Scaling Scientific Knowledge Discovery with Neuro-Symbolic AI and Large Language Models

Wilma Johanna Schmidt[1,4], Diego Rincon-Yanez[2,6], Evgeny Kharlamov[1,4] and Adrian Paschke[3,5]

[1]Bosch Center for AI, Robert Bosch GmbH, Renningen, Germany
[2]University of Salerno, Fisciano, Italy
[3]AG Corporate Semantic Web, Freie Universität Berlin, Berlin, Germany
[4]SIRIUS, Centre for Scalable Data Access, University of Oslo, Oslo, Norway
[5]Data Analytics Center, Fraunhofer FOKUS, Berlin, Germany
[6]Universidad de Santander, Facultad de Ingenierías y Tecnologías, Cucuta, Colombia

Abstract

The increasing amount of available research data leads to the need to scale scientific knowledge discovery, e.g., the conduction of systematic literature reviews (SLRs), to keep up with fast developments in research and further support decision-making in the industry. AI-based methods are gaining importance in these tasks and have been integrated into many SLR tools. Yet, several challenges are still open on applying especially neural methods on scientific knowledge discovery tasks. To address this, we evaluate various neural and neuro-symbolic scenarios on a specific generative writing task. While confirming existing concerns on pure Large Language Model (LLM) approaches for these tasks, we obtain a heterogeneous picture of Retrieval-Augmented Generation (RAG) approaches. The most promising candidate is a Knowledge Graph (KG) based context-enhanced LLM approach for Knowledge Discovery.

Keywords

Neuro-Symbolic AI, Knowledge Graph, Large Language Model, Retrieval-Augmented Generation (RAG), Systematic Literature Review

1. Introduction

Recent AI approaches are drastically impacting solutions and the ways of working in several industries at an additional fast-ongoing development pace. Yet, at least two trends are expected to remain predictable: (i) the high, even increasing need for fast decision-making and (ii) the continuously increasing amount of available data to make decisions. This is highly reflected in the growing research field of data-driven decision-making.

Large language models (LLMs) are a novel generative AI approach that shows promising results on various industrial challenges, yet LLMs tend to encounter limitations on reliability [1] and interpretability [2] [1]. Fortunately, e.g., smart prompting techniques may "*enhance the model's ability to explain their reasoning and justify their decision*" [2]. With context-enhanced prompts, LLMs can be more strongly guided toward suitable responses. The versatility and capability of LLMs mark a paradigm shift in how we interact with machines, making these interactions more intuitive and resembling human-like conversations. However, a notable challenge with LLMs is their occasional tendency to produce information not rooted in reality or their training data, a phenomenon often termed "hallucinations" [3] [1]. To mitigate these hallucinations, the concept of Retrieval Augmented Generation (RAG) has arisen as the ability of the LLM to analyze text with the capacity to retrieve relevant information from selected external sources; this enhances the accuracy and reliability of the produced answer.

On the other hand, neuro-symbolic AI, as a combination of neural and symbolic methods [4], positions itself as a promising candidate for industrial applications[5]. One benefit of neuro-symbolic solutions

First International Workshop on Scaling Knowledge Graphs for Industry, co-located with 20th International Conference on Semantic Systems (SEMANTICS) - Amsterdam, Sept. 17–19, 2024
Wilma.Schmidt@de.bosch.com (W. J. Schmidt); drinconyanez@unisa.it (D. Rincon-Yanez); Evgeny.Kharlamov@de.bosch.com (E. Kharlamov); adrian.paschke@fu-berlin.de (A. Paschke)
0000-0002-8982-1678 (D. Rincon-Yanez); 0000-0003-3247-4166 (E. Kharlamov); 0000-0003-3156-9040 (A. Paschke)

includes the integration of domain knowledge [6], e.g., in the form of Knowledge Graphs (KGs) [1]. Integrating KGs as a structured and symbolic knowledge representation into RAG-type applications offers a powerful approach to addressing the challenge of reducing the hallucinations[7] by combining the ability of language models to analyze text with the capability to retrieve relevant information from external sources, such as knowledge bases.

Nowadays, it is virtually impossible to keep track of new research, considering the overload in scientific publications worldwide [8]. Research needs to support the decision-making process at an industrial scale, meaning the engineering of scientific knowledge and discovery that comes with the necessity of analyzing a massive corpus of data. There are established methods in research that can be applied for systematic analyses of a large landscape of publications, such as a *Systematic Literature Review (SLR)* [9]. Yet, SLRs are time-consuming if conducted manually. AI methods have shown to be effective for increasing efficiency, such as paper selection[10], yet recent research has not fully exploited these capabilities [10]. Specifically, LLMs open up new steps to automate SLRs further with knowledge representation and smart prompting. While some open challenges in scientific knowledge discovery are addressed by AI-based techniques [10], neuro-symbolic approaches have not been explicitly assessed on their potential and limitations in this field.

Considering this potential, this paper identifies the benefits and limitations of different approaches for scientific knowledge discovery, specifically answering research questions of an SLR. We evaluate LLM-based and neuro-symbolic, specifically document-based RAG and RDF-KG-based context-enhanced LLM-based approaches. Additionally, a prompt engineering process was conducted based on different neuro-symbolic approaches drafted as systematic experimentation scenarios.

Moreover, this work tackles the missing transparency on proprietary SLR tools with AI support ([2] [10]); For this reason and unpredictability concerns, a GitHub repository[1] with the used system and user prompts was prepared including different specific scientific knowledge discovery questions and the respective answers.

The further parts of this paper are structured as follows: we analyze and discuss research on the status and open challenges of AI-supported SLRs in Section 2. We present our approach in Section 3. In Section 4, we first describe the different scenarios of our experiment. Second, we show the obtained results and analyze the benefits and limitations. After discussing open challenges on scaling scientific knowledge discovery with neuro-symbolic AI in Section 5, we conclude in Section 6 and point to the limitations of our work and future steps.

2. Related Work

This section shows relevant related work on scaling scientific knowledge discovery, with a focus on neuro-symbolic AI.

One of the most prominent LLM challenges is hallucination reduction. An ML-oriented method to solve this is fine-tuning, but this comes at a high cost in terms of time and effort [11]. It is possible to develop a model that allows for the prediction of multiple tail or head entities for a given relation and entity, leveraging the relevant neighbors of the entities[12]. This has resulted in improved efficiency and effectiveness of LLMs in utilizing KG information in specialized or personalized domains. However, both cases generate new challenges, such as increased costs due to the need for fine-tuning on LLMs, although it is significantly lower than other methods since very specific, compressed, and previously validated information is mapped. An additional challenge is the risk of information loss in the graphs due to the difficulty in leveraging the most relevant neighbors because of the large number of connections a node can have.

As one example of scientific knowledge discovery, SLRs have proven valuable. An SLR consists of three main phases: *planning*, *conduction*, and *reporting*. De la Torre-López *et al.* [10] show in an SLR that most AI-based support in automating SLRs is on the *conduction* phase of SLRs, specifically the task of paper selection. Phase *planning* is semi-automated with traditional methods (see, e.g., [13] on duplicate

[1] GitHub Repository - https://github.com/d1egoprog/KG-SLR4LLM

identification), and the *reporting* phase is commonly done manually. The authors see accordingly a gap in more research on AI-driven writing tasks [10].

Bolaños *et al.* reviewed AI opportunities and challenges for literature reviews [2] by reviewing existing SLR tools. The authors stress the importance of the research direction on integrating advanced NLP technologies to replace possibly outdated methodologies in available SLR tools and the "*promising research direction*" of "*the use of semantic technologies [...]*" particularly knowledge graphs, to enhance the characterization and classification of research papers [2]. An interesting work on integrating advanced NLP technologies by Jansen *et al.* employs LLMs in survey research [14]. The authors see "*potential advantages to using LLMs like ChatGPT for survey research to generate survey responses*" and discuss potential issues such as bias and lack of contextual understanding of LLMs. Our work addresses the latter by evaluating neuro-symbolic approaches to knowledge injection.

Further work (e.g., [15] [16]) shows research interest in this field, yet still lacks research on neuro-symbolic, e.g., RAG and Graph RAG, Memory-based, to improve the reporting phase in scientific knowledge discovery.

Focused on the medical domain, Yun *et al.* [17] summarizes that "*further research is warranted for using LLMs for literature reviews in other domains as our study only focused on the task of writing medical systematic reviews.*" While van Dinter *et al.* [18] extend the domain view in their work, the focus still remained on the medical and computer science domain, leading to no SLRs evaluated from the manufacturing domain.

In summary, the related work shows interest in the AI-support for scientific knowledge discovery. The exploration focuses on SLRs as a method and general medical or computer science as a domain. To the best of our knowledge, no SLR has been conducted manually and then challenged against LLM capabilities in any way. Further, no AI-based support for SLRs started with a KG, but only on metadata of publications or texts containing the respective content of a publication. With our work, we address the previously mentioned gaps.

3. Building Neuro-Symbolic AI Frameworks for Scientific Knowledge Discovery

In this section, we describe the underlying neuro-symbolic approaches and the architectural pattern employed in our work's neuro-symbolic scenarios.

In order to address scalability in the realm of scientific knowledge discovery, we evaluate different approaches on the example of an SLR's generative writing task. In addition to the human and LLM-based responses to specific research questions, an evaluation of neuro-symbolic potentials and challenges is needed. In this section, we describe a document-based RAG approach and a framework for an RDF-KG-based context-enhanced LLM; these are the basis of the selected neuro-symbolic scenarios in our experiments.

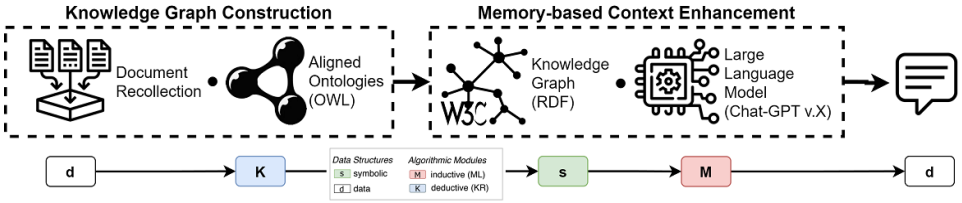


Figure 1: Neuro-Symbolic AI Enhancement approach for ingesting Knowledge Graphs into the LLM; NeuroSymbolic AI Architecture {d-K-s-M-d}, using the boxology notation [19]

Lewis *et al.* [7] introduce Retrieval-Augmented Generation (RAG) as the combination of "*pre-trained, parametric-memory generation models with a non-parametric memory through a general-purpose fine-tuning approach*". In our work, the RAG approach is based on an LLM for the parametric-memory model based on a folder of text files for the non-parametric memory, see Figure 1. The LLM is executed in scenarios with different GPT models from OpenAI[2].

[2]https://platform.openai.com

The document base contains *49* text files of the final search corpus from a recently conducted SLR [20]. Each text file was scrapped, and the text was extracted from the main publication website. With the selected document base, a Knowledge Graph construction process was performed using the extracted paper content and the paper metadata and a schema was assembled by leveraging existing ontologies such as BIBO[3], SWRC[4], ORKG[5] and others.

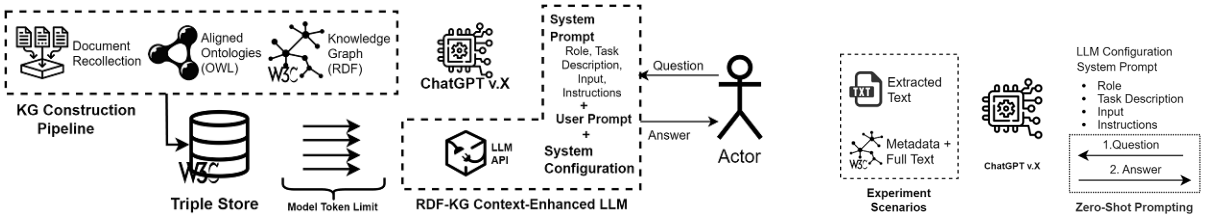
To test the RDF-KG-based context-enhanced LLM, see Figure 1, the public API of OpenAI was employed, specifically on the GPT-4-turbo model. The KG includes entities from the *49* assessed publications, authors, venues, and identified research fields. The complete publication list (*49*) can be found in the GitHub repository[6]; as well as the assembled schema and the fully populated KG.

4. Scaling Knowledge Discovery with Knowledge Graphs and Neuro-Symbolic AI

In this section, we describe the experimental framework 4.1 conducted, RAGs and RDF-KG-based context-enhanced LLMs. We conclude with the results of our experiments 4.2.

4.1. Experimental Configuration

The evaluation was centered on evaluating two approaches on LLMs, RAGs (1) and RDF-KG-based context-enhanced (2). The main goal is to scale scientific knowledge discovery as can be detailed in Figure 2a. The performed evaluation was centered on the results of five research questions (a main research question and an additional four) drafted for the selected document base. With the scope of assessing the generative writing capabilities and knowledge discovery by leveraging research questions of an SLR.



(a) RDF-KG-Based Context-Enhanced LLM (Scenario: S5); Zero-Shot Prompting

(b) RAG (Scenarios: S2, S4); Zero-Shot Prompting

Figure 2: High-Level Architecture View

To increase comparability between the LLM with no knowledge and the RAG-based approach, *GPT-3.5-turbo* and *GPT-4-turbo* were employed in both scenarios, the scenario detail is listed in Table 1. The approach on an RDF-KG-based context-enhanced LLM is conducted only on GPT-4-turbo. The GPT-4-turbo serves as the basis for the evaluation across the neural and neuro-symbolic approaches.

Table 1
Model Information

Scenario	Model	Setup
S1	gpt-3.5-turbo	temperature 0.5; zero shot
S2	gpt-3.5-turbo	temperature 0.5; zero shot; 10 message sources
S3	gpt-4-turbo	temperature 0.5; zero shot
S4	gpt-4	temperature 0.5; zero shot; 10 message sources
S5	gpt-4-turbo	temperature 0.5; zero shot; five times 9-10 message sources in context, then the summary of 5 responses in additional prompt

[3]Namespace: http://purl.org/ontology/bibo/
[4]Namespace: http://swrc.ontoware.org/ontology#
[5]Namespace: http://orkg.org/core
[6]https://github.com/wAIlma/SLR-NeSyAI-KGC-I40/data

In each scenario, five steps are undertaken, each of them addressing the research questions (RQ) from [20]: (1) *Which role play neuro-symbolic AI approaches in knowledge graph construction for Smart Manufacturing?* (Main RQ), (2) *What are publication characteristics on neuro-symbolic AI in knowledge graph construction for Smart Manufacturing?* (RQ1), (3) *In which steps of the knowledge graph construction process are neuro-symbolic AI methods applied in Smart Manufacturing?* (RQ2), (4) *What are common neuro-symbolic AI architectures in knowledge graph construction?* (RQ3), and (5) *For which manufacturing use cases are knowledge graphs constructed with neuro-symbolic AI?* (RQ4).

Considering that, the scenario *5* holds the model token constraint. Hence, the KG containing 49 documents is split into *five* SubKGs with a separate context, each, and asked to merge the five responses.

4.2. Evaluation

In this Section, we present the evaluation approach and the analysis of the results. The underlying framework of all scenarios is shown in Figure 2. The conducted LLM-based and neuro-symbolic scenarios are listed as follows:

1. **LLM only:** No further data provided. Scenarios: S1 and S3
2. **Document-based RAG** Files contain the retrieved text retrieved from the manually selected 49 publications. Scenarios: S2 and S5
3. **RDF-KG-based context-enhanced LLM** An RDF KG is provided as the context in addition to a system prompt and user prompt to LLM. Scenario: S5

Considering the lack of gold standards for evaluating an LLM response, an evaluation model was selected that reflects on the known weaknesses of LLMs and yet might not cover all requirements for answering a scientific research question. The selected evaluation criteria were adapted from [14], each with a score from *1* to *5*, on the scenarios, see Table 2.

Table 2
Evaluation Criteria.

Id	Criterion Name	Description	Score
C1	Domain-specific vocabulary	Use of neuro-symbolic- and manufacturing domain-specific vocabulary	1 (specific vocabulary not used or used in the wrong context) to 5 (specific vocabulary correctly employed)
C2	Contextual understanding (hallucination)	Degree of "nonsensical or inappropriate responses"	1 (completely inappropriate response) to 5 (appropriate response)
C3	Compelling misinformation	Share of "highly convincing text that is factually wrong"	1 (at least *50%* of response is factually wrong) to 5 (response is completely true)
C4	Lack of transparency	Degree of increasing transparency caused by "disclosing LLM participation and intractability of LLM training and the text-generation process"	1 (no or ineligible sources provided) to 5 (all relevant sources provided and all cited in-text)

We show our results in Figure 3. Based on the results, we see that scaling scientific knowledge with LLMs and improving this approach with RAGs is at an interesting yet not applicable level. On the one hand, the responses vary significantly across scenarios and research questions. On the other hand, scientific criteria are not met as hallucinations occur, and references are handled unreliably. In contrast, we obtain promising results from the RDF-KG-based context-enhanced LLM. We discuss these specific points in our next section 5.

5. Discussion

Overall, the responses across the different scenarios show a wide range from disappointing to promising answers. Some responses (e.g. *S2-RQ1*, *S2-RQ2*) do not attempt an answer although the relevant context

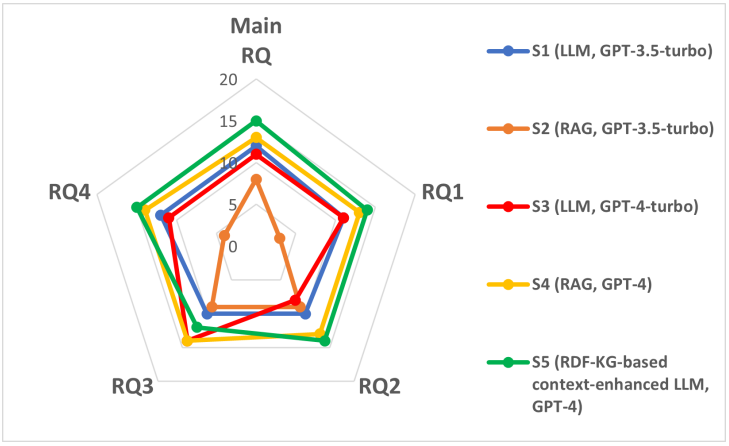


Figure 3: Results on scenarios for scaling scientific knowledge discovery

is provided via text chunks and the LLM is trained on general knowledge to at least return a more complex answer. On the contrary, one of the best answers (*S4-RQ3*) includes an outlook on evolutionary knowledge which is not explicitly requested by the prompts. Underneath the variety, at least two common flaws can be identified, that apply to all scenarios: (i) missing (references to) definitions and (ii) missing tables, charts or figures to illustrate the statements.

We see on LLM-based and RAG-based scenarios severe challenges. With consistent system prompts and varying research questions, the responses vary unexpectedly on several factors: (i) the reference list (*S1-RQ1* and *S1-RQ4* contain no references at all), (ii) in-text citations (none provided by e.g. *S4-RQ4*) and, (iii) whether the provided references are not made up (e.g. *S1-RQ2* returns a template for references with no actual values included). *S2-mainRQ* quotes directly from a provided source, yet omits quote indication and citation. The RDF-KG-based context-enhanced LLM is a promising direction, yet it also needs further improvement to ensure responses on a scientific level.

Neuro-symbolic approaches are one way of reducing hallucinations in LLMs. Our results show a good performance of *S1* and *S4*, yet a disappointing performance of *S2*. The RAG-based approach with a GPT-3.5-turbo model (*S2*) describes neuro-symbolic AI as a combination of "merits of statistical learning with semantical knowledge and reasoning", omitting the neural perspective, which is crucial.

6. Conclusion

In summary, our work shows a promising neuro-symbolic approach of an RDF-KG-based context-enhanced LLM for scaling scientific knowledge discovery. One further benefit of this approach is the foundation for handling evolutionary knowledge. Via the KG the knowledge can be updated and made available for future scientific queries to the LLM with minimal effort.

Our results show a need for caution when working with RAG-based approaches. Based on the overall results, we see that scaling scientific knowledge with LLMs and improving this approach with simple RAGs is not at an applicable level. On the other hand, scientific criteria are not met as hallucinations occur, and references are treated unreliably. RDF-KG-based context-enhanced LLMs appear to be better suited for this task based on our results, yet also require further improvements before being applicable.

Our experiment sheds light on scientific knowledge discovery from research data from the manufacturing domain yet is applicable to SLRs across industries.

Our work does not cover the whole area of scientific knowledge discovery, omitting, e.g., paper selection tasks in SLR or expert interviews as approaches.

Lastly, token processing is a costly parameter. As a research paper may contain about ten thousand tokens, processing a large data corpus quickly runs into a token issue. Smart prompting and suitable neuro-symbolic architectures are needed to address this.

In future work, we plan to evaluate different parameter configurations, especially temperature and number of message sources on RDF-KG-based context-enhanced LLMs.

Acknowledgements

We want to thank Valentin Knappich and Cem Akdag for their helpful support and insights during our work.

References

[1] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering 36 (2024) 3580–3599. doi:10.1109/TKDE.2024.3352100.

[2] F. Bolanos, A. Salatino, F. Osborne, E. Motta, Artificial Intelligence for Literature Reviews: Opportunities and Challenges (2024). arXiv:2402.08565.

[3] K. Sanderson, GPT-4 is here: what scientists think, Nature 615 (2023) 773. doi:10.1038/d41586-023-00816-5.

[4] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, L. Zhou, Neuro-symbolic approaches in artificial intelligence, National Science Review 9 (2022) nwac035. doi:10.1093/nsr/nwac035.

[5] D. Rincon-Yanez, M. H. Gad-Elrab, D. Stepanova, K. T. Tran, C. Chu Xuan, B. Zhou, E. Karlamov, Addressing the Scalability Bottleneck of Semantic Technologies at Bosch, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 13998 LNCS, 2023, pp. 177–181. doi:10.1007/978-3-031-43458-7_33.

[6] D. Yu, B. Yang, D. Liu, H. Wang, S. Pan, A survey on neural-symbolic learning systems, Neural Networks 166 (2023) 105–126. doi:10.1016/j.neunet.2023.06.028.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. T. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, Advances in Neural Information Processing Systems December (2020).

[8] E. Landhuis, Scientific literature: Information overload, Nature 535 (2016) 457–458. doi:10.1038/nj7612-457a.

[9] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering – a systematic literature review, Information and Software Technology 51 (2009) 7–15. doi:10.1016/j.infsof.2008.09.009.

[10] J. de la Torre-López, A. Ramírez, J. R. Romero, Artificial intelligence to automate the systematic review of scientific literature, Computing 105 (2023) 2171–2194. doi:10.1007/s00607-023-01181-x.

[11] N. Dziri, S. Milton, M. Yu, O. Zaiane, S. Reddy, On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?, in: NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, Association for Computational Linguistics, Stroudsburg, PA, USA, 2022, pp. 5271–5285. doi:10.18653/v1/2022.naacl-main.387.

[12] Y.-H. Lin, H.-T. Shieh, C.-Y. Liu, K.-T. Lee, H.-C. Chang, J.-L. Yang, Y.-S. Lin, Retrieval-Augmented Language Model for Extreme Multi-Label Knowledge Graph Link Prediction (2024). arXiv:2405.12656.

[13] A. Carrera-Rivera, W. Ochoa, F. Larrinaga, G. Lasa, How-to conduct a systematic literature review: A quick guide for computer science research, MethodsX 9 (2022) 101895. doi:10.1016/j.mex.2022.101895.

[14] B. J. Jansen, S.-g. Jung, J. Salminen, Employing large language models in survey research, Natural Language Processing Journal 4 (2023) 100020. doi:10.1016/j.nlp.2023.100020.

[15] A. M. Sami, Z. Rasheed, K.-K. Kemell, M. Waseem, T. Kilamo, M. Saari, A. N. Duc, K. Systä, P. Abrahamsson, System for systematic literature review using multiple AI agents: Concept and an empirical evaluation (2024). arXiv:2403.08399.

[16] B. D. Lund, T. Wang, N. R. Mannuru, B. Nie, S. Shimray, Z. Wang, ChatGPT and a new academic

reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing, Journal of the Association for Information Science and Technology 74 (2023) 570–581. doi:10.1002/asi.24750.

[17] H. S. Yun, T. A. Trikalinos, I. J. Marshall, B. C. Wallace, Appraising the Potential Uses and Harms of Large Language Models for Medical Systematic Reviews, in: EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings, Association for Computational Linguistics, Stroudsburg, PA, USA, 2023, pp. 10122–10139. doi:10.18653/v1/2023.emnlp-main.626.

[18] R. van Dinter, B. Tekinerdogan, C. Catal, Automation of systematic literature reviews: A systematic literature review, Information and Software Technology 136 (2021) 106589. doi:10.1016/j.infsof.2021.106589.

[19] F. van Harmelen, A. ten Teije, A Boxology of Design Patterns forHybrid Learning and Reasoning Systems, Journal of Web Engineering 18 (2019) 97–124. doi:10.13052/jwe1540-9589.18133.

[20] W. Schmidt, D. Rincon-Yanez, E. Kharlamov, A. Paschke, Systematic Literature Review on Neuro-Symbolic AI in Knowledge Graph Construction for Manufacturing, Semantic Web Journal TBD (2024).