# Framing Few-Shot Knowledge Graph Completion with Large Language Models

Adrian M.P. Braşoveanu[1,2,*], Lyndon J.B. Nixon[1,2], Albert Weichselbraun[3,4] and Arno Scharl[1,3]

[1]*Modul University Vienna, Am Kahlenberg 1, 1190, Vienna, Austria*
[2]*Modul Technology GmbH, Am Kahlenberg 1, 1190, Vienna, Austria*
[3]*webLyzard technology GmbH, Liechtensteinstrasse 41/26, 1090 Vienna, Austria*
[4]*University of Applied Sciences of the Grisons, Pulvermühlestrasse 57, CH-7004 Chur, Switzerland*

**Abstract**

Knowledge Graph Completion (KGC) from text involves identifying known or unknown entities (nodes) as well as relations (edges) among these entities. Recent work has started to explore the use of Large Language Models (LLMs) for entity detection and relation extraction, due to their Natural Language Understanding (NLU) capabilities. However, LLM performance varies across models and depends on the quality of the prompt engineering. We examine specific relation extraction cases and present a set of examples collected from well-known resources in a small corpus. We provide a set of annotations and identify various issues that occur when using different LLMs for this task. As LLMs will remain a focal point of future KGC research, we conclude with suggestions for improving the KGC process.

**Keywords**

knowledge graph completion, slot filling, relation extraction, large language models,

## 1. Introduction

Building a Knowledge Graph (KG) is a time-consuming process. While dumps from mainstream KGs like Wikidata or DBpedia are freely available, they are not complete (entities and relations are missing, especially when considering domain-specific or regional use cases) nor are they consistent in their knowledge modelling (e.g., use of different properties for the same purpose, the existence inverse relations, etc.). This means that further steps must be taken to maintain a graph and ensure it is correct and complete regarding its entities and relations. One needs to carefully consider which entities to include, which kinds of properties and relations, as well as various update scenarios to ensure the graph remains consistent (e.g., the inclusion of temporal restrictions on relations). Given that public KGs are unlikely to provide the full coverage that is required by a local, domain-specific knowledge graph, large collections of online text documents

*SEMANTICS 2023 EU: 19th International Conference on Semantic Systems, September 20-22, 2023, Leipzig, Germany*
*Corresponding author.
adrian.brasoveanu@modul.ac.at (A. M.P. Braşoveanu); lyndon.nixon@modul.ac.at (L. J.B. Nixon); albert.weichselbraun@fhgr.ch (A. Weichselbraun); arno.scharl@modul.ac.at (A. Scharl)
0000-0002-3439-4736 (A. M.P. Braşoveanu); 0000-0001-7091-4543 (L. J.B. Nixon); 0000-0001-6399-045X (A. Weichselbraun); 0000-0001-5346-9521 (A. Scharl)

(e.g. Wikipedia articles, news stories or social media posts) present themselves as an alternative source to identify and include entities and relations as part of a KG completion (KGC) task.

This work started with a KGC task for capturing information about public communication around the Sustainable Development Goals (SDGs). The SDGs cover domains such as social action, health, energy, economy, sustainability, climate, biodiversity, and politics. While we used Wikidata to seed an initial knowledge graph, we quickly found that various people, organisations, and locations were missing, especially in the national and regional context. In Austria, for example, we could not find the 'Schlichtungsstelle Österreich' (organisation), 'Markus Klamminger' (person, director of a health agency) or the 'Recyclingcenter Himberg' (location and organisation!). Furthermore, relations of importance to us are either inconsistently used or missing, for example, 'position held' by individuals, the 'objective' of social movement organisations or the 'operator' of a location of energy infrastructure.

While these entities and relations may not occur in global KGs, they are of importance to a complete understanding of SDG communication at the regional level. While previously KGC approaches have relied on various NLP algorithms to identify candidates both for entities and relations between entities in unstructured text, the emergence of LLMs trained on huge amounts of online text has led to an impressive level of Natural Language Understanding (NLU) demonstrated beyond the state of the art in various benchmarks. It seems reasonable to assume LLMs might also improve on the state of the art in KGC and first explorations of LLMs like Flan-T5 or GPT-3 in entity and relation extraction suggest their effectiveness [1].

Taking this work a step further, we examine specific cases of missing relations in our data that are a bit more difficult to extract in KGC. We present a small set of examples that are also gathered in a corpus and identify the various issues discovered when using different LLMs for this task. The examples discussed in this work are based on English documents, but we plan to extend our collection to multiple languages, the German language being the next in line. We close with some thoughts on improving the KGC process using LLMs in the future.

## 2. Related Work

Relation extraction (RE) is a vast field, covered by many articles and surveys. We, therefore, start the discussion of related work with several surveys that address topics related to LLMs and KGs.

A recent survey of large language models [2] splits them into public and closed source models, provides a timeline of their recent history, and lists of commonly used resources (e.g., extensive collections of web resources, conversations, scientific data or code). The survey by Hogan et al. [3] discusses the state of the art on knowledge graphs and provides a good overview of the problems that can be solved with KGs, as well as the open research challenges in the field. Pan et al. [4] discuss approaches for unifying knowledge extraction tasks using LLMs and KGs. They review both augmented LLMs and KG-enhanced LLMs, as well as the synergies between LLMs and KGs.

An early survey focused on RE [5] also elaborates on the state of the art before the emergence of LLMs, as it also surveys the various datasets available at the time of its writing (circa 2020).

Shen et al. [6] survey KGC, identifying two large areas of contributions: embeddings and

text-based. KG embeddings might contain both the entities and relations, or just the entities. The evaluations of such methods typically use graph datasets that only contain entities and relations instead of textual datasets; therefore, we are less interested in them for the current work, as we want to observe how well missing relations can be extracted from texts.

We restrict the rest of this section to reviewing work that is closer to our contribution, namely: (i) strategies for using large language models for knowledge extraction tasks, and (ii) filling in missing relations in a knowledge graph.

Chain of Thought (CoT) [7], one of the earliest prompt mechanisms, is based on the idea of providing the entire chain of reasoning behind arriving at a solution. Such a mechanism for problem-solving makes sense for any kind of problem that does not have a simple answer, regardless of the domain. For relation extraction, one quick method to implement this mechanism consists of adding explanations to the few-shot examples fed to an LLM. When only provided with a set of relations, the LLM might start creating numerous relations, even with undefined entities (e.g., entities defined through undefined pronouns), whereas if explanations are added, it is possible to restrict the number of extracted relations, as it will now only pick relations between clearly defined entities. A longer analysis of the CoT mechanisms is presented in [8], LLMs being considered greedy reasoners that have difficulties picking up the right answer when multiple solutions are available. Several alternatives to CoT have recently appeared. ReACT [9] combines reasoning and acting, generally leading to better results than classic CoT. Tree of Thoughts (ToT) [10] expands on the ideas from the previous papers, generalizes the CoT mechanism, and allows for multiple paths of reasoning, all intermediate actions being considered separate thoughts, similar to how humans think. Besides the different problem-solving styles reflected by mechanisms like CoT or ToT, it is also important to phrase the prompts in a certain manner that will lead the LLM towards a certain type of solution. The effect of different prompt patterns is discussed in several papers about their usage in common tasks [11] or code generation tasks [12]. Since LLMs are now capable of using elaborate problem-solving mechanisms, a frequent question that arises is if we can completely remove human evaluators from the chain. According to a recent publication [13], the scores for the high-rated answers for humans and LLMs are similar, and therefore, it seems LLM evaluations can indeed be fully automated.

Using LLMs for KGC is a topic of its own. Since we are mostly interested in sustainability, one of the works that drew our attention was [14]. The paper examined how to use pre-trained models like REBEL [15] compared to foundational LLMs like ChatGPT to generate a sustainability KG. The two experiments in the paper examine sustainability concept extraction and sustainability ontology generation using ChatGPT. The main findings of the paper are included in a set of principles formulated after a qualitative evaluation. The principles cover both the data preprocessing (e.g., context should not be captured, multi-field data sources, mapping of synonyms to help with disambiguation) and organization of the KG (e.g., triples, scalability, avoidance of contradictions, updating rules). Another important question is: How much additional (or auxiliary in-context or out-of-context) information or explanation is needed when implementing various ChatGPT heuristics? Partially, this question is addressed in [16], but only in a NER setting. Two recent articles suggest that providing additional explanations in the few-shot examples can significantly improve the results. Dunn [17] showcases how to build CoT prompts for extracting scientific relations from texts. A recent article [1] considers

**Table 1**
Relation extraction scores, including common missing relations categorized by their main entity type.

| Entity Type | Relation | Condition |
|---|---|---|
| Person | employer | - |
| | field of work | - |
| | jurisdiction | where appropriate, e.g., judge |
| | member of party | where appropriate, e.g. politician |
| | occupation | - |
| | position held | qualified by where and when |
| Organisation | director | - |
| | field of work | - |
| | jurisdiction | where appropriate, e.g., legal and politics |
| | member of | - |
| | part of | qualified by country or region |
| | objective | where appropriate, e.g. social movement |
| | opposing | where appropriate, e.g., social movement |
| Location | country | - |
| | located in | for location included in other location |
| | operator | where appropriate, e.g., energy infrastructure |
| | part of | for location type associated with the same type |
| Event | location | - |
| | main subject | - |
| | organizer | - |
| | part of | if part of event series |
| | point of time | this can be a single point or range |
| | interval | - |

the problem of relation extraction with LLMs under different levels of supervision. Few-shot settings on GPT-3 are found to be closer to state-of-the-art (SOTA) performance, but adding CoT explanations to Flan-T5 is found to lead to SOTA results. Since our knowledge graph completion task can be seen as a restricted version of their task, in which we are required to only extract a limited number of pre-defined relations, we started our exploration of various settings with the CoT-style explanations.

## 3. Method

During the process of building our internal knowledge graph for SDG communication, we assessed information from various sources, including Wikidata and online text such as news stories. Table 1 presents a set of common missing relations from Wikidata and the various additional conditions or information that might need to be considered when collecting such relations from textual data sources. In some cases, data needs to be collected only if it is appropriate, whereas in other cases it needs to be qualified by certain attributes (e.g., location and time or country and region).

**Table 2**
Example inputs for the tasks.

| Task Type | Example Input |
|---|---|
| Relations | Text, Relations |
| Explanations | Text, Relations, Explanations |
| KGC | Text, Relations (only missing), Explanations |
| Self-scoring | Any of the output from the previous tasks |

After identifying these typically missing relations, we created a small relation extraction gold standard that specifically focuses on the observed problems. The corpus examined in this work draws upon English-speaking sentences harvested from well-known and trusted sources such as the New York Times and The Guardian. Provenance data (e.g., URLs of the original articles) was collected for each sentence. The example sentences have been selected from articles related to sustainability, clean energy, climate change, law, and politics. The examples were selected based on the list of relations described in 1. Several examples were selected for each relation. The annotated gold standard dataset is available on our GitHub page[1] together with prompts and example runs from various LLMs. Currently, 12 sentences were used as examples for the few-shot training, as this was the minimal number suggested in [1] and 50 sentences for evaluating the results. The sentences were annotated collaboratively, as we wanted to improve our annotation process over time.

The relation extraction task with an LLM can be modelled as a seq2seq (sequence to sequence) problem, in which a text input known as a prompt is used to seed responses or completions from LLM pipelines [17]. The prompts themselves can be raw text or include more complicated data structures (e.g., texts with annotations, texts with explanations, JSON objects, etc.). The collected sentences were annotated with relations and simple explanations (CoT) to be used as few-shot training examples with the LLM.

We have selected various LLMs for an initial evaluation based on their inclusion in the Chatbot Arena Leaderboard [2] and their general performance in live applications that integrated chatbots for each of these models[3]. Only the models that showcased promising results in an early-stage evaluation with several examples were included in our evaluation. Most of the models selected also had public APIs, which we considered important to ensure later reproducibility.

The structure of the evaluation is based on four task types. Table 2 showcases the inputs for the tasks, whereas Table 3 presents the prompts used for the evaluation. Three of the tasks require the model to produce output based on specific prompts, whereas the last task requires the model to examine its answers against a small golden standard.

The approach we used was to prompt the LLM with examples of 'correct' responses for some sentences (as recommended in [1]), then give it batches of new sentences and ask it to respond.

[1]https://github.com/modultechnology/few-shot-kgc-with-llms/
[2]https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard
[3]The initial list included https://chat.lmsys.org/, https://chat.petals.dev/ and https://nat.dev/. All the runs discussed in this paper were generated using https://nat.dev/, as after testing we considered it to be safe, reliable, and consistent for handling larger runs.

**Table 3**
The prompts used for the general evaluation. X = [Person, Location, Organisation, Event, Date, Work, Other]. Y = [Per, Loc, Org, Event, Date, Work, Other]. Z = [employer, field of work, jurisdiction, member of, occupation, position, director, part of, objective, opposes, country, located in, operator, location, main subject, organizer, point of time] . [EXAMPLES] represents a list of 12 examples.

| Task Type | Prompt |
|---|---|
| Relations | List relations among the entities [X] in the given text and provide a reasonable explanation. Make sure to mark each entity with its abbreviated name [Y]. Relations should be expressed as triples that have the form [subject, predicate, object]. Use the following texts as examples. [EXAMPLES] Using the previous texts as examples, compute the relations for the following texts |
| Explanations | List relations among the entities [X] in the given text and provide a reasonable explanation. Make sure to mark each entity with its abbreviated name [Y]. Relations should be expressed as triples that have the form [subject, predicate, object]. Provide a reasonable explanation for the relations extracted, as presented in the following examples. Provide a reasonable explanation for the relations extracted, as presented in the following examples. [EXAMPLES] Using the previous texts as examples, compute the relations for the following texts |
| Completions | List relations of types [Z] among the entities [X] in the given text and provide a reasonable explanation. Make sure to mark each entity with its abbreviated name [Y]. Relations should be expressed as triples that have the form [subject, predicate, object]. Provide a reasonable explanation for the relations extracted, as presented in the following examples. Provide a reasonable explanation for the relations extracted, as presented in the following examples. [EXAMPLES] Using the previous texts as examples, compute the relations for the following texts |
| Self-scoring | Please compare the previous answers with the following list and compute the F1 score, considering that the newly provided list represents the gold standard. |

If the models were able to generate answers for all sentences, then a single batch was used; otherwise, we used batches of 5 to 10 documents. Each batch was opened in a new chat so that the runs with multiple batches could be considered equivalent to single-batch runs.

## 4. Experiments and Discussion

Our experiments focused on models developed by OpenAI, Anthropic, and Mosaic. For OpenAI we included both 3.5 and 4.0 models.

ChatGPT[4] or GPT-3.5, the latest public iteration of the GPT3+ series [18], is an LLM with

[4]https://openai.com/blog/chatgpt

175B parameters designed to generate human-like responses for a wide variety of domains. The context window of GPT-3.5 is generally known to be around 4096 tokens, whereas for GPT-4 it can get to 32K tokens. In special cases, it can also be extended through prompt engineering.

Claude[5] is an LLM designed by the Anthropic team using their Constitutional AI philosophy [19]. The main idea behind this philosophy is to create harmless AIs with minimal human supervision. The total number of parameters is undisclosed. However, the model has the largest context window (up to 100k) and was designed to work similarly to ChatGPT.

Finally, Mosaic's MPT-30B[6] is pre-trained on a 1T token corpus and has a context window of 8K.

We preferred to use a single testing interface for the current experiment, therefore well-known models, including LLaMa [20], Bloom [21] and Flan-T5 [22] were not included due to the lack of public interfaces which support larger context windows for them. Future experiments will change the testing model to allow for both models that need to be deployed locally and for models that can be accessed through public interfaces.

Each task was repeated multiple times, and we generally followed the principle of saving several runs (e.g., generally 3 runs for each task). Whenever possible, we opted for runs on different days. We kept the same parameters for the runs. In principle, these were the default parameters provided by the interface with minimal changes (e.g., temperature was set to 1 or as close to 1 if the default for a model was not 1). Despite including a reference to explanations on the relations run, no explanations were generated for the relations runs. This was as expected, and it indicated that explanations were only generated when the few-shot examples included them. The explanation runs included the explanations, but the texts for them varied widely depending on the models. The completion runs that only generated the missing relations have led to fewer overall relations, but the models have also generated relations with similar meanings and different names, usually called aliases. The self-scoring runs have often resulted in the model's refusal to generate or explain their scores. Table 4 presents a set of differences between the provided gold and a single set of model runs for a given sentence.

After several runs, a list of frequent issues emerged. These issues are presented in the following list, and examples for each of them are provided in Table 5:

1. *Inverse relation.* Instead of generating relations of the form (s1, pred, s2), the model generates inverse relations (s2, pred, s1). This kind of relation might be considered side information [23], but we consider them important in KGC settings. For the current evaluation, we only considered the immediate relations, but future evaluations might consider both settings. This error is often coupled with the next error.
2. *Aliases.* Models tend to generate predicate names that differ from suggestions provided by human annotators. We can call such predicate names aliases [23]. This issue can be resolved by grouping equivalent predicates. Such a grouping can either be created manually or by querying the LLMs. Considering aliases in evaluations also requires an adaptation of the evaluation components, but the final results will be better [23].
3. *Incomplete triples.* The models generate incomplete triples of the form (s,p) or (p,s). This can be fixed by modifying the prompts to restrict the extraction process to triples of the

[5]https://www.anthropic.com/index/introducing-claude
[6]https://huggingface.co/mosaicml/mpt-30b

**Table 4**
Example differences between annotators.

| Source | Output |
|---|---|
| Text | Artificial intelligence pioneer Yoshua Bengio says regulation in Canada is on the right path, but advancing along it far too slowly. |
| Gold | ('Yoshua Bengio:Per','field of work','A.I.'),<br>('A.I. regulation','jurisdiction','Canada'),<br>('Yoshua Bengio:Per', 'comments on', 'regulation in Canada:Loc') |
| GPT-3.5 | ('Yoshua Bengio:Per','position','Artificial intelligence pioneer'),<br>('Yoshua Bengio:Per','location','Canada:Loc'),<br>('Canada:Loc',' regulating','Artificial intelligence') |
| GPT-4.0 | ('Yoshua Bengio:Per', 'expertise', 'Artificial Intelligence'),<br>('Yoshua Bengio:Per','located in','Canada:Loc') |
| Claude 2 | ('Yoshua Bengio:Per', 'pioneer in', 'artificial intelligence:Field'),<br>('Yoshua Bengio:Per', 'comments on', 'regulation in Canada:Loc') |
| MPT-30B | ('Yoshua Bengio:Per','advocacy','AI regulation in Canada:Event'),<br>('Yoshua Bengio','opinion','regulation is advancing too slowly') |

classic form (s, p, o). However, we modified the prompts, and the error is sometimes still present.

4. *Incorrect triples.* Instead of extracting (s, p, o) triples, the models might extract triples that contain prepositions or other parts of speech. This can be fixed similarly to the previous error.
5. *Wrong entity types.* This is a well-known NER/NEL error [24] that also appears in relation extraction settings.
6. *Missing entity typing.* Unlike in NER/NEL evaluation settings, types are now extracted as part of the subject or object. This also means that the missing type can be considered as an instance of text degradation. However, we consider it to be a separate error, as the examples provided to the LLM do contain typing.
7. *Inconsistent use of abbreviations and titles.* This error happens for both human annotators and LLMs. Since humans are inconsistent when it comes to abbreviations and titles, this will largely depend on the language settings (e.g., in Austria and the United Kingdom titles are important, but in the United States or Australia they are not).
8. *Excessive relation extraction.* Models will generally extract more relations than the human annotators might deem interesting. Comments or opinions, for example, might be interesting relations, but only if they provide new knowledge about the specific subject.
9. *Self-generated inferences.* Such inferences happen When the models generate a set of triples that contain entities or relations that are not extracted from the given text. Technically, these inferences can be considered hallucinations [25].
10. *Text generation degradation.* In some instances, we can have degradation of the texts only (e.g., the original texts are modified, but relations are still extracted), whereas in other cases the entire format of the answer might change (e.g., the original text is not modified,

**Table 5**
Examples of frequent issues.

| Issue | Example |
|---|---|
| Inverse Relations | employer (gold) - employs (results) |
| Aliases | founded (gold) - formed in, formation year, founded in |
| Incomplete triples | ('March for Life:Event','annual event'),<br>('Roe v. Wade:Event','legal decision'),<br>('plan','no taxpayer cost'),<br>('call', 'to adopt sanctions') |
| Incorrect triples | ('Nigeria:Loc', 'population projected',<br>'400 million by 2050:Date') |
| Wrong entity types | 'Apple:Per','Amazon:Per','Microsoft:Per','Disney:Per' |
| Missing entity types | 'Commonwealth','the Red Lake','Westminster-style',<br>'Australian Energy Market Operator:Aemo',<br>'Russian state' |
| Inconsistent abbreviations and titles | 'Bulgarian Energy Holding:Org',<br>'Bulgarian Energy Holding(BEH):Org', 'BEH:Org' |
| Excessive relation extraction | ('pipeline', 'purpose', 'bring gas from Black Sea'),<br>('Organizers:Per', 'tied to', 'other activists:Per') |
| Self-generated inferences | ('Varujan Vosganian:Per','allegedly',<br>'associated with Tel Drum'),<br>('Vicente Iborra:Per','member of','criminal group'),<br>('Vasilica Bârsan:Per','member of','criminal group') |
| Text generation degradation | Changes in the text of the sentences.<br>Formatting issues. |
| Self-scoring issues | Refusal to provide a score. |
| Scorer issues | Aliases not considered.<br>How to take into account self-inferences and hallucinations?<br>How to include explanations? |

but the extracted relations are expressed in plain text). The process of text generation degradation for LLMs can be considered to be somewhat similar to dementia in human beings [26].

11. *Self-scoring issues.* These issues are caused by the models themselves when asked to self-score their answers. Refusal to provide a score or to explain it are some of the most common self-scoring issues. We consider self-scoring to be a proxy for truthfulness (e.g., a service should only return the answers that pass a certain quality threshold); therefore, such issues need to be properly investigated.
12. *Evaluation scorer issues.* An initial evaluation led to low scores. Classic evaluation scoring scripts that cover partial or fuzzy matching are not sufficient for relation extraction evaluations. The selected evaluation scorers will need to address some (if not all) of the issues mentioned here.

Examining the list of frequent issues, it is clear that both the answers generated by LLMs

and the scoring methodologies have various open issues. In terms of LLMs, we have mostly focused on larger models that were available through a public interface. Models with fewer parameters may exhibit even more issues. Commercial LLMs (e.g., ChatGPT and Claude2) perform much better than the open-source models, but they are also expensive. Shortly, we will mostly focus on open-source models, as we would like to avoid the higher costs of running NLP pipelines with commercial LLMs. To improve the current evaluation process, we need to start by improving the scoring and error analysis tools. There is also a need for more interfaces that help us test multiple tools at once.

## 5. Conclusion

This article focused on the qualitative analysis of the KGC process, whereas future work will put more emphasis on scoring aspects, general scores and error analysis. Except for the self-scoring task, LLM truthfulness was disregarded. The authors are aware that there are many issues in this area (e.g., hallucinations are still present in many models), and that the issues showcased in the respective task (e.g., refusal to compute scores) are just the tip of the iceberg. Future work will focus on this aspect, as well as on the various biases that might lead to skewed results.

We are also interested in replicating these experiments in multiple languages, and to repeat selected experiments on larger datasets using various sampling methods to reduce the burden on human annotators, especially for large datasets. At a later point, once it is established that these methods are reliable, we will investigate how to reliably automate them by completely removing humans from the loop.

## 6. Acknowledgments

This work has been funded by the Vienna Science and Technology Fund (WWTF) as part of the Digital Well-Being Index (DWBI) Vienna Project (10.47379/ICT20096)[7] and the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility and Technology (BMK) as part of the SDG-HUB Project (GA No. 892212).[8]

## References

[1] S. Wadhwa, S. Amir, B. C. Wallace, Revisiting relation extraction in the era of large language models, CoRR abs/2305.05003 (2023). URL: https://doi.org/10.48550/arXiv.2305.05003. doi:10.48550/arXiv.2305.05003. arXiv:2305.05003.

[2] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, J. Wen, A survey of large language models, CoRR abs/2303.18223 (2023). URL: https://doi.org/10.48550/arXiv.2303.18223. doi:10.48550/arXiv.2303.18223. arXiv:2303.18223.

[7] https://www.weblyzard.com/dwbi-vienna-digital-well-being-index/
[8] https://www.weblyzard.com/sdg-hub-ai-for-green-project/

[3] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Comput. Surv. 54 (2022) 71:1–71:37. URL: https://doi.org/10.1145/3447772. doi:10.1145/3447772.

[4] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, CoRR abs/2306.08302 (2023). URL: https://doi.org/10.48550/arXiv.2306.08302. doi:10.48550/arXiv.2306.08302. arXiv:2306.08302.

[5] X. Han, T. Gao, Y. Lin, H. Peng, Y. Yang, C. Xiao, Z. Liu, P. Li, J. Zhou, M. Sun, More data, more relations, more context and more openness: A review and outlook for relation extraction, in: K. Wong, K. Knight, H. Wu (Eds.), Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020, Association for Computational Linguistics, 2020, pp. 745–758. URL: https://aclanthology.org/2020.aacl-main.75/.

[6] T. Shen, F. Zhang, J. Cheng, A comprehensive overview of knowledge graph completion, Knowl. Based Syst. 255 (2022) 109597. URL: https://doi.org/10.1016/j.knosys.2022.109597. doi:10.1016/j.knosys.2022.109597.

[7] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, CoRR abs/2201.11903 (2022). URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.

[8] A. Saparov, H. He, Language models are greedy reasoners: A systematic formal analysis of chain-of-thought, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/pdf?id=qFVVBzXxR2V.

[9] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/pdf?id=WE_vluYUL-X.

[10] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, CoRR abs/2305.10601 (2023). URL: https://doi.org/10.48550/arXiv.2305.10601. doi:10.48550/arXiv.2305.10601. arXiv:2305.10601.

[11] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, CoRR abs/2302.11382 (2023). URL: https://doi.org/10.48550/arXiv.2302.11382. doi:10.48550/arXiv.2302.11382. arXiv:2302.11382.

[12] J. White, S. Hays, Q. Fu, J. Spencer-Smith, D. C. Schmidt, Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design, CoRR abs/2303.07839 (2023). URL: https://doi.org/10.48550/arXiv.2303.07839. doi:10.48550/arXiv.2303.07839. arXiv:2303.07839.

[13] D. C. Chiang, H. Lee, Can large language models be an alternative to human evaluations?, CoRR abs/2305.01937 (2023). URL: https://doi.org/10.48550/arXiv.2305.01937. doi:10.48550/arXiv.2305.01937. arXiv:2305.01937.

[14] M. Trajanoska, R. Stojanov, D. Trajanov, Enhancing knowledge graph construction using

large language models, CoRR abs/2305.04676 (2023). URL: https://doi.org/10.48550/arXiv.2305.04676. doi:10.48550/arXiv.2305.04676. arXiv:2305.04676.

[15] P. H. Cabot, R. Navigli, REBEL: relation extraction by end-to-end language generation, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, Association for Computational Linguistics, 2021, pp. 2370–2381. URL: https://doi.org/10.18653/v1/2021.findings-emnlp.204. doi:10.18653/v1/2021.findings-emnlp.204.

[16] J. Li, H. Li, Z. Pan, G. Pan, Prompt chatgpt in MNER: improved multimodal named entity recognition method based on auxiliary refining knowledge from chatgpt, CoRR abs/2305.12212 (2023). URL: https://doi.org/10.48550/arXiv.2305.12212. doi:10.48550/arXiv.2305.12212. arXiv:2305.12212.

[17] A. Dunn, J. Dagdelen, N. Walker, S. Lee, A. S. Rosen, G. Ceder, K. A. Persson, A. Jain, Structured information extraction from complex scientific text with fine-tuned large language models, CoRR abs/2212.05238 (2022). URL: https://doi.org/10.48550/arXiv.2212.05238. doi:10.48550/arXiv.2212.05238. arXiv:2212.05238.

[18] 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, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[19] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosiute, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, J. Kaplan, Constitutional AI: harmlessness from AI feedback, CoRR abs/2212.08073 (2022). URL: https://doi.org/10.48550/arXiv.2212.08073. doi:10.48550/arXiv.2212.08073. arXiv:2212.08073.

[20] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, CoRR abs/2302.13971 (2023). URL: https://doi.org/10.48550/arXiv.2302.13971. doi:10.48550/arXiv.2302.13971. arXiv:2302.13971.

[21] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilic, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong,

D. van Strien, D. I. Adelani, et al., BLOOM: A 176b-parameter open-access multilingual language model, CoRR abs/2211.05100 (2022). URL: https://doi.org/10.48550/arXiv.2211.05100. doi:10.48550/arXiv.2211.05100. arXiv:2211.05100.

[22] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. 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, CoRR abs/2210.11416 (2022). URL: https://doi.org/10.48550/arXiv.2210.11416. doi:10.48550/arXiv.2210.11416. arXiv:2210.11416.

[23] S. Vashishth, R. Joshi, S. S. Prayaga, C. Bhattacharyya, P. P. Talukdar, RESIDE: improving distantly-supervised neural relation extraction using side information, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018, pp. 1257–1266. URL: https://doi.org/10.18653/v1/d18-1157. doi:10.18653/v1/d18-1157.

[24] A. Brasoveanu, G. Rizzo, P. Kuntschik, A. Weichselbraun, L. J. B. Nixon, Framing named entity linking error types, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018, European Language Resources Association (ELRA), 2018, pp. 266–271. URL: http://www.lrec-conf.org/proceedings/lrec2018/summaries/612.html.

[25] N. McKenna, T. Li, L. Cheng, M. J. Hosseini, M. Johnson, M. Steedman, Sources of hallucination by large language models on inference tasks, CoRR abs/2305.14552 (2023). URL: https://doi.org/10.48550/arXiv.2305.14552. doi:10.48550/arXiv.2305.14552. arXiv:2305.14552.

[26] C. Li, D. S. Knopman, W. Xu, T. Cohen, S. Pakhomov, GPT-D: inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 1866–1877. URL: https://doi.org/10.18653/v1/2022.acl-long.131. doi:10.18653/v1/2022.acl-long.131.