Benchmarking Large Language Models for Sustainable Development Goals Classification: Evaluating In-Context Learning and Fine-Tuning Strategies

Andrea Cadeddu[1,†], Alessandro Chessa[1,†], Vincenzo De Leo[1,2,†], Gianni Fenu[2,†], Enrico Motta[3,†], Francesco Osborne[3,4,†], Diego Reforgiato Recupero[2,*,†], Angelo Salatino[3,†] and Luca Secchi[1,2,†]

[1]*Linkalab s.r.l., Cagliari, Italy*
[2]*Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy*
[3]*Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom*
[4]*Department of Business and Law, University of Milano Bicocca, Milan, Italy*

Abstract

In 2012, the United Nations set 17 Sustainable Development Goals (SDGs) to build a better future by 2030, but monitoring progress is challenging due to data complexity. Recent Large Language Models (LLMs) have significantly improved Natural Language Processing tasks, including text classification. This study evaluates only open-weight LLMs for single-label, multi-class SDG text classification, comparing Zero-Shot, Few-Shot, and Fine-Tuning approaches. Our goal is to determine whether smaller, resource-efficient models, optimized through prompt engineering, can obtain competitive results on a challenging dataset. Using a benchmark dataset from the Open SDG initiative, our findings show that with effective prompt engineering, small models can significantly achieve competitive performance.

Keywords

Sustainable Development Goals, Large Language Models, Text Classification, United Nations

1. Introduction

The Sustainable Development Goals (SDGs) consist of 17 interlinked global objectives established at the 2012 United Nations Conference on Sustainable Development in Rio de Janeiro, serving as a "*blueprint to achieve a better and more sustainable future for all*"[1] and aimed for achievement by 2030 [1, 2]. These goals underpin the 2030 Agenda for Sustainable Development, endorsed unanimously by all United Nations (UN) Member States. Monitoring SDG progress is challenging[2] due to the vast and complex data involved [3, 4, 5]. Manual methods are no longer sufficient; automated text classification has become essential for extracting and categorizing relevant information from reports, news, social media, scientific articles and official documents [6, 7, 8, 9]. Such models enable real-time monitoring, rapid crisis response, and support data-driven decision-making. Recent advances in Large Language Models (LLMs) have transformed Natural Language Processing (NLP) by delivering state-of-the-art performance in text classification, sentiment analysis, and language understanding [10, 11, 12]. In this context, understanding whether the latest LLMs can support toward identifying SDGs is crucial.

In this work, we conduct a comparative study of several proprietary and open-weight LLMs applied to a single-label, multi-class SDG text classification task. Our primary aim is to assess whether

ESWC 2025: Empowering Knowledge through Semantics: From Knowledge Graphs to Neurosemantics, June 01–05, 2025, Portoroz, Slovenia

*Corresponding author.
†These authors contributed equally.
andrea.cadeddu@linkalab.it (A. Cadeddu); alessandro.chessa@linkalab.it (A. Chessa); vincenzo.deleo@linkalab.it (V. D. Leo); fenu@unica.it (G. Fenu); enrico.motta@open.ac.uk (E. Motta); francesco.osborne@open.ac.uk (F. Osborne); diego.reforgiato@unica.it (D. R. Recupero); angelo.salatino@open.ac.uk (A. Salatino); luca.secchi@linkalab.it (L. Secchi)

[1]https://www.un.org/sustainabledevelopment/sustainable-development-goals/
[2]https://unece.org/sites/default/files/2021-04/2012761_E_web.pdf

smaller models, optimized via prompt engineering techniques and requiring fewer resources, can obtain competitive results on a challenging dataset. To this end, we introduce a benchmark dataset from the open-source initiative Open SDG (OSDG[3]) and explore optimization strategies including in-context learning Zero-Shot (ZS), Few-Shot (FS), and Fine-Tuning (FT). Our experiments demonstrate that with proper fine-tuning and optimization, resource-efficient models can achieve competitive performance. In summary, our main contributions are:

- The introduction of a novel benchmark for SDG text classification;
- A preliminary evaluation of several LLMs;
- An in-depth analysis of optimization strategies covering ZS, FS, and FT approaches.

The overarching objective of this project is to transform the OSDG dataset, along with the associated classification labels for each document, into a structured Knowledge Graph, which will be deployed via a SPARQL endpoint. This approach facilitates interoperability with Semantic Web and knowledge graph technologies [13], which have demonstrated significant efficacy in enhancing AI systems across a variety of domains in recent years [14, 15, 16, 17]. Advances in information extraction have led to the development of numerous effective pipelines for constructing knowledge graphs from text, employing either fully automated pipelines [18, 19, 20] or human-in-the-loop methodologies [21, 22]. As part of our methodology, we also intend to leverage a combination of LLMs and SPARQL queries to efficiently retrieve and align classified documents with established SDG taxonomies. This integration will enable real-time monitoring and scalable data-driven decision-making while supporting conversational agents that leverage the knowledge graph for advanced reasoning and contextual understanding [23, 24].

This research was carried out in collaboration with Ovum S.r.l., an Italian startup specializing in Artificial Intelligence, Cloud Computing, and Big Data, focused on developing tools for efficient interpretation of large-scale textual data through the lens of the SDGs. The remainder of the paper is organized as follows. Section 2 presents the literature review. Section 3 describes the task, detailing the benchmark dataset and outlining the characteristics of the employed LLMs. Section 4 explains the experimental setup, and Section 5 presents the results. Finally, Section 6 concludes the paper and offers recommendations for future work.

2. Related Work

We reviewed the literature on SDG classification and related taxonomies, with an emphasis on employing LLMs for text categorization. Specifically, we found that various NLP methods have been used to classify documents in the SDG context, including the development of ontologies and classification models. For example, the OSDG initiative [25] combines features from previous work, such as the keyword ontology by Bautista-Puig et al. [26] and data from the FP7-4-SD project[4], to construct a comprehensive ontology that is then mapped to the Microsoft Academic Graph[5]. A formal Knowledge Organization System (KOS) has also been proposed [27, 28] to model the Global SDG Indicator Framework, covering 17 Goals, 169 Targets, and 231 indicators, while linking to resources such as UNBIS and EuroVoc. Other studies have employed pre-trained deep learning models, like the Universal Sentence Encoder [29], to classify SDG-related texts in legal and other domains, while research in [30] evaluated the Aurora SDG model version 5 against previous systems and the Elsevier model. Commercial efforts are also notable. Collaborations among Springer Nature, Digital Science Consultancy[6], and Dimensions[7] have led to a system that categorizes publications into SDGs using supervised machine learning. Similarly, SDG Juicer[8], developed by Ovum, Linkalab[9], and AB Innovation Consulting[10], uses AI to extract SDGs from

[3] https://open-sdg.org/
[4] https://www.fp7-4-sd.eu/
[5] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
[6] https://www.digital-science.com/product/consultancy/
[7] https://www.dimensions.ai/
[8] https://sdgjuicer.com/
[9] https://www.linkalab.it/
[10] https://www.abinnovationconsulting.com/

corporate documents, thereby aligning business strategies with the UN Agenda 2030.

Over the past five years, LLMs have transformed NLP by processing vast amounts of text from various sources [31, 32, 33, 34, 35, 13, 36]. Text classification, in particular, has benefited from these advances, improving tasks such as sentiment analysis [37], research topic identification [38], intent recognition [39], and automated fact-checking [40]. More broadly, LLMs have played a crucial role in enhancing information extraction pipelines across various domains [41], including engineering [42] and financial sustainability [43]. For instance, Angioni et al. [44] combined linguistic pattern analysis with transformer models to construct a knowledge graph from a collection of news articles, aiming to identify key trends related to ESG factors.

The advent of transformer architectures, initiated by models like BERT and GPT in 2018 [10, 45], has been pivotal, with transformer-based models now dominating top NLP tasks [46]. The rise of platforms such as HuggingFace[11] and the rapid adoption of ChatGPT[12] have further spurred research in text classification. Recent advances have also explored alternative methodologies for SDG classification by leveraging quantized, instruction-tuned LLMs. In particular, Fankhauser et al. [47] proposed the Decompose-Synthesize-Refine-Extract (DSRE) framework, which employs advanced prompt decomposition techniques to break down the classification task into more manageable subtasks. This framework not only enhances the zero-shot capabilities of LLMs for both single-label and multi-label classification but also addresses computational efficiency through model quantization. Their results indicate that, even with minimal fine-tuning, such strategies can achieve competitive performance while significantly reducing computational overhead. These insights further underscore the potential of incorporating instruction tuning and prompt engineering to overcome data imbalance and scalability challenges in automated SDG classification. Despite these advancements, challenges remain, including high computational costs and the need for robust hardware, which restrict access to state-of-the-art LLMs [48, 49]. In light of these issues, a comparative evaluation of LLMs specifically for SDG classification is still lacking, an issue that our work aims to address.

3. Background

The experiments described in this work were performed using 3 LLMs on data from the OSDG Community Dataset[13] (OSDG-CD), a dataset obtained thanks to the collaborative work of a thousand volunteers distributed in more than 110 countries, who validated thousands of text extracts for training models on SDG classification tasks; the October 2023 version, used in this work, includes more than 40k excerpts with more than 300k labels, each averaging around 90 words and drawn from public documents (including UN sources such as SDG-Pathfinder[14] and SDG Library[15]). Nine volunteers analyzed each text and consensus was measured by an "agreement" score defined as $|LP - LN|/(LP + LN)$, where L stands for "Label", P stands for "Positive" and N stands for "Negative". In a preliminary dataset cleaning phase, only texts with at least 3 validations, a positive-negative ratio greater than 2:1, and an agreement greater than 0.75 were selected. Since the original dataset contains texts related only to SDGs 1 to 16, adding another 400 texts to represent the category "Other" (which, for uniformity of nomenclature, we called "SDG 0") was necessary. The final balanced dataset was then divided into a training set of 4,760 texts (70%), and validation and test set both of 1,020 texts (15%), all equally distributed among the 17 classes mentioned above.

The three LLMs that were used in the experiments described in this work are: i) **Llama-2-7b-chat-hf**[16], ii) **Mistral-7B-instruct**[17] and iii) **Phi-3-mini-4k-instruct**[18]. These LLMs were selected because,

[11] https://huggingface.co/

[12] https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

[13] https://zenodo.org/records/5550238

[14] https://sdg.iisd.org/news/oecd-tool-applies-sdg-lens-to-international-organizations-policy-content/

[15] https://www.sdglibrary.ca/

[16] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[17] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[18] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

Table 1
Results (in percentages) of the experiments when employing ZS learning. Values are in percentages.

MODEL NAME	Pre	Rec	Acc	F1
Llama-2-7b-chat-hf	63.3	54.7	54.7	53.8
Mistral_7B_instruct	64.3	55.6	55.6	51.6
Phi-3-mini-4k-instruct	**69.7**	**59.2**	**59.2**	**59.8**

Table 2
Results (in percentages) of the experiments when employing FS learning with 3 random examples.

MODEL NAME	Pre	Rec	Acc	F1
Llama-2-7b-chat-hf	61.4	48.1	48.1	48.5
Mistral_7B_instruct	66.5	58.2	58.2	54.9
Phi-3-mini-4k-instruct	**75.0**	**72.6**	**72.6**	**72.7**

at the onset of this study, they were among the highest-performing open-weight models in classification tasks, and the diversity in their architectures reflected the state-of-the-art in open-weight generative AI at that time. *LLaMa-2* is a decoder-only Transformer developed by Meta and available in sizes from 7B to 70B parameters [50]. Its Chat variants are optimized via Supervised Fine-Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF) on a 2T token corpus[19] and public training datasets. One of its strengths is Grouped-Query Attention (GQA) [51], which significantly optimizes inference scalability. *Mistral-7B* is the first decoder-only model developed by the French company Mistral AI, founded in 2023 [52]. Among the key innovations are the Sliding Window Attention mechanism, which supports up to 8,000 tokens and a fixed cache of 128,000 tokens, grouped query attention, and a BPE Byte-fallback tokenizer. The *Phi-3* models were developed by Microsoft in 2024 and offer high performance in a compact decoder-only design [53]. The smallest model, with nearly 4 billion parameters, supports large context windows of up to 128K tokens without any quality loss. These models are instruction-optimized and built according to the Microsoft Responsible AI standard, making them particularly suitable for resource-constrained environments.

4. Experimental evaluation

The primary aim of this study was to assess the 3 LLMs cited in the previous sections using the OSDG Community Dataset benchmark discussed above for SDG text classification. We employed three learning approaches: ZS learning, FS learning, and FT.

In the ZS learning setting, models processed the test set with a basic prompt (no examples), while in FS learning the prompt was augmented with 3 random training texts. For FT, the pre-trained models underwent an additional training phase using the training and validation sets from the OSDG-CD (see Section 3); evaluation was then performed on the corresponding test set.

All prompt templates are available at https://github.com/vincenzodeleo/sdg_classification_prompts.

For FT, standard parameter settings were used. Specifically, Llama-2 was fine-tuned for 1 epoch with a learning rate of $2 \cdot 10^{-4}$, batch size 4, LoRA attention dimension 64, LoRA scaling factor 16, dropout 0.1, using float16 for 4-bit base models with nf4 quantization. Mistral was fine-tuned for 10 epochs with a learning rate of 10^{-5} and batch size 8, whereas Phi-3 was fine-tuned for 1 epoch with a learning rate of $5 \cdot 10^{-6}$, batch size 4, LoRA scaling factor 32, and dropout 0.05.

Table 3
Results (in percentages) of the experiments when employing FT. Values are in percentages.

MODEL NAME	Pre	Rec	Acc	F1
Llama-2-7b-chat-hf	86.8	85.3	85.3	85.5
Mistral_7B_instruct	**88.0**	**88.2**	**88.2**	**88.1**
Phi-3-mini-4k-instruct	70.7	62.7	62.7	63.5

5. Results

In this section, we present the experimental outcomes for the 3 models Llama-2-7b-chat-hf, Mistral-7B-instruct, and Phi-3-mini-4k-instruct, evaluated using ZS learning, FS learning with 3 random examples, and FT. Performance was assessed via macro-averaged Precision, Recall, Accuracy, and F1-score.

Tables 1 and 2 show the ZS and FS results, respectively. Notably, the Phi-3 model achieved 59.8% in ZS and 72.7% in FS, indicating its strong performance in settings with minimal examples probably thanks to its optimized architecture and training methodologies on optimized dataset. This finding aligns with the result from [53], where Phi-3 is demonstrated to outperform larger models in specific benchmarks despite being trained on fewer parameters

Table 3 presents the FT results, where all models (except Phi-3) improved significantly, with Mistral attaining an F1-score of 88.1%. These results are consistent with findings already shown in other experiments[20], which claim that the fine-tuned Mistral-7B can also outperform GPT-4.

6. Conclusions

The experiments shown in this work demonstrate that open-weight LLMs performance in text classification against SDGs can be significantly improved through prompt engineering and fine-tuning strategies. In zero-shot and few-shot settings, the Phi-3-mini-4k-instruct model consistently outperformed the other two models under evaluation, achieving F1-scores of around 60% and 70%, respectively; this was possible thanks to its optimized architecture and efficient training methodologies, which provide a competitive advantage in low-information situations. On the other hand, fine-tuning significantly improved the performance of the Mistral-7B-instruct model, which managed to achieve F1-scores of almost 90%; this result underlines the effectiveness of domain-specific supervised training. These results not only highlight the potential of small open-weight LLMs as valid alternatives to proprietary models that are vastly larger in terms of parameters, but also highlight the importance of tailoring prompting and training strategies to the specific requirements of the task at hand. Building on the promising results obtained in our current experiments, future work will extend our investigation in two significant directions. First, we plan to broaden the portfolio of open-weight LLMs by incorporating additional models, such as BERT [10], T5 [54], Mixtral [55], Zephir [56] and also to other state-of-the-art models such as DeepSeek [57], to evaluate whether these alternatives can further enhance performance in SDG text classification tasks under various learning paradigms (zero-shot, few-shot, and fine-tuning). Second, we intend to integrate experiments with proprietary models, like OpenAI's GPT-4 [58] and DistilGPT-2 [59], to conduct a direct and comprehensive comparison between smaller, resource-efficient models and their larger, closed-weight counterparts. This expanded study aims to provide deeper insights into the trade-offs between efficiency, scalability, and domain-specific performance, ultimately guiding the selection of the most suitable model for sustainable development applications.

[19] https://llama.meta.com/llama2/

[20] https://agentissue.medium.com/mistral-7b-outperforms-gpt-4-in-specialized-tasks-d75fec6803e2
https://predibase.com/blog/lora-land-fine-tuned-open-source-llms-that-outperform-gpt-4

Acknowledgments

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No.3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] G. Schmidt-Traub, et al., Indicators and a monitoring framework for the sustainable development goals, 2015. URL: https://sdgs.un.org/publications/indicators-and-monitoring-framework-sustainable-development-goals-17958.

[2] J. Espey, et al., Data for development - a needs assessment for sdg monitoring and statistical capacity development, 2015. URL: https://sdgs.un.org/publications/data-development-needs-assessment-sdg-monitoring-and-statistical-capacity-development.

[3] T. Hirai, F. Comim, Measuring the sustainable development goals: A poset analysis, Ecological Indicators 145 (2022) 109605. URL: https://www.sciencedirect.com/science/article/pii/S1470160X22010780. doi:10.1016/j.ecolind.2022.109605.

[4] D. Ziegler, S. Wolff, A.-B. Agu, G. Cortiana, M. Umair, F. de Durfort, E. Neumann, G. Walther, J. Kristiansen, M. Lienkamp, How to measure sustainability? an open-data approach, Sustainability 15 (2023) 3203–. URL: https://ideas.repec.org/a/gam/jsusta/v15y2023i4p3203-d1063481.html.

[5] M. Mishra, S. Desul, C. A. G. Santos, S. K. Mishra, A. H. M. Kamal, S. Goswami, A. M. Kalumba, R. Biswal, R. M. da Silva, C. A. C. dos Santos, K. Baral, A bibliometric analysis of sustainable development goals (sdgs): a review of progress, challenges, and opportunities, Environment, Development and Sustainability (2023) 1–43. doi:10.1007/s10668-023-03225-w.

[6] J. E. Guisiano, R. Chiky, J. D. Mello, Sdg-meter: A deep learning based tool for automatic text classification of the sustainable development goals, in: N. T. Nguyen, T. K. Tran, U. Tukayev, T.-P. Hong, B. Trawiński, E. Szczerbicki (Eds.), Intelligent Information and Database Systems, Springer International Publishing, Cham, 2022, pp. 259–271.

[7] T. Matsui, K. Suzuki, K. Ando, Y. Kitai, C. Haga, N. Masuhara, S. Kawakubo, A natural language processing model for supporting sustainable development goals: translating semantics, visualizing nexus, and connecting stakeholders, Sustainability Science 17 (2022) 969–985. URL: https://doi.org/10.1007/s11625-022-01093-3. doi:10.1007/s11625-022-01093-3.

[8] J. Guisiano, R. Chiky, Automatic classification of multilabel texts related to sustainable development goals (sdgs), in: TECHENV EGC2021, Montpellier, France, 2021. URL: https://hal.archives-ouvertes.fr/hal-03154261.

[9] A. Salatino, F. Osborne, E. Motta, Cso classifier 3.0: a scalable unsupervised method for classifying documents in terms of research topics, International Journal on Digital Libraries (2022) 1–20.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[11] T. Brown, et al., Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[12] J. Lehmann, A. Meloni, E. Motta, F. Osborne, D. R. Recupero, A. A. Salatino, S. Vahdati, Large language models for scientific question answering: An extensive analysis of the sciqa benchmark, in: European Semantic Web Conference, Springer, 2024, pp. 199–217.

[13] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: opportunities and challenges, Artificial Intelligence Review (2023) 1–32.

[14] D. Fensel, U. Şimşek, K. Angele, E. Huaman, E. Kärle, O. Panasiuk, I. Toma, J. Umbrich, A. Wahler, Knowledge Graphs Methodology, Tools and Selected Use Cases, 2020. URL: http://lib.ugent.be/catalog/ebk01:4100000010122122.

[15] A. Borrego, D. Dessì, D. Ayala, I. Hernández, F. Osborne, D. R. Recupero, D. Buscaldi, D. Ruiz, E. Motta, Research hypothesis generation over scientific knowledge graphs, Knowledge-Based Systems 315 (2025) 113280.

[16] D. Greco, F. Osborne, S. Pusceddu, D. Reforgiato Recupero, Modelling big data platforms as knowledge graphs: the data platform shaper, Journal of Big Data 12 (2025) 64.

[17] A. Chessa, G. Fenu, E. Motta, F. Osborne, D. R. Recupero, A. Salatino, L. Secchi, Data-driven methodology for knowledge graph generation within the tourism domain, IEEE Access 11 (2023) 67567–67599.

[18] 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) 109945.

[19] L. Zhong, J. Wu, Q. Li, H. Peng, X. Wu, A comprehensive survey on automatic knowledge graph construction, ACM Computing Surveys 56 (2023) 1–62.

[20] P. Manghi, A. Mannocci, F. Osborne, D. Sacharidis, A. Salatino, T. Vergoulis, New trends in scientific knowledge graphs and research impact assessment, 2021.

[21] S. Tsaneva, D. Dessì, F. Osborne, M. Sabou, Knowledge graph validation by integrating llms and human-in-the-loop, Information Processing & Management 62 (2025) 104145.

[22] A. Brack, A. Hoppe, M. Stocker, S. Auer, R. Ewerth, Analysing the requirements for an open research knowledge graph: use cases, quality requirements, and construction strategies, International Journal on Digital Libraries 23 (2022) 33–55.

[23] A. Meloni, S. Angioni, A. Salatino, F. Osborne, D. Reforgiato Recupero, E. Motta, Integrating conversational agents and knowledge graphs within the scholarly domain, IEEE Access 11 (2023) 22468–22489. doi:10.1109/ACCESS.2023.3253388.

[24] X. Ren, T. Chen, Q. V. H. Nguyen, L. Cui, Z. Huang, H. Yin, Explicit knowledge graph reasoning for conversational recommendation, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–21.

[25] L. Pukelis, N. Bautista-Puig, M. Skrynik, V. Stanciauskas, OSDG - open-source approach to classify text data by UN sustainable development goals (sdgs), CoRR abs/2005.14569 (2020). URL: https://arxiv.org/abs/2005.14569. arXiv:2005.14569.

[26] N. Bautista, SDG ontology (2019). URL: https://figshare.com/articles/dataset/SDG_ontology/11106113. doi:10.6084/m9.figshare.11106113.v1.

[27] A. Joshi, L. G. G. Morales, S. Klarman, A. Stellato, A. Helton, C. S. Lovell, A. Haczek, A knowledge organization system for the united nations sustainable development goals, in: Eighteenth Extended Semantic Web Conference - Resources Track, 2021. URL: https://openreview.net/forum?id=6Y_I9qc6JBD.

[28] A. Salatino, T. Aggarwal, A. Mannocci, F. Osborne, E. Motta, A survey on knowledge organization systems of research fields: Resources and challenges, Quantitative Science Studies (2025) 1–37.

[29] F. Sovrano, M. Palmirani, F. Vitali, Deep learning based multi-label text classification of unga resolutions, in: Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance, ICEGOV '20, Association for Computing Machinery, New York, NY, USA,

2020, pp. 686–695. URL: https://doi.org/10.1145/3428502.3428604. doi:10.1145/3428502.3428604.
[30] F. Schmidt, M. Vanderfeesten, Evaluation on accuracy of mapping science to the united nations' sustainable development goals (sdgs) of the aurora sdg queries, 2021. URL: https://doi.org/10.5281/zenodo.4964606. doi:10.5281/zenodo.4964606.
[31] F. Bolanos, A. Salatino, F. Osborne, E. Motta, Artificial intelligence for literature reviews: Opportunities and challenges, Artificial Intelligence Review 57 (2024).
[32] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, R. Daneshjou, Large language models in medicine: the potentials and pitfalls: a narrative review, Annals of internal medicine 177 (2024) 210–220.
[33] C. W. Kosonocky, C. O. Wilke, E. M. Marcotte, A. D. Ellington, Mining patents with large language models elucidates the chemical function landscape, Digital Discovery 3 (2024) 1150–1159.
[34] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, S. Ananiadou, Mentallama: interpretable mental health analysis on social media with large language models, in: Proceedings of the ACM Web Conference 2024, 2024, pp. 4489–4500.
[35] E. Motta, F. Osborne, M. M. Pulici, A. Salatino, I. Naja, Capturing the viewpoint dynamics in the news domain, in: International Conference on Knowledge Engineering and Knowledge Management, Springer, 2024, pp. 18–34.
[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.
[37] M. S. U. Miah, M. M. Kabir, T. B. Sarwar, M. Safran, S. Alfarhood, M. Mridha, A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm, Scientific Reports 14 (2024) 9603.
[38] A. Cadeddu, A. Chessa, V. D. Leo, G. Fenu, E. Motta, F. Osborne, D. R. Recupero, A. Salatino, L. Secchi, Optimizing tourism accommodation offers by integrating language models and knowledge graph technologies, Information 15 (2024) 398.
[39] P. Wang, K. He, Y. Wang, X. Song, Y. Mou, J. Wang, Y. Xian, X. Cai, W. Xu, Beyond the known: Investigating llms performance on out-of-domain intent detection, arXiv preprint arXiv:2402.17256 (2024).
[40] L. Tang, P. Laban, G. Durrett, Minicheck: Efficient fact-checking of llms on grounding documents, arXiv preprint arXiv:2404.10774 (2024).
[41] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, H. Chen, Z. Liu, Z. Dou, J.-R. Wen, Large language models for information retrieval: A survey, arXiv preprint arXiv:2308.07107 (2023).
[42] T. Aggarwal, A. Salatino, F. Osborne, E. Motta, Large language models for scholarly ontology generation: An extensive analysis in the engineering field, arXiv preprint arXiv:2412.08258 (2024).
[43] M. Birti, F. Osborne, A. Maurino, Optimizing large language models for esg activity detection in financial texts, arXiv preprint arXiv:2502.21112 (2025).
[44] S. Angioni, S. Consoli, D. Dessí, F. Osborne, D. R. Recupero, A. Salatino, Exploring environmental, social, and governance (esg) discourse in news: An ai-powered investigation through knowledge graph analysis, IEEE Access (2024).
[45] A. Radford, K. Narasimhan, Improving language understanding by generative pre-training, 2018. URL: https://api.semanticscholar.org/CorpusID:49313245.
[46] J. Fields, K. Chovanec, P. Madiraju, A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe?, IEEE Access 12 (2024) 6518–6531. doi:10.1109/ACCESS.2024.3349952.
[47] T. Fankhauser, S. Clematide, SDG classification using instruction-tuned LLMs, in: C. Corsin, C. Mark, W. Albert, M. Claudiu, M. Elisabeth, Z. Lucas (Eds.), Proceedings of the 9th edition of the Swiss Text Analytics Conference, Association for Computational Linguistics, Chur, Switzerland, 2024, pp. 148–156. URL: https://aclanthology.org/2024.swisstext-1.13/.
[48] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, X. Hu, Harnessing the power of llms in practice: A survey on chatgpt and beyond (2024). URL: https://doi.org/10.1145/3649506. doi:10.1145/3649506, just Accepted.
[49] E. Bisong, Building Machine Learning and Deep Learning Models on Google Cloud Platform, 1 ed.,

Apress, Berkeley, CA, 2019. doi:10.1007/978-1-4842-4470-8, published: 28 September 2019.
[50] H. Touvron, et al., Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.
[51] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebron, S. Sanghai, GQA: Training generalized multi-query transformer models from multi-head checkpoints, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 4895–4901. URL: https://aclanthology.org/2023.emnlp-main.298. doi:10.18653/v1/2023.emnlp-main.298.
[52] A. Q. Jiang, et al., Mistral 7b, 2023. arXiv:2310.06825.
[53] M. I. Abdin, et al., Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, Technical Report MSR-TR-2024-12, Microsoft, 2024. URL: https://www.microsoft.com/en-us/research/publication/phi-3-technical-report-a-highly-capable-language-model-locally-on-your-phone/.
[54] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, 2022. arXiv:2210.11416.
[55] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mixtral of experts, 2024. arXiv:2401.04088.
[56] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, T. Wolf, Zephyr: Direct distillation of lm alignment, 2023. arXiv:2310.16944.
[57] DeepSeek-AI, D. Guo, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL: https://arxiv.org/abs/2501.12948. arXiv:2501.12948.
[58] OpenAI, J. Achiam, et al., Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.
[59] M. Ottaviani, S. Stahlschmidt, On the performativity of sdg classifications in large bibliometric databases, 2024. URL: https://arxiv.org/abs/2405.03007. arXiv:2405.03007.