applying the proposed method, these documents can be systematically transformed into structured knowledge entries, enabling advanced search, analytics, and integration with DSS. Unlike the previous use cases, this scenario emphasises long-term value creation rather than immediate operational efficiency. It also introduces challenges related to consistency across documents, version control, and evolving schemas, all of which require careful governance.

Across all use cases, a common pattern emerges: the effectiveness of AI-driven extraction depends not only on model capability, but also on system design. Prompt engineering, schema definition, and validation workflows play critical roles in determining output quality. The findings suggest that full automation is neither necessary nor optimal. A hybrid approach-combining AI efficiency with human oversight-offers the most robust solution.

In conclusion, the method demonstrates strong potential to reduce time-consuming manual processes and unlock the value of unstructured data. Its applicability spans simple administrative tasks to complex KM systems. However, its successful deployment depends on acknowledging and mitigating the limitations of generative models, ensuring that efficiency gains do not come at the expense of reliability or trust.

5. Conclusion

This paper considers the growing need to integrate AI as a core service within knowledge externalisation processes, particularly in the transformation of unstructured data into structured, machine-processable formats. As organisations continue to accumulate vast amounts of textual information, the limitations of manual data entry and traditional transformation techniques become increasingly evident. In this context, generative AI emerges not merely as a supportive technology but as a foundational component capable of interpreting, structuring, and operationalising organisational knowledge.

By proposing an AI-driven knowledge externalisation method, the study positions generative models as powerful tools applicable across a wide range of knowledge-intensive business processes. The approach demonstrates how unstructured sources, such as documents and communications, can be systematically mapped to predefined schemas, enabling their integration into structured databases and enterprise systems. This has direct implications for multiple domains, including human resources, healthcare, legal services, financial

administration, and broader KM practices. As such, the method opens new avenues for research and application, extending beyond isolated use cases toward more comprehensive, system-level integration.

A key contribution of the paper is its balanced perspective on automation. While the method significantly reduces the need for manual data entry and improves efficiency, it does not eliminate the role of human expertise and redefines it. Employees shift from performing repetitive data input tasks to assuming responsibility for validation and verification, ensuring the accuracy and reliability of AI-generated outputs. This human-in-the-loop approach is essential for maintaining trust and integrity, particularly in domain-sensitive environments.

The experimental results support the central hypothesis of the study. Participants reported substantial time savings, high usability, and generally strong accuracy, confirming both the practicality and the perceived value of the method. At the same time, the findings highlight the importance of validation mechanisms, reinforcing the conclusion that full automation, while promising, must be approached incrementally and responsibly.

Ultimately, the proposed method contributes to improving connectivity and data mapping not only between unstructured documents and structured databases but also across complex systems with heterogeneous knowledge representations. By enabling more seamless integration of information, it lays the groundwork for more adaptive, intelligent, and scalable digital infrastructures, where knowledge can flow more freely and be leveraged more effectively across organisational boundaries.

Acknowledgements

The authors gratefully acknowledge the support provided by the project UNITe BG16RFPR002-1.014-0004 funded by PRIDST.

Ethics Declaration: The study involved human participants who voluntarily completed the survey questionnaire. No personally identifiable information or sensitive personal data was collected, stored, or used during the research process, and all responses were analysed anonymously.

AI Declaration: AI tools were used in a limited capacity to support language refinement and editing. All conceptual development, analysis, and interpretation are the sole responsibility of the authors.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp.1877-1901.

Castro, A., Pinto, J., Reino, L., Pipek, P. and Capinha, C., 2024. Large language models overcome the challenges of unstructured text data in ecology. *Ecological informatics*, 82, p.102742.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, June. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies,* Vol. 1, pp. 4171-4186.

Dunn, A., Dagdelen, J., Walker, N., Lee, S., Rosen, A.S., Ceder, G., Persson, K. and Jain, A., 2022. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv*:2212.05238.

Han, R., Yang, C., Peng, T., Tiwari, P., Wan, X., Liu, L. and Wang, B., 2023. An empirical study on information extraction using large language models. *arXiv preprint arXiv*:2305.14450.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S. and Ngomo, A.C.N., 2021. Knowledge graphs. ACM Computing Surveys (Csur), 54(4), pp.1-37.

Izacard, G. and Grave, E., 2021, April. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics,* pp. 874-880.

Kimball, R. and Ross, M., 2013. *The data warehouse toolkit: The definitive guide to dimensional modeling*. John Wiley & Sons.

Kumar, S., 2017. A survey of deep learning methods for relation extraction. *arXiv preprint arXiv*:1705.03645.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel, S., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, pp.9459-9474.

Maynard, D., Bontcheva, K., Augenstein, I. (2017). Named Entity Recognition and Classification. In: *Natural Language Processing for the Semantic Web. Synthesis Lectures on Data, Semantics, and Knowledge*. Springer, Cham. https://doi.org/10.1007/978-3-031-79474-2_3

Nasar, Z., Jaffry, S.W. and Malik, M.K., 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys* (CSUR), 54(1), pp.1-39.

Nonaka, I. and Takeuchi, H., 1995. *The Knowledge-Creating Company*. Oxford University Press.

Pai, L., Gao, W., Dong, W., Ai, L., Gong, Z., Huang, S., Zongsheng, L., Hoque, E., Hirschberg, J. and Zhang, Y., 2024. A survey on open information extraction from rule-based model to large language model. *Findings of the association for computational linguistics*: EMNLP 2024, pp.9586-9608.

Paulheim, H., 2016. Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web, 8(3), pp.489-508.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. and Chi, E.H., 2022. Emergent abilities of large language models. *arXiv preprint arXiv*:2206.07682.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y. and Chen, E., 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6), p.186357.

Yang, Y., Wu, Z., Yang, Y., Lian, S., Guo, F. and Wang, Z., 2022. A survey of information extraction based on deep learning. *Applied Sciences*, 12(19), p.9691.

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. and Du, Y., 2023. A survey of large language models. *arXiv preprint arXiv*:2303.18223, 1(2), pp.1-124.

Zhao, X., Deng, Y., Yang, M., Wang, L., Zhang, R., Cheng, H., Lam, W., Shen, Y. and Xu, R., 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11), pp.1-39.