

Assessing Large Language Models for SPARQL Query Generation in Scientific Question Answering

Antonello Meloni¹, Diego Reforgiato Recupero^{1,*}, Francesco Osborne², Angelo Salatino², Enrico Motta², Sahar Vahdati³ and Jens Lehmann^{3,†}

¹Department of Mathematics and Computer Science, University of Cagliari, Italy

²Knowledge Media Institute, Open University, United Kingdom

³ScaDS.AI - TU Dresden, Germany

Abstract

Scientific question answering remains a significant challenge for the current generation of large language models (LLMs) due to the requirement of engaging with highly specialised concepts. A promising solution is to integrate LLMs with knowledge graphs of research concepts, ensuring that responses are grounded in structured, verifiable information. One effective approach involves using LLMs to translate questions posed in natural language into SPARQL queries, enabling the retrieval of relevant data. In this paper, we analyse the performance of several LLMs on this task using two scientific question-answering benchmarks: SciQA and DBLP-QuAD. We explore both few-shot learning and fine-tuning strategies, investigate error patterns across different models, and propose directions for future research.

Keywords

Knowledge Graphs, Large Language Models, Machine Translation, SPARQL

1. Introduction

Answering scientific questions poses a significant challenge for current LLMs due to the need to engage with highly specialised and complex concepts. In this domain, common limitations of LLMs, such as hallucinations [1] and the “long-tail” issue [2], where LLMs struggle with rare or less frequently occurring concepts, become especially crucial. While a new generation of LLM-based systems for literature reviews and scientific writing support has emerged [3], their output still falls short of the standards expected in high-quality scientific literature.

One promising solution is the integration of LLMs with Knowledge Graphs (KGs) of research concepts, which helps ensure that responses are grounded in structured, verifiable information [4, 5]. The scientific domain benefits from a wide array of knowledge organization systems [6], such as taxonomies and ontologies of research topics, which play a crucial role in categorizing, managing, and retrieving information. Additionally, numerous knowledge graphs have been developed in this space, providing machine-readable, semantically rich, and interlinked descriptions of the content of research publications [7, 8, 9, 10]. Notable examples include the Open Research Knowledge Graph (ORKG)¹ [7], the Computer Science Knowledge Graph (CS-KG)², and Nanopublications³ [11, 9].

An effective approach for integrating LLMs with KGs involves using LLMs to translate scientific questions, posed in natural language, into SPARQL queries [12]. This allows for the retrieval of relevant data from the KG, which can either be presented directly to the user or further refined by the LLM. This

ISWC 2024 Special Session on Harmonising Generative AI and Semantic Web Technologies, November 13, 2024, Baltimore, Maryland

*Corresponding author.

†Work done outside of Amazon.

✉ antonello.meloni@unica.it (A. Meloni); diego.reforgiato@unica.it (D. R. Recupero); francesco.osborne@open.ac.uk (F. Osborne); angelo.salatino@open.ac.uk (A. Salatino); enrico.motta@open.ac.uk (E. Motta); sahar.vahdati@tu-dresden.de (S. Vahdati); jens.lehmann@tu-dresden.de (J. Lehmann)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹Open Research Knowledge Graph - <https://www.orkg.org/>

²Computer Science Knowledge Graph - <http://w3id.org/cskg/>

³Nanopublications - <https://nanopub.org/>

solution also enables less technically proficient users to query and navigate complex knowledge graphs of scientific concepts through a natural language interface.

In this paper, we evaluate the performance of several LLMs in translating scientific questions into SPARQL queries. We first conduct a comprehensive evaluation on the SciQA benchmark [13], followed by testing the best-performing methods on the DBLP-QuAD benchmark [14]. Our goal is to assess how effectively LLMs perform this task and determine whether current training or prompting methods are sufficient or if more advanced techniques are required. We explore the effects of fine-tuning and various prompting strategies, including zero-shot and few-shot learning, using different example selection methods such as semantic similarity [15] and diversity [16, 17]. Furthermore, we analyse error patterns across models to identify areas for improvement. The insights gained from this study provide a foundation for advancing the field by developing more comprehensive benchmarks and designing systems better equipped to answer complex scientific questions.

This short paper extends [18] by introducing several additional experiments. These include the evaluation of Mistral, enhanced error analysis, and the integration of the DBLP-QuAD benchmark. It should be noted that here we intentionally focus on testing off-the-shelf LLMs using general-purpose optimisation strategies that can be widely applied across diverse tasks and datasets. In contrast, other researchers have focused on developing specialised approaches for SciQA, often incorporating additional components to integrate information from the ORKG ontological schema [19, 20, 21].

In summary, the key contributions of this study are as follows: i) we conduct performance analysis of five language models, evaluated across zero-shot, few-shot, and fine-tuning approaches; ii) we demonstrate that the best models can achieve an F1 score exceeding 97% on both benchmarks; and iii) we release the complete codebase of our experiments to support further research into LLMs performance on similar benchmarking tasks⁴.

2. Experiments on the SciQA dataset

The SciQA dataset contains 2,565 pairs of natural language questions and corresponding SPARQL queries, designed to retrieve relevant information from the ORKG, which includes 170,000 resources detailing research from nearly 15,000 scholarly articles across 709 topics. The dataset consists of both manually curated and automatically generated question-query pairs. Specifically, 100 pairs were manually created, revealing eight distinct question templates. Using these templates, an additional 2,465 pairs were generated by GPT-3 and verified by human experts [22]. The SciQA benchmark dataset is divided into three parts: 70% for training (1,795 samples), 10% for validation (257 samples), and 20% for testing (513 samples).

For our experiments, we evaluated five LLMs: T5-base⁵[23], GPT-2-large⁶[24], Dolly-v2-3b⁷[25], Mistral-7B-v0.1⁸[26], and GPT-3.5 Turbo⁹[22]. We examined three optimization methods for LLMs: fine-tuning (FT), zero-shot learning (ZSL), and few-shot learning (FSL). In FSL, to evaluate a question from the test set, we used different methods from the literature to select the most relevant samples for each question.

Random: Select S samples randomly from the training set for each test question.

Similarity: Order samples by their semantic similarity to the test question, and choose the top S most similar samples.

Diversity - Test A (All Diverse Templates): Rank samples by semantic similarity to the test question and select the top S samples, ensuring they represent different templates.

Diversity - Test B (Same Template for All): Rank samples by semantic similarity and select the top S samples that share the same template as the first sample.

⁴Codebase and prompts - <https://github.com/paper-support-materials/Analysis-of-the-SciQA-Benchmark>

⁵T5-base - <https://huggingface.co/t5-base>

⁶GPT-2-large - <https://huggingface.co/gpt2-large>

⁷Dolly-v2-3b - <https://huggingface.co/databricks/dolly-v2-3b>

⁸Mistral-7B-v0.1 - <https://huggingface.co/mistralai/Mistral-7B-v0.1>

⁹GPT-3.5 Turbo - <https://platform.openai.com/docs/models/gpt-3-5>

Table 1

Summary of F1 scores and exact matches (in parentheses), organised by various strategies (Strat.) and criteria (C). *S* denotes the number of samples, with the best results of each model displayed in **bold**.

Strat.	C	Test name	S	T5-base	GPT2-large	Dolly-v2-3b	Mistral-7B	GPT-3.5-turbo
FT				0.9751 (483)	0.9669 (430)	0.9658 (483)	0.9769 (480)	
ZSL					0.0653 (0)	0.1087 (0)	0.2043 (0)	0.2632 (0)
FSL	Similarity		1		0.2718 (0)	0.8792 (167)	0.8529 (259)	0.9368 (356)
			3		0.4051 (2)	0.8304 (182)	0.9409 (382)	0.9667 (451)
			5			0.8242 (180)	0.9386 (403)	0.9709 (464)
			7			0.8052 (181)	0.9470 (409)	0.9736 (475)
	Random		1		0.2005 (0)	0.5659 (27)	0.5055 (28)	0.7362 (45)
			3		0.2187 (0)	0.5900 (31)	0.6835 (75)	0.8259 (113)
			5			0.6242 (51)	0.7342 (115)	0.8675 (165)
			7			0.6576 (69)	0.7799 (135)	0.8905 (189)
	Diversity	Test A	3		0.2215 (0)	0.7000 (43)	0.8589 (154)	0.9378 (315)
			5			0.6525 (39)	0.8447 (120)	0.9428 (328)
			7			0.6729 (46)	0.8033 (84)	0.9375 (313)
		Test B	3		0.2988 (1)	0.8025 (171)	0.9179 (368)	0.9561 (412)
			5			0.8181 (201)	0.9095 (378)	0.9566 (417)
			7			0.8261 (212)	0.9087 (389)	0.9562 (422)
		DPPs	7			0.3912 (54)	0.7277 (84)	0.9452 (389)

Diversity - DPPs (Determinantal Point Processes): Start with the most similar sample to the question and select additional samples with minimal semantic similarity to each other and the initial sample [27].

Table 1 provides a comparative analysis of all configurations according to their F1 scores and number of exact match rates. The fine-tuned Mistral achieved the highest F1 score (97.69%), slightly surpassing the fine-tuned T5 (97.51%) and GPT-3.5, using the 7-sample few-shot method based on similarity (97.36%). Next is the fine-tuned Dolly (96.58%) and GPT-2 (96.69%). In terms of exact matches, T5 and Dolly attained the highest score (483/513, 94.1%). The model utilising FSL performed well overall, though it did not achieve the same level of performance as the fine-tuned model. For example, Mistral, with a 7-sample FSL approach based on similarity, achieved a solid 94.7% F1. Semantic similarity is the most effective method for FSL across all models. Notably, the benchmark proved highly challenging for all models under ZSL conditions, as none achieved any exact matches and F1 scores were under 26%.

We performed an in-depth review of the queries generated by the top three models that were classified as incorrect due to their deviation from the benchmark responses. The most common error category involved the generation of incorrect predicates (56.8% of erroneous queries on average), followed by semantic errors, where the query failed to accurately reflect the user’s question (51.4%), and misspelled entities (51.3%). A query can be associated with multiple categories, meaning the total percentage across all categories exceeds 100%.

The most common error types for both T5 and GPT-3.5 stemmed from a limited understanding of the underlying ontological schema. A prevalent issue was the generation of misspelled entities, which constituted 60.0% of T5’s and 60.5% of GPT-3.5’s incorrect outputs. Furthermore, both models had difficulty accurately assigning entity types, contributing to 36.6% of T5’s errors and 52.6% of GPT-3.5’s errors.

In contrast, Mistral performed significantly better in these two categories, with error rates of 33.3% and 15.2%, respectively. However, Mistral exhibited a much higher rate of semantic misunderstandings, with 69.7% of its errors resulting from incorrect interpretations of the query.

The first type of error appears easier to address, potentially by incorporating an additional entity recognition component. To investigate this approach further, we evaluated the top three models (fine-tuned Mistral, fine-tuned T5, and GPT-3.5 with 7-shot learning) on the DBLP-QuAD benchmark. DBLP-QuAD is similar to SciQA but also includes relevant entities and relationships as part of the input. This enabled us to assess whether providing the correct entity could help reduce the occurrence of

Table 2

F1-scores and exact matches of the best models on DBLP-QuAD dataset. In **bold** the most performing model.

	T5-base	Mistral-7B	GPT-3.5 Turbo
F1-score	0,9746	0,8866	0,9473
Exact Match	1693	1214	1364

these types of errors.

3. Experiments on the DBLP-QuAD dataset

The DBLP-QuAD benchmark¹⁰ [14] includes 10,000 distinct question-query pairs, divided into training, validation, and test sets in a 7:1:2 ratio. The dataset covers 13,348 entities (creators and publications) and 11 predicates from the DBLP Knowledge Graph¹¹. It offers 10 query types, each with 1,000 question-query pairs, equally split between creator-focused and publication-focused queries. Additionally, 2,350 of the questions in DBLP-QuAD are temporal, requiring the analysis of statistics across a specified timeframe, e.g., “in the last five years”. The key difference from the SciQA dataset is that the entities and relationships involved in the natural language query are also provided.

We evaluated the top three models from our previous experiments on the DBLP-QuAD dataset. Please note that, unlike the original paper that introduced the benchmark [14], we fine-tuned the T5 model for 20 epochs instead of 5, and applied a slightly modified prompt. Full implementation details can be found in the GitHub repository linked in the introduction.

As reported in Table 2, T5-base achieved the best results (97.5%), slightly outperforming GPT-3.5 (94.7%). Mistral exhibited lower performance on this benchmark (88.7%). The advantage of T5-base becomes even more evident when considering exact matches. In this case, T5-base leads significantly (1,693/2,000, 84.6%), outperforming both GPT-3.5 at (68.2%) and Mistral (60.7%).

These findings are consistent with previously observed error patterns. T5 and GPT-3.5, which had previously struggled the most with entity identification, now seem to leverage the provided entities effectively and outperform Mistral. Additionally, the results suggest that a lightweight encoder-decoder model such as T5, which can be fine-tuned effectively for translating natural language into SPARQL, has significant potential as a scalable, resource-efficient, and effective method for this task, particularly when paired with an entity resolution mechanism.

4. Conclusions

The experiments reported in this short paper provide several valuable insights about the capability of LLMs to address scientific question answering on KGs.

First, it appears that current benchmarks are not sufficiently challenging for the latest generation of LLMs, as the best configurations achieved over 97% F1 on both benchmarks. This may be attributed to the regularities within the benchmarks, allowing fine-tuned models to learn and reproduce patterns. To advance this field further, it is crucial to develop more diverse and challenging datasets that cover a broader range of realistic query types, while actively involving human users in the dataset creation process to minimise the risk of LLMs learning only a limited set of templates. Conducting user studies with real-world applications could also be beneficial, as users tend to formulate more varied and complex questions [28].

Second, incorporating additional components to resolve entities and relations seems to be highly useful, particularly for encoder-decoder models like T5 and generalist LLMs using few-shot learning, such as GPT-3.5. This approach allows LLMs to focus on semantic interpretation and the generation of

¹⁰<https://huggingface.co/datasets/awalesushil/DBLP-QuAD?row=89>

¹¹<https://blog.dblp.org/tag/knowledge-graph/>

accurate, well-formed SPARQL queries without needing to understand the specific schema of a given KG. A possible enhancement would be to also provide systems with an ontological schema representation as context, as explored in recent specialised approaches [19, 20].

We are currently developing a new and more challenging benchmark, building upon the Academia/Industry DynAmics (AIDA) Knowledge Graph [29] and the Computer Science Knowledge Graph (CS-KG) [30]. To expand the diversity of question types, we are leveraging various question templates drawn from large-scale question-answering benchmarks, such as Mintaka [31]. We are also analysing the performance of large language models across several tasks relevant to scientific research, including the construction of scientific knowledge graphs [32], link prediction between research concepts [33, 34], research paper classification [35], citation recommendation [36], and the generation of literature reviews [3].

References

- [1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, *ACM Computing Surveys* 55 (2023) 1–38.
- [2] X. Zhou, K. Kim, B. Xu, J. Liu, D. Han, D. Lo, The devil is in the tails: How long-tailed code distributions impact large language models, *arXiv preprint arXiv:2309.03567* (2023).
- [3] F. Bolanos, A. Salatino, F. Osborne, E. Motta, Artificial intelligence for literature reviews: Opportunities and challenges, *arXiv preprint arXiv:2402.08565* (2024).
- [4] A. Hogan, E. Blomqvist, M. Cochez, C. d’Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, *ACM Computing Surveys (Csur)* 54 (2021) 1–37.
- [5] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges, *Artificial Intelligence Review* (2023) 1–32.
- [6] A. Salatino, T. Aggarwal, A. Mannocci, F. Osborne, E. Motta, A survey on knowledge organization systems of research fields: Resources and challenges, *arXiv preprint arXiv:2409.04432* (2024).
- [7] M. Y. Jaradeh, A. Oelen, K. E. Farfar, et al., Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge, in: *Proceedings of the 10th International Conference on Knowledge Capture*, 2019, pp. 243–246.
- [8] A. A. Salatino, F. Osborne, A. Birukou, E. Motta, Improving editorial workflow and metadata quality at springer nature, in: *The Semantic Web – ISWC 2019*, Springer International Publishing, Cham, 2019, pp. 507–525.
- [9] M. Wijkstra, T. Lek, T. Kuhn, K. Welbers, M. Steijaert, Living literature reviews, *arXiv preprint arXiv:2111.00824* (2021).
- [10] S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, E. Motta, AIDA: A knowledge graph about research dynamics in academia and industry, *Quantitative Science Studies* 2 (2021) 1356–1398. URL: https://doi.org/10.1162/qss_a_00162. doi:10.1162/qss_a_00162. arXiv:https://direct.mit.edu/qss/article-pdf/2/4/1356/2007973/qss_a_00162.pdf.
- [11] P. Groth, A. Gibson, J. Velterop, The anatomy of a nanopublication, *Information Services & Use* 30 (2010) 51–56.
- [12] L.-P. Meyer, J. Frey, F. Brei, N. Arndt, Assessing sparql capabilities of large language models, 2024. URL: <https://arxiv.org/abs/2409.05925>. arXiv:2409.05925.
- [13] S. Auer, D. A. C. Barone, C. Bartz, E. G. Cortes, M. Y. Jaradeh, O. Karras, M. Koubarakis, D. Mourontsev, D. Pliukhin, D. Radyush, I. Shilin, M. Stocker, E. Tsalapati, The sciq scientific question answering benchmark for scholarly knowledge, *Scientific Reports* 13 (2023) 7240. URL: <https://doi.org/10.1038/s41598-023-33607-z>. doi:10.1038/s41598-023-33607-z.
- [14] D. Banerjee, S. Awale, R. Usbeck, C. Biemann, Dbp-quad: A question answering dataset over the DBLP scholarly knowledge graph, in: I. Frommholz, P. Mayr, G. Cabanac, S. Verberne, J. Brennan (Eds.), *Proceedings of the 13th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 45th European Conference on Information Retrieval (ECIR 2023)*, Dublin,

- Ireland, April 2nd, 2023, volume 3617 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 37–51. URL: <https://ceur-ws.org/Vol-3617/paper-05.pdf>.
- [15] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, W. Chen, What makes good in-context examples for gpt-3?, arXiv preprint arXiv:2101.06804 (2021).
 - [16] I. Levy, B. Bogin, J. Berant, Diverse demonstrations improve in-context compositional generalization, arXiv preprint arXiv:2212.06800 (2022).
 - [17] A. Kulesza, B. Taskar, et al., Determinantal point processes for machine learning, *Foundations and Trends® in Machine Learning* 5 (2012) 123–286.
 - [18] D. Iter, R. Pryzant, R. Xu, S. Wang, Y. Liu, Y. Xu, C. Zhu, In-context demonstration selection with cross entropy difference, arXiv preprint arXiv:2305.14726 (2023).
 - [19] L. Jiang, X. Yan, R. Usbeck, A structure and content prompt-based method for knowledge graph question answering over scholarly data, *CEUR Workshop proceedings* 3592 (2023). URL: <https://ceur-ws.org/Vol-3592/paper3.pdf>.
 - [20] T. A. Taffa, R. Usbeck, Leveraging llms in scholarly knowledge graph question answering, in: *Scholarly-QALD-23: Scholarly QALD Challenge at The 22nd International Semantic Web Conference (ISWC 2023)*(Athens, Greece, volume 3592, 2023, pp. 1–10. URL: <https://ceur-ws.org/Vol-3592/paper5.pdf>.
 - [21] D. Pliukhin, D. Radyush, L. Kovriguina, D. Mouromtsev, Improving subgraph extraction algorithms for one-shot sparql query generation with large language models, in: *Scholarly-QALD-23: Scholarly QALD Challenge at The 22nd International Semantic Web Conference (ISWC 2023)*(Athens, Greece, volume 3592, 2023, pp. 1–10. URL: <https://ceur-ws.org/Vol-3592/paper6.pdf>.
 - [22] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.
 - [23] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.* 21 (2020).
 - [24] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, *OpenAI blog* 1 (2019) 9.
 - [25] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, R. Xin, Free dolly: Introducing the world’s first truly open instruction-tuned llm, 2023. URL: <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>.
 - [26] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, *ArXiv abs/2310.06825* (2023). URL: <https://api.semanticscholar.org/CorpusID:263830494>.
 - [27] A. Kulesza, Determinantal point processes for machine learning, *Foundations and Trends® in Machine Learning* 5 (2012) 123–286. URL: <http://dx.doi.org/10.1561/22000000044>. doi:10.1561/22000000044.
 - [28] A. Meloni, S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, E. Motta, Integrating conversational agents and knowledge graphs within the scholarly domain, *Ieee Access* 11 (2023) 22468–22489.
 - [29] S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, E. Motta, AIDA: A knowledge graph about research dynamics in academia and industry, *Quantitative Science Studies* 2 (2021) 1356–1398. URL: https://doi.org/10.1162/qss_a_00162. doi:10.1162/qss_a_00162. arXiv:https://direct.mit.edu/qss/article-pdf/2/4/1356/2007973/qss_a_00162.pdf.
 - [30] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, Cs-kg: A large-scale knowledge graph of research entities and claims in computer science, in: U. Sattler, A. Hogan, M. Keet, V. Presutti, J. P. A. Almeida, H. Takeda, P. Monnin, G. Pirrò, C. d’Amato (Eds.), *The Semantic Web – ISWC 2022*, Springer International Publishing, Cham, 2022, pp. 678–696.
 - [31] P. Sen, A. F. Aji, A. Saffari, Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering, arXiv preprint arXiv:2210.01613 (2022).

- [32] D. Dessí, F. Osborne, D. R. Recupero, D. Buscaldi, E. Motta, Scicero: A deep learning and nlp approach for generating scientific knowledge graphs in the computer science domain, *Knowledge-Based Systems* 258 (2022). URL: <https://oro.open.ac.uk/85472/>. doi:10.1016/j.knosys.2022.109945.
- [33] M. Nayyeri, G. M. Cil, S. Vahdati, F. Osborne, M. Rahman, S. Angioni, A. Salatino, D. R. Recupero, N. Vassilyeva, E. Motta, et al., Trans4e: Link prediction on scholarly knowledge graphs, *Neurocomputing* 461 (2021) 530–542.
- [34] A. Borrego, D. Dessì, I. Hernández, F. Osborne, D. R. Recupero, D. Ruiz, D. Buscaldi, E. Motta, Completing scientific facts in knowledge graphs of research concepts, *IEEE Access* 10 (2022) 125867–125880.
- [35] A. Cadeddu, A. Chessa, V. De Leo, G. Fenu, E. Motta, F. Osborne, D. R. Recupero, A. Salatino, L. Secchi, A comparative analysis of knowledge injection strategies for large language models in the scholarly domain, *Engineering Applications of Artificial Intelligence* 133 (2024) 108166.
- [36] D. Buscaldi, D. Dessí, E. Motta, M. Murgia, F. Osborne, D. R. Recupero, Citation prediction by leveraging transformers and natural language processing heuristics, *Information Processing & Management* 61 (2024) 103583.