NFDI4DS Shared Tasks for Scholarly Document Processing

Raia Abu Ahmad[1], Rana Abdulla[2], Tilahun Abedissa Taffa[2], Sören Auer[7], Hamed Babaei Giglou[7], Ekaterina Borisova[1], Zongxiong Chen[5], Stefan Dietze[6], Jennifer D'Souza[7], Mayra Elwes[3], Genet-Asefa Gesese[4], Shufan Jiang[4], Ekaterina Kutafina[3], Philipp Mayr[6], Georg Rehm[1], Sameer Sadruddin[7], Sonja Schimmler[5], Daniel Schneider[8], Kanishka Silva[6], Sharmila Upadhyaya[6], and Ricardo Usbeck[2]

Abstract: Shared tasks are powerful tools for advancing research through community-based standardised evaluation. As such, they play a key role in promoting findable, accessible, interoperable, and reusable (FAIR), as well as transparent and reproducible research practices. This paper presents an updated overview of twelve shared tasks developed and hosted under the German National Research Data Infrastructure for Data Science and Artificial Intelligence (NFDI4DS) consortium, covering a diverse set of challenges in scholarly document processing. Hosted at leading venues, the tasks foster methodological innovations and contribute open-access datasets, models, and tools for the broader research community, which are integrated into the consortium's research data infrastructure.

Keywords: Shared Tasks, NFDI4DS, NFDI

1 Introduction

Shared tasks are community-driven challenges that compare computational methods on specific problems using shared datasets and metrics [Pa17]. They advance research by fostering innovation, encouraging collaboration, and establishing state-of-the-art methods across tasks and fields [FU18]. Shared tasks offer multiple benefits to the community by promoting transparency and reproducibility, allowing for a fair comparison of methodologies, and resulting in publicly available resources (e. g., datasets, open-source software) [Es21].

[1] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin, Germany, raia.abu_ahmad@dfki.de; ekaterina.borisova@dfki.de; georg.rehm@dfki.de
[2] Leuphana University of Lüneburg, Germany, rana.abdullah@leuphana.de; tilahun.taffa@leuphana.de; ricardo.usbeck@leuphana.de
[3] University Hospital Cologne, University of Cologne, Institute for Biomedical Informatics, Germany, mayra.elwes@uk-koeln.de; ekaterina.kutafina@uni-koeln.de
[4] FIZ-Karlsruhe – Leibniz-Institute for Information Infrastructure, Germany, genet-asefa.gesese@fiz-karlsruhe.de; shufan.jiang@fiz-karlsruhe.de
[5] Fraunhofer Institute for Open Communication Systems FOKUS, Berlin, Germany, zongxiong.chen@fokus.fraunhofer.de; sonja.schimmler@fokus.fraunhofer.de
[6] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany, stefan.dietze@gesis.org; philipp.mayr@gesis.org; kanishka.silva@gesis.org; sharmila.upadhyaya@gesis.org
[7] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany, auer@tib.eu; hamed.babaei@tib.eu; jennifer.dsouza@tib.eu; sameer.sadruddin@tib.eu
[8] Innovation Center Computer Assisted Surgery (ICCAS), Leipzig University, Germany, daniel.schneider@uni-leipzig.de

doi:10.18420/inf2025_103

NFDI4DS (National Research Data Infrastructure for Data Science and Artificial Intelligence) is a German consortium that focuses on supporting the research data lifecycle in Data Science and Artificial Intelligence (AI) [Sc23].[9] This includes collecting, processing, analysing, publishing, and reusing artifacts such as datasets, models, code, and publications. NFDI4DS pays special attention to the FAIR principles as well as to transparency and reproducibility through open-source tools and rich documentation.

Shared tasks play an important role in NFDI4DS, serving as mechanisms for community engagement and driving the development and refinement of tools and methods for its FAIR research data infrastructure. They also support NFDI's broader mission of building a sustainable, interoperable, and accessible national research data infrastructure by promoting interdisciplinary collaboration, standardisation efforts, and long-term reuse of research outputs.

To reach a broader audience and encourage widespread participation, NFDI4DS shared tasks are usually hosted at well-known venues. Currently, these include 1. the Natural Scientific Language Processing and Research Knowledge Graphs (NSLP) workshop series, initiated by NFDI4DS;[10] 2. the Scholarly Document Processing (SDP) workshop series [Ch20];[11] 3. the Conference and Labs of the Evaluation Forum (CLEF);[12] 4. the Workshop on Semantic Evaluation (SemEval);[13] and 5. the Int. Semantic Web Conference (ISWC) challenges.[14]

This paper provides an overview of shared tasks organised by NFDI4DS, extending [Ab23]. The twelve tasks demonstrate how these competitions can drive methodological progress and contribute to the development of a FAIR research data infrastructure.

2 Completed Shared Tasks

2.1 SOMD 2025: Software Mention Detection

Organiser: GESIS – Leibniz Institute for the Social Sciences
Although software is integral to scientific discovery and analysis, it is rarely cited with the same consistency or formality as literature. Automatic identification and disambiguation of software mentions and related attributes (URLs, Versions, License, etc.) enhance accessibility and reproducibility but remain challenging [Sc21]. SOMD 2025 [Up25] is a shared task focused on extracting software mentions from scholarly articles. Following SOMD 2024 [KKD24], the task focused on joint learning of software mentions, attribute

[9] https://www.nfdi4datascience.de
[10] https://nfdi4ds.github.io/nslp2025/ and Rehm et al. [Re24]
[11] https://sdproc.org
[12] https://clef2025.clef-initiative.eu/
[13] https://semeval.github.io
[14] https://iswc2025.semanticweb.org/#/program/challenges

extraction, and classification of the their relations. The task was hosted at SDP 2025,[15] and included two phases: Phase I involved model development using a gold standard training set [Sc21], evaluated using the average score of the entities extraction and relation classification tasks, while phase II tested generalisability on an out-of-distribution set. The competition (March/April 2025) attracted six teams, the best one achieved F1 scores of 0.89 (Phase I) and 0.63 (Phase II), surpassing baselines of 0.804 and 0.491, respectively.

2.2 SciVQA: Scientific Visual Question Answering

Organiser: Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
The first SciVQA shared task [BRR25], part of SDP 2025,[16] is about developing question answering (QA) systems based on images of scientific figures, their captions, and metadata. The challenge aims to evaluate the recognition and reasoning capabilities of multimodal large language models (LLMs) on closed-ended visual vs. non-visual questions. The former refers to questions addressing visual attributes such as colour, shape, direction, height, size or position of objects, while the latter does not incorporate any of these aspects. The novel and large-scale SciVQA dataset[17] comprises 3000 images of figures from arXiv pre-prints and ACL Anthology publications. Each figure is semi-automatically annotated according to a custom QA schema featuring seven types, resulting in a total of 21000 English QA pairs. The benchmark contains several types of figures (e. g., line graph, bar chart, venn diagram, confusion matrix), also distinguishing between compound and non-compound figures. The evaluation is based on precision, recall and F1-scores of ROUGE-1, ROUGE-L [Li04], and BertScore [Zh20].[18] The competition attracted 20 registrations, with seven leaderboard submissions and five papers reporting their solutions. The highest-ranking team achieved scores of 0.80 for ROUGE-1 and ROUGE-L and 0.98 for BertScore.

2.3 ClimateCheck: Scientific Fact-checking of Social Media Posts on Climate Change

Organiser: Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
ClimateCheck [AUR25b] was hosted at SDP 2025.[19] The task addressed the growing challenge of misinformation in online discourse about critical issues, taking climate change as an example, by linking claims made on social media to relevant scholarly articles. Two datasets were constructed for the task [AUR25a], the first consists of English climate-related claims gathered from existing resources, while the second is a corpus of 394,269 publications in climate and environmental sciences. The claims were filtered for scientific check-worthiness and rephrased to be atomic tweet-styled statements. A TREC-like [Vo05]

[15] https://sdproc.org/2025/somd25.html
[16] https://sdproc.org/2025/scivqa.html
[17] https://huggingface.co/datasets/katebor/SciVQA
[18] https://www.codabench.org/competitions/5904/
[19] https://sdproc.org/2025/climatecheck.html

linking approach was then used to create claim-abstract pairs for annotation, resulting in a dataset of 435 unique claims, each connected to abstracts annotated as *supports*, *refutes*, or *not enough information*.[20] Subtask I of ClimateCheck focused on retrieving the top 10 relevant abstracts per claim, evaluated using Recall@K as well as Binary Preference (Bpref), while subtask II asked participants to classify the relation of each claim-abstract pair they retrieved, evaluated using F1-score with additional scaling based on successful retrieval. The task was hosted on CodaBench [Xu22], attracting 27 participants with ten leaderboard submissions. The top team achieved a Recall@10 score of 0.66, Bpref of 0.49 (subtask I), and and F1 of 0.73 (subtask II).

2.4 FoRC2025: Field of Research Classification

Organiser: Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
The second FoRC shared task [Fr25] was hosted at NSLP 2025,[21] running in February/April 2025 on CodaBench. Building on FoRC 2024 [ABR24a], the 2025 iteration reused the FoRC4CL dataset with 1,500 ACL Anthology papers annotated using a taxonomy of 170 topics [ABR24b]. For 2025, a weakly-labeled dataset of 41,000 automatically annotated articles was added. Given an ACL Anthology article, the task was to predict all labels depicting its main contributions. While only one team submitted results, they significantly outperformed the baselines and previous systems using a two-stage approach: k-nearest neighbors for label retrieval, followed by zero-shot prompting for refinement. Their model yielded a micro-F1 of 0.68, a macro-F1 of 0.66, and a weighted-F1 of 0.69, showing the potential of hybrid retrieval-generation pipelines. The FoRC4CL 2025 dataset offers a valuable benchmark for future research on hierarchical multi-label classification, weak supervision methods, and the integration of LLMs in scientific document understanding.

2.5 LLMs4Subjects@SemEval-2025: LLM-based Automated Subject Tagging for a National Technical Library's Open-Access Catalog

Organiser: TIB Leibniz Information Centre for Science and Technology
The first edition of LLMs4Subjects, organised as SemEval-2025 Task 5,[22] focused on automated subject tagging of scientific and technical records using LLMs [DS25]. The task targeted records in English and German from the TIBKAT catalog, Germany's national technical library (5.7 million records). The dataset comprised 123,589 open-access records annotated with subject terms from the Gemeinsame Normdatei (GND) [DS24]. Two variants were offered: (1) the *all-subjects* collection (~100k records, full GND coverage) and (2) the *tib-core* subset (~48k records, 14 key technical domains), each accompanied by a refined

[20] https://huggingface.co/datasets/rabuahmad/climatecheck
[21] https://nfdi4ds.github.io/nslp2025/docs/forc_shared_task.html
[22] https://semeval.github.io/SemEval2025/tasks.html

GND taxonomy. In total, 14 teams participated with approaches ranging from extreme multi-label text classification with LLM augmentation (e. g., [SIL25]), to few-shot LLM ensembles [KK25], to retrieval-augmented generation (e. g., [Ba25]). Quantitative and qualitative evaluations were conducted, with subject specialists reviewing model outputs. Key insights included the benefits of synthetic data generation, LLM-assisted translation, multilingual model design, and contrastive learning.[23] The task established a strong foundation for scalable, bilingual subject tagging in digital libraries. The next edition[24] will focus on *energy- and compute-efficient* LLM systems as part of the GermEval workshop series.

2.6 LLMs4OL-2024: Large Language Models for Ontology Learning

Organiser: TIB Leibniz Information Centre for Science and Technology
The first LLMs4OL challenge [GDA24] was organised at ISWC 2024 and co-supported by the SCINEXT project. The goal was to evaluate and advance the use of LLMs for ontology learning (OL) tasks, including term typing, taxonomy induction, and non-taxonomic relation extraction [BDA23]. The challenge comprised three tasks and 21 subtasks, using datasets derived from ontologies such as WordNet, GeoNames, and UMLS. Datasets were annotated for both few-shot and zero-shot setups, allowing for robust evaluation of model generalisation [Ba24]. The challenge attracted 14 participating teams on Codalab,[25] eight of which reached the final leaderboard. Winning approaches combined traditional machine learning with LLM prompt-tuning, rules-based augmentation, or retrieval-enhanced methods. The diversity of methods and breadth of domains addressed established a strong benchmark for future research, highlighting the importance of hybrid strategies and domain adaptation for effective ontology construction. A second edition, LLMs4OL 2025,[26] is planned to be co-located with ISWC 2025. A major addition is a new task, *text2onto*, which evaluates systems for their ability to generate structured ontologies from unstructured text.

2.7 Scholarly QALD: Scholarly Question Answering over Linked Data

Organiser: Leuphana University of Lüneburg, Germany
The QALD challenge addresses the development of natural language interfaces over scholarly knowledge graphs (KGs), enabling users to access and query them without requiring expertise in semantic technologies. It was held at ISWC 2023 and ISWC 2024, aiming to foster the development and evaluation of QA systems tailored to scholarly data. The 2023 edition comprised two main tasks: 1. DBLP-QUAD (KGQA over DBLP), in which participants worked with a dataset of 10,000 question-SPARQL pairs, designed for queries over DBLP; and 2. SciQA (KGQA over ORKG), which focused on answering comparative scholarly

[23] https://sites.google.com/view/llms4subjects/team-results-leaderboard
[24] https://sites.google.com/view/llms4subjects-germeval/
[25] https://codalab.lisn.upsaclay.fr/competitions/19547
[26] https://sites.google.com/view/llms4ol2025

questions over the ORKG. The task attracted seven teams [Ba23] and was evaluated using the F1 score.[27] In 2024, a new Hybrid Question Answering (Hybrid-QA) task was introduced, requiring participants to fetch answers by querying multiple scholarly KGs and textual resources, reflecting more realistic, multi-source user needs in the scholarly domain [Ta24]. Three teams submitted results, and the evaluation was expanded to a broader set of metrics, including Exact Match, F-Score, and METEOR.[28]

3 Ongoing Shared Tasks

3.1 CheckThat!Lab Task 4: Scientific Web Discourse

Organiser: GESIS – Leibniz Institute for the Social Sciences
Scientific web discourse has increased substantially throughout the past years [BLW20; Du21]. It is usually informal and exhibits incomplete citation habits, where the actual study is never explicitly cited. This poses challenges both from a computational perspective, when mining social media or computing Altmetrics, and a societal perspective, leading to poorly informed online debates [Ro21]. To address this, two tasks are introduced as part of the CLEF 2025 CheckThat! Lab.[29] Subtask 4.1, SciWeb Discourse Detection, aims to classify the different forms of science-related online discourse [Ha22]. Given a tweet, the task is to detect if it contains a scientific claim, a scientific reference, or is referring to science contexts or entities. The SciTweets corpus [Ha22] will be used with an additional test set of 500-1000 tweets. Systems will be evaluated using macro-F1. Subtask 4.2, SciWeb Claim-Source Retrieval, aims to retrieve the scientific paper that serves as the source for a claim with an implicit scientific reference. A ground truth dataset containing tweet-study pairs will be provided, using the CORD-19 corpus [Wa20]. Tweets are collected using the Altmetric corpus[30] and a pretrained classifier [Ha22], resulting in a collection of 15,967 tweet-study pairs. Systems will be evaluated using Mean Reciprocal Rank (MRR@5).

4 Shared Tasks Under Development

4.1 Dataset and Machine Learning Model Mention Detection

Organiser: GESIS – Leibniz Institute for the Social Sciences
Machine Learning (ML) models and datasets are the foundation of scientific discovery and innovation in AI. Their occurrence in scholarly articles is crucial for understanding and reproducing of research. With the intent of creating novel tasks and corpora of scholarly work

[27] https://kgqa.github.io/scholarly-QALD-challenge/2023/
[28] https://kgqa.github.io/scholarly-QALD-challenge/2024/
[29] https://checkthat.gitlab.io/clef2025/
[30] https://www.altmetric.com

to contribute to the fine-grained representation of research work, the GSAP-NER corpora [Ot23] were designed. This set of gold-standard datasets provides a nuanced understanding of how ML models and datasets are mentioned and utilised in scholarly articles. The shared task will use GSAP-NER and encourage researchers to work on novel solutions for detecting ML models and datasets, optimising the automated extraction of these entities from scholarly articles.

4.2 ReadMe2KG: Github ReadMe to Knowledge Graph

Organiser: FIZ Karlsruhe - Leibniz Institute for Information Infrastructure
GitHub[31] is a popular platform for hosting and collaborating on software projects. Researchers use GitHub repositories to share datasets, models, and the source code of their experiments. These repositories can provide implementation details and facilitate the exploration and reproduction of research results, each typically including a README.md file, which serves as an introductory document. READMEs are usually written in Markdown format and provide key information such as the project's purpose, setup instructions, usage examples, and often links to the original research paper. Aiming to enhance the NDFI4DS-KG[32] with information from GitHub README files, two subtasks are proposed. In Subtask I, Fine-grained Named Entity Recognition, participants will develop classifiers that take README files as input and output the mentions of the ∼15 classes in the NFDI4DS Ontology [Ge24]. A dataset with ∼160 README.md files will be made available to train classifiers, which will be evaluated using micro and macro scores of recall, precision, and F1. Subtask II, Entity Linking, will task participants with developing a method to link entities mentioned in README files to entities in the NFDI4DS-KG. A dataset with ∼200 README.md files will be made available. The systems will be evaluated using micro and macro scores of recall, precision, and F1.

4.3 CURED: Clinical Uncertainty and Reliable Evidence-based Decision-making

Organiser: Innovation Center Computer Assisted Surgery (ICCAS)
CURED is a planned shared task on trustworthy, evidence-based clinical prognosis of patient trajectories. In safety-critical settings like clinical decision-making, AI applications must operate under expert supervision. Clinical decision support (CDS) systems fulfill this role by processing large volumes of data to generate actionable insights, while ensuring that clinicians retain final control. However, AI-driven CDS models frequently exhibit overconfidence when faced with outliers or distribution shifts [Gu17]. To foster reliable human–AI collaboration and support evidence-based decisions, CDS systems must transparently communicate their uncertainty and limits of their knowledge [Li25]. However,

[31] https://github.com
[32] 1https://nfdi.fiz-karlsruhe.de/4ds/sparql, https://nfdi.fiz-karlsruhe.de/4ds/shmarql

no adequate standardised benchmark currently exists for evaluating knowledge uncertainty estimates in clinical settings. To address this gap, we propose CURED, a shared task to evaluate collaborative clinical decision making in multi-agent settings by leveraging uncertainty quantification to efficiently integrate each agent's expertise. CURED will challenge participants to build prognostic models on longitudinal electronic health records that output both a prediction and calibrated epistemic confidence estimate. Submissions will be ranked according to both discriminative performance and the quality of their uncertainty estimates. CURED is currently in the dataset creation phase, with a plan to run in 2026.

4.4 Algorithm2Domain

Organiser: University Hospital Cologne, University of Cologne
Reusability of data science algorithms is crucial for efficient and scalable research. To maximise their generalisability to unseen data and adaptability across diverse distributions, promoting best-practices in algorithm development is essential. Simultaneously, new projects need effective ways to discover existing algorithms that generalise well to their specific data. To address this, we initiated the development of a meta-benchmarking platform to evaluate algorithm generalisability and test domain adaptation methods. The concept is built on four key pillars: 1. Decomposition of algorithm development into "degrees of freedom" for modular benchmarking of computational components; 2. Collaborative development of best-practice benchmarking guidelines; 3. Establishment of FAIR standards for benchmarking, drawing on existing frameworks such as Croissant[33]; 4. Co-creation of a shared knowledge base on algorithm generalisation and adaptation. Algorithm2Domain,[34] builds on insights from the ADATIME benchmarking suite [Ra23], mapping its structure to the "degrees of freedom" framework, testing the process of adding new modules, and exploring no-code GUI solutions for rapid testing. We aim to establish a development roadmap for these pillars and build the infrastructure to support community-driven contributions in shared tasks, such as the preparation of algorithms, benchmarking, and expansion of the collaborative knowledge base.

5 Conclusion

This paper presented a total of twelve shared tasks, all of which are part of the NFDI4DS consortium, which aims to contribute a FAIR, transparent, and reproducible research data infrastructures. The presented shared tasks were split into seven completed, one ongoing, and four future tasks, demonstrating the consortium's efforts in tackling challenging problems in the scholarly information processing domain using a community-based approach. These efforts directly contribute to the broader NFDI mission of enabling sustainable and interoperable research data practices across different disciplines.

[33] https://github.com/mlcommons/croissant
[34] https://github.com/BI-K/Algorithm2Domain/

References

[Ab23] Abu Ahmad, R. et al.: NFDI4DS Shared Tasks. Gesellschaft für Informatik eV, 2023.

[ABR24a] Abu Ahmad, R.; Borisova, E.; Rehm, G.: FoRC@ NSLP2024: Overview and Insights from the Field of Research Classification Shared Task. In: International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs. Springer, pp. 189–204, 2024.

[ABR24b] Ahmad, R. A.; Borisova, E.; Rehm, G.: FoRC4CL: a fine-grained field of research classification and annotated dataset of NLP articles. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Pp. 7389–7394, 2024.

[AUR25a] Abu Ahmad, R.; Usmanova, A.; Rehm, G.: The ClimateCheck Dataset: Mapping Social Media Claims About Climate Change to Corresponding Scholarly Articles. In: Proceedings of the 5th Workshop on Scholarly Document Processing (SDP). Vienna, Austria, 2025.

[AUR25b] Abu Ahmad, R.; Usmanova, A.; Rehm, G.: The ClimateCheck Shared Task: Scientific Fact-Checking of Social Media Claims about Climate Change. In: Proceedings of the 5th Workshop on Scholarly Document Processing (SDP). Vienna, Austria, 2025.

[Ba23] Banerjee, D. et al., eds.: Joint Proceedings of Scholarly QALD 2023 and SemREC 2023: 1st Scholarly QALD Challenge 2023 and 4th SeMantic Answer Type, Relation and Entity Prediction Tasks Challenge 2023, vol. 3592, Athens, Greece: CEUR Workshop Proceedings, 2023.

[Ba24] Babaei Giglou, H. et al.: LLMs4OL 2024 Datasets: Toward Ontology Learning with Large Language Models. Open Conference Proceedings 4, pp. 17–30, 2024.

[Ba25] Bayrami Asl Tekanlou, H. et al.: Homa at SemEval-2025 Task 5: Aligning Librarian Records with OntoAligner for Subject Tagging. In: Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025). Association for Computational Linguistics, Vienna, Austria, pp. 2371–2377, 2025.

[BDA23] Babaei Giglou, H.; D'Souza, J.; Auer, S.: LLMs4OL: Large language models for ontology learning. In: International Semantic Web Conference. Springer, pp. 408–427, 2023.

[BLW20] Brüggemann, M.; Lörcher, I.; Walter, S.: Post-normal science communication: exploring the blurring boundaries of science and journalism. Journal of Science Communication 19 (3), A02, 2020.

[BRR25] Borisova, E.; Rauscher, N.; Rehm, G.: SciVQA 2025: Overview of the First Scientific Visual Question Answering Shared Task. In: Proceedings of the 5th Workshop on Scholarly Document Processing (SDP). Vienna, Austria, 2025.

[Ch20] Chandrasekaran, M. K. et al.: Overview of the First Workshop on Scholarly Document Processing (SDP). In: Proceedings of the First Workshop on Scholarly Document Processing. Association for Computational Linguistics, Online, pp. 1–6, 2020.

[DS24] D'Souza, J. et al.: The SemEval 2025 LLMs4Subjects Shared Task Dataset, 2024, https://github.com/jd-coderepos/llms4subjects/, accessed: 07/24/2025.

[DS25] D'Souza, J. et al.: SemEval-2025 Task 5: LLMs4Subjects–LLM-based Automated Subject Tagging for a National Technical Library's Open-Access Catalog. arXiv preprint arXiv:2504.07199, 2025.

[Du21] Dunwoody, S.: Science journalism: Prospects in the digital age. In: Routledge handbook of public communication of science and technology. Routledge, pp. 14–32, 2021.

[Es21] Escartín, C. P. et al.: Towards transparency in NLP shared tasks. arXiv preprint arXiv:2105.05020, 2021.

[Fr25] Francis, M. et al.: Overview of the FoRC@NSLP2025 Shared Task: Field of Research Classification for Computational Linguistics and Natural Language Processing Publications. In: Joint Proceedings of the ESWC 2025 Workshops and Tutorials co-located with 22nd Extended Semantic Web Conference (ESWC 2025). 2025.

[FU18] Filannino, M.; Uzuner, Ö.: Advancing the State of the Art in Clinical Natural Language Processing through Shared Tasks. Yearbook of medical informatics 27 (1), pp. 184–192, 2018.

[GDA24] Giglou, H. B.; D'Souza, J.; Auer, S.: LLMs4OL 2024 Overview: The 1st Large Language Models for Ontology Learning Challenge. In: Open Conference Proceedings. Vol. 4, pp. 3–16, 2024.

[Ge24] Gesese, G. A. et al.: NFDI4DSO: Towards a BFO Compliant Ontology for Data Science. arXiv preprint arXiv:2408.08698, 2024.

[Gu17] Guo, C. et al.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17, JMLR.org, Sydney, NSW, Australia, pp. 1321–1330, 2017.

[Ha22] Hafid, S. et al.: SciTweets-A Dataset and Annotation Framework for Detecting Scientific Online Discourse. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. Pp. 3988–3992, 2022.

[KK25] Kluge, L.; Kähler, M.: DNB-AI-Project at SemEval-2025 Task 5: An LLM-Ensemble Approach for Automated Subject Indexing. In: Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025). Association for Computational Linguistics, Vienna, Austria, pp. 1114–1124, 2025.

[KKD24] Krüger, F.; Karmakar, S.; Dietze, S.: SOMD@NSLP2024: Overview and Insights from the Software Mention Detection Shared Task. In: Natural Scientific Language Processing and Research Knowledge Graphs: First International Workshop, NSLP 2024, Hersonissos, Crete, Greece, May 27, 2024, Proceedings. Springer-Verlag, Hersonissos, Crete, Greece, pp. 247–256, 2024.

[Li04] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain, pp. 74–81, 2004.

[Li25] Lindenmeyer, A. et al.: Towards Trustworthy AI in Healthcare: Epistemic Uncertainty Estimation for Clinical Decision Support. Journal of Personalized Medicine 15 (2), 2025.

[Ot23] Otto, W. et al.: GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets. In (Bouamor, H.; Pino, J.; Bali, K., eds.): Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp. 8166–8176, 2023.

[Pa17] Parra Escartín, C. et al.: Ethical Considerations in NLP Shared Tasks. In (Hovy, D. et al., eds.): Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. Association for Computational Linguistics, Valencia, Spain, pp. 66–73, 2017.

[Ra23] Ragab, M. et al.: Adatime: A benchmarking suite for domain adaptation on time series data. ACM Transactions on Knowledge Discovery from Data 17 (8), pp. 1–18, 2023.

[Re24] Rehm, G. et al., eds.: Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024): First International Workshop. Hersonissos, Crete, Greece, 27 May 2024, Springer, Cham, Switzerland, 2024.

[Ro21] Rocha, Y. M. et al.: The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. Journal of Public Health, pp. 1–10, 2021.

[Sc21] Schindler, D. et al.: SoMeSci- A 5 Star Open Data Gold Standard Knowledge Graph of Software Mentions in Scientific Articles. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. CIKM '21, Association for Computing Machinery, Virtual Event, Queensland, Australia, pp. 4574–4583, 2021.

[Sc23] Schimmler, S. et al.: NFDI4DS infrastructure and services. Gesellschaft für Informatik eV, 2023.

[SIL25] Suominen, O.; Inkinen, J.; Lehtinen, M.: Annif at SemEval-2025 Task 5: Traditional XMTC augmented by LLMs. In: Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025). Association for Computational Linguistics, Vienna, Austria, pp. 2395–2402, 2025.

[Ta24] Taffa, T. A. et al.: Hybrid-SQuAD: Hybrid Scholarly Question Answering Dataset. arXiv preprint arXiv:2412.02788, 2024.

[Up25] Upadhyaya, S. et al.: SOMD2025: A Challenging Shared Tasks for Software Related Information Extraction. In: Proceedings of the 5th Workshop on Scholarly Document Processing (SDP). Vienna, Austria, 2025.

[Vo05] Voorhees, E.: TREC: Experiment and evaluation in information retrieval, 2005.

[Wa20] Wang, L. L. et al.: CORD-19: The COVID-19 Open Research Dataset. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. Association for Computational Linguistics, Online, 2020.

[Xu22] Xu, Z. et al.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns 3 (7), p. 100543, 2022.

[Zh20] Zhang*, T. et al.: BERTScore: Evaluating Text Generation with BERT. In: International Conference on Learning Representations. 2020.