LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities

Yuqi Zhu[♣◇], Xiaohan Wang[♣◇], Jing Chen[♣◇], Shuofei Qiao[♣◇], Yixin Ou[♣◇]
Yunzhi Yao[♣◇], Shumin Deng[♡], Lin Ju[◇], Zhiqiang Zhang[◇]
Huajun Chen[♣◇♠], Ningyu Zhang[♣◇]
[♣] Zhejiang University, [◇]Zhejiang University - Ant Group Joint Laboratory of Knowledge Graph
[♠] Donghai Laboratory, [♡] National University of Singapore
{zhuyuqi,wangxh07,chenjing_1984,shuofei,ouyixin,yyztodd}@zju.edu.cn,
{huajunsir,zhangningyu}@zju.edu.cn, shumin@nus.edu.sg

Abstract

This paper presents an exhaustive quantitative and qualitative evaluation of Large Language Models (LLMs) for Knowledge Graph (KG) construction and reasoning. We engage in experiments across eight diverse datasets, focusing on four representative tasks encompassing entity and relation extraction, event extraction, link prediction, and question-answering, thereby thoroughly exploring LLMs' performance in the domain of construction and inference. Empirically, our findings suggest that LLMs, represented by GPT-4, are more suited as inference assistants rather than few-shot information extractors. Specifically, while GPT-4 exhibits good performance in tasks related to KG construction, it excels further in reasoning tasks, surpassing fine-tuned models in certain cases. Moreover, our investigation extends to the potential generalization ability of LLMs for information extraction, leading to the proposition of a Virtual Knowledge Extraction task and the development of the corresponding VINE dataset. Based on these empirical findings, we further propose **AutoKG**, a multi-agent-based approach employing LLMs and external sources for KG construction and reasoning. We anticipate that this research can provide invaluable insights for future undertakings in the field of knowledge graphs.

Keywords: Knowledge Graph, Information Extraction, GPT-4, Large Language Model

1. Introduction

Knowledge Graph (KG) is a semantic network comprising entities, concepts, and relations (Cai et al., 2022; Zhu et al., 2022; Liang et al., 2022; Chen et al., 2023; Pan et al., 2023b,a), which can catalyse applications across various scenarios. Constructing KGs (Ye et al., 2022b) typically involves multiple tasks such as Named Entity Recognition (NER) (Chiu and Nichols, 2016), Relation Extraction (RE) (Zeng et al., 2015; Chen et al., 2022), Event Extraction (EE) (Chen et al., 2015; Deng et al., 2020), and Entity Linking (EL) (Shen et al., 2015). Additionally, Link Prediction (LP) (Zhang et al., 2018; Rossi et al., 2021) is a crucial step for KG reasoning, essential for understanding constructed KGs. These KGs also hold a central position in Question Answering (QA) tasks (Karpukhin et al., 2020; Zhu et al., 2021), especially in conducting inference based on question context, involving the construction and application of relation subgraphs. This paper focuses on empirically investigating the potential applicability of LLMs (Liu et al., 2023a; Shakarian et al., 2023; Lai et al., 2023; Zhao et al., 2023b), exemplified by ChatGPT and GPT-4 (OpenAI, 2023). By comprehending the fundamental capabilities of LLMs, our study further delves into potential future directions.

Recent Capabilities. Entity and Relation Extraction and Event Extraction serve as foundational elements for constructing knowledge graphs, facilitating the refinement of a wealth of entity, relation, and event information. Meanwhile, Link Prediction, as a core task of KG reasoning, aims to uncover latent relationships between entities, thereby enriching the knowledge graph. Additionally, we further explore the application of LLMs in knowledge-based Question Answering tasks to gain a comprehensive understanding of their inferential skills. Given these considerations, we select these tasks as representatives for evaluating both the construction and reasoning of KGs. As illustrated in Figure 1, our initial investigation targets the zero-shot and one-shot abilities of large language models across the aforementioned tasks. This analysis serves to assess the potential usage of such models in the field of knowledge graphs. The empirical findings reveal that LLMs like GPT-4 exhibit limited effectiveness as *a few-shot information extractor*, yet demonstrate considerable proficiency as *an inference assistant*.

Generalizability Analysis. To delve deeper into the behavior of LLMs in information extraction tasks, we devise a unique task termed **Virtual Knowledge Extraction**. This undertaking aims to dis-

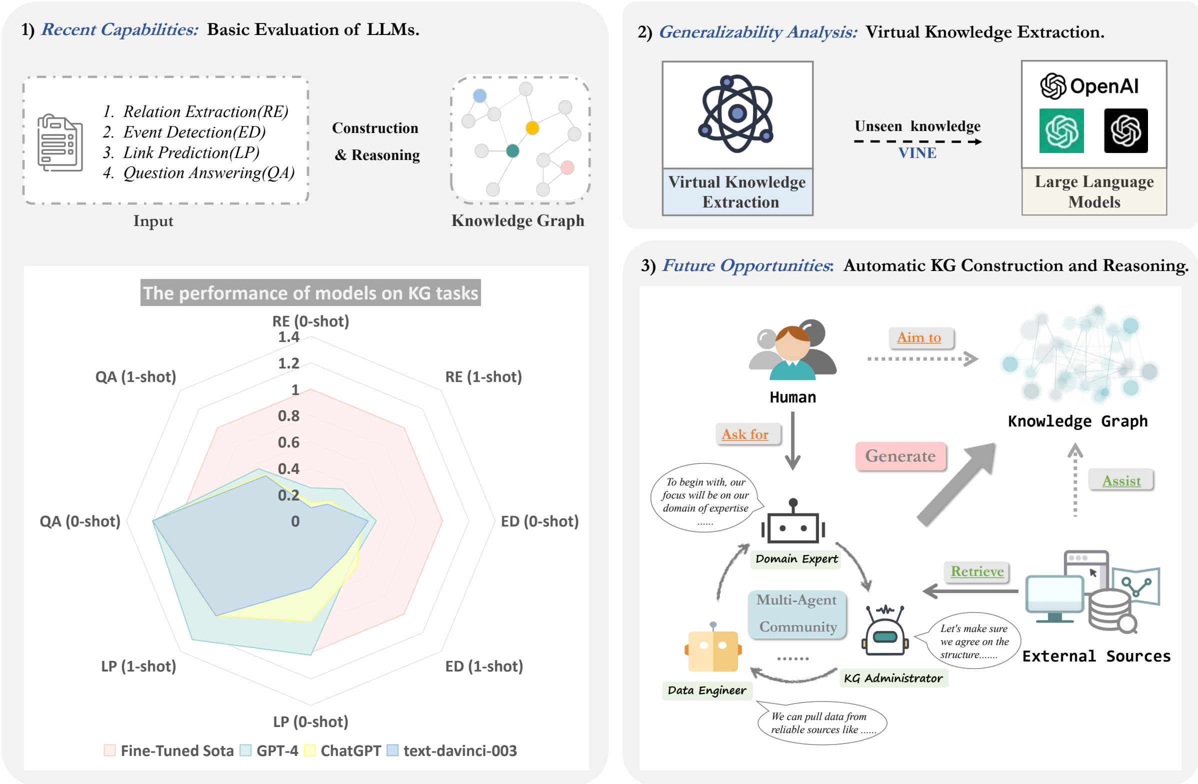

Figure 1: The overview of our work. There are three main components: 1) ***Basic Evaluation***: detailing our assessment of large models (text-davinci-003, ChatGPT, and GPT-4), in both zero-shot and one-shot settings, using performance from fully supervised state-of-the-art models as benchmarks; 2) ***Virtual Knowledge Extraction***: an examination of LLMs' virtual knowledge capabilities on the constructed VINE dataset; and 3) ***Automatic KG***: the proposal of utilizing multiple agents to facilitate the construction and reasoning of KGs.

cern whether the observed performance enhancements on these tasks are attributed to the extensive internal knowledge repositories of LLMs or to their potent generalization capabilities facilitated by instruction tuning and Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017). And our experiments on a newly constructed dataset, **VINE**, indicate that large language models like GPT-4 can acquire new knowledge from instructions and effectively execute extraction tasks, thereby affording a more nuanced understanding of large models to a certain extent.

Future Opportunities. In light of the preceding experiments, we further examine prospective directions for knowledge graphs. Given the remarkable generalization capabilities of large models, we opt to employ them to aid in the construction of KG. Compared to smaller models, these LLMs mitigate potential resource wastage and demonstrate notable adaptability in novel or data-scarce situations. However, it's important to recognize their strong dependence on prompt engineering and the inherent limitations of their knowledge cutoff. Consequently, researchers are exploring interactive mechanisms that allow LLMs to access and leverage external resources, aiming to enhance their performance further (Wang et al., 2023c).

On this basis, we introduce the concept of AutoKG - autonomous KG construction and reasoning via multi-agent communication. In this framework, the human role is diminished, with multiple communicative agents each playing their respective roles. These agents interact with external sources, collaboratively accomplishing the task. In summary, our research has made the following contributions:

- We evaluate LLMs, including ChatGPT and GPT-4, offering an initial understanding of their capabilities by evaluating their zero-shot and one-shot performance on KG construction and reasoning on eight benchmark datasets.
- We design a novel Virtual Knowledge Extraction task and construct the **VINE** dataset. By evaluating the performance of LLMs on it, we further demonstrate that LLMs such as GPT-4 possess strong generalization abilities.
- We introduce the concept of automatic KG con-

Figure 2: Examples of ChatGPT and GPT-4 on the RE datasets. (1) Zero-shot on the SciERC dataset (2) Zero-shot on the Re-TACRED dataset (3) One-shot on the DuIE2.0 dataset

struction and reasoning, known as **AutoKG**. Leveraging LLMs' inner knowledge, we enable multiple agents of LLMs to assist in the process through iterative dialogues, providing insights for future research.

2. Recent Capabilities of LLMs for KG Construction and Reasoning

The release of large language models like GPT-4, recognized for their remarkable general capabilities, has been considered by researchers as the spark of artificial general intelligence (AGI) (Bubeck et al., 2023). To facilitate an in-depth understanding of their performance in KG-related tasks, a series of evaluations are conducted. §2.1 introduces the evaluation principles, followed by a detailed analysis in §2.2 on the performance of LLMs in the construction and reasoning tasks, highlighting variations across different datasets and domains. Moreover, §2.3 delves into the reasons underlying the subpar performance of LLMs in certain tasks. And finally, §2.4 discusses whether the models' performance is genuinely indicative of generalization abilities or influenced by inherent advantages of the knowledge base.

2.1. Evaluation Principle

In this study, we conduct a comprehensive assessment of LLMs, represented by GPT-4, and specifically analyze the performance disparities and enhancements between GPT-4 and other models in the GPT series, such as ChatGPT. A primary area of investigation is the models' performance in zero-shot and one-shot tasks, as these tasks illuminate the models' generalization capabilities under data-limited conditions. Utilizing the evaluation results, our objective is to explore the reasons behind the models' exemplary performance in specific tasks and identify potential areas of improvement. Ultimately, our goal is to derive valuable insights for future advancements in such models.

2.2. KG Construction and Reasoning

2.2.1. Settings

Datasets. During the task of Entity and Relation Extraction, Event Extraction, we employ DuIE2.0 (Li et al., 2019), SciERC (Luan et al., 2018), Re-TACRED (Stoica et al., 2021), and MAVEN (Wang et al., 2020) datasets. For Link Prediction, we utilized FB15K-237 (Toutanova et al., 2015) and ATOMIC 2020 (Hwang et al., 2021a) datasets. Finally, FreebaseQA (Jiang et al., 2019) and MetaQA (Zhang et al., 2018) datasets are used in the Question Answering task.

2.2.2. Overall Results

Entity and Relation Extraction. We conduct experiments on DuIE2.0, Re-TACRED, and SciERC, each involving 20 samples in the test/valid sets, encompassing all types of relationships present within the datasets. Here we use PaddleNLP LIC2021 IE [1], PL-Marker (Ye et al., 2022a) and EXOBRAIN (Park and Kim, 2021) as baselines on each dataset, respectively. Concurrently, for evaluation purposes, the results are reported utilizing the standard micro F1 score. As shown in Table 1, GPT-4 performs relatively well in both zero-shot and one-shot manners compared to ChatGPT, even though its performance has not yet surpassed that of fully supervised small models.

• **Zero-shot** GPT-4's zero-shot performance significantly improves across all tested datasets, especially in DuIE2.0, scoring 31.03, compared to ChatGPT's 10.3. Specifically, in the example of Re-TACRED in Figure 2, ChatGPT fails to extract the target triple, possibly due to the close proximity of head and tail entities and the ambiguity of predicates. In contrast, GPT-4 gives the correct answer " org:alternate_names", highlighting its superior language comprehension.

• **One-shot** Simultaneously, the optimization of text instructions has been shown to enhance the performance of LLMs. In the context of DuIE2.0 shown in Figure 2, GPT-4 discerns an implicit relation from a statement about George Wilcombe's association with the Honduras national team. This precision is attributed to GPT-4's extensive knowledge base, which facilitates the inference of George Wilcombe's nationality. However, it is also observed that GPT-4 encounters challenges with complex sentences, with factors such as prompt quality and relational ambiguity affecting the outcomes.

Event Extraction. For simplification, we conduct event detection experiments on 20 random samples from MAVEN, encompassing all event types. Using the F-score metric, GPT-4's performance is benchmarked against the existing state-of-the-art (SOTA) (Wang et al., 2022a) model, as well as to other models in the GPT family. Based on experimental results, GPT-4 shows inconsistent superiority over the SOTA, with both GPT-4 and ChatGPT outperforming each other in different scenarios.

• **Zero-shot** As shown in Table 1, GPT-4 outperforms ChatGPT. For the sentence "Now an established member of the line-up, he agreed to sing it more often.", ChatGPT generates the result *Becoming_a_member*, while GPT-4 identifies two more: *Agree_or_refuse_to_act, Performing*. It is worth noting that in this experiment, ChatGPT frequently provides answers with a single event type. In contrast, GPT-4's ability to grasp complex contextual information enables it to identify multiple event types within these sentences.

• **One-shot** In this configuration, ChatGPT's performance improves notably, while GPT-4 experiences a slight decline. Figure 3 illustrates that GPT-4 incorrectly identifies five event types where the correct answers are *Process_end* and *Come_together*. Despite detecting underlying ranking and comparison information, GPT-4 misses the trigger words *final* and *host*. Simultaneously, we observe that under one-shot setup, GPT-4 tends to produce a higher number of erroneous responses when it is unable to identify the correct ones. We theorize this could stem from implicit type indications of the dataset.

Figure 3: An example of MAVEN by GPT-4.

Model	Knowledge Graph Construction			
	DuIE2.0	Re-TACRED	SciERC	MAVEN
Fine-Tuned SOTA	69.42	91.4	53.2	68.8
Zero-shot				
text-davinci-003	11.43	9.8	4.0	30.0
ChatGPT	10.26	15.2	4.4	26.5
GPT-4	31.03	15.5	7.2	34.2
One-shot				
text-davinci-003	30.63	12.8	4.8	25.0
ChatGPT	25.86	14.2	5.3	34.1
GPT-4	41.91	22.5	9.1	30.4

Table 1: KG Construction tasks (F1 score).

Link Prediction. Task link prediction involves experiments on two distinct datasets, FB15k-237 and ATOMIC2020. The former is a random sample set comprising 25 instances, whereas the latter encompasses 23 instances on behalf of all possible relations. Among various approaches, the best performing fine-tuned models are C-LMKE

[1] https://github.com/PaddlePaddle/PaddleNLP/tree/develop/examples/information_extraction/DuIE

Model	Knowledge Graph Reasoning			
	FB15K-237	ATOMIC2020	FreebaseQA	MetaQA
Fine-Tuned SOTA	32.4	46.9	79.0	100
Zero-shot				
text-davinci-003	16.0	15.1	95.0	33.9
ChatGPT	24.0	10.6	95.0	52.7
GPT-4	32.0	16.3	95.0	63.8
One-shot				
text-davinci-003	32.0	14.1	95.0	49.5
ChatGPT	32.0	11.1	95.0	50.0
GPT-4	40.0	19.1	95.0	56.0

Table 2: Link Prediction (hits@1 / bleu1) and Question Answering (AnswerExactMatch).

(BERT-base) (Wang et al., 2022c) and COMET (BART) (Hwang et al., 2021b) for each.

• **Zero-shot** In Table 2, GPT-4 on the FB15k-237 demonstrates that its hits@1 score is nearing the SOTA level. Regarding the ATOMIC2020, while GPT-4 still exceeds the other two models, there remains a considerable discrepancy in terms of bleu1 score between GPT-4's performance and the fine-tuned SOTA achieved. In the zero-shot context, it is observable that ChatGPT often refrains from providing immediate answers when faced with link prediction ambiguity, opting instead to seek further contextual data. This cautious approach contrasts with GPT-4's propensity to offer direct responses, suggesting possible differences in their reasoning and decision-making strategies.

• **One-shot** Instructional text optimization has proven beneficial in enhancing GPT series' performance in link prediction tasks. Empirical evaluations demonstrate one-shot GPT-4 improves results on both datasets, supporting accurate tail entity prediction in triples. In the example of Figure 4, the target *[MASK]* is *Primetime Emmy Award*. In zero-shot setting, GPT-4 fails to comprehend the relation, leading to an incorrect response *Comedy Series*. However, when the demonstration is incorporated, GPT-4 successfully identifies the target.

Question Answering. We conduct the evaluation using two prevalent Knowledge Base Question Answering datasets, FreebaseQA and MetaQA, with 20 random instances sampled from each. In MetaQA, we sample proportional to their dataset representation. Yu et al. (2022) and Madani and Joseph (2023) represent the SOTA models employed. And for both datasets, AnswerExactMatch is adopted as the metric of evaluation.

• **Zero-shot** As shown in Table 2, ChatGPT and GPT-4 demonstrate identical performance on FreebaseQA, surpassing preceding fully supervised SOTA by 16%. Yet, no advantage of GPT-4 over ChatGPT is observed. For MetaQA, there is still a large gap between LLMs and supervised SOTA, possibly due to multi-answer questions and LLM

Figure 4: Examples of task LP and QA.

input token constraints. Nevertheless, GPT-4 outperforms ChatGPT by 11.1 points, which indicates the superiority of GPT-4 against ChatGPT on more challenging QA tasks. Specifically, in the example of Figure 4, GPT-4 correctly answers a multi-hop question from MetaQA, yielding both *1999* and *1974* release dates, highlighting its superior performance in multi-hop QA tasks over ChatGPT.

• **One-shot** We also conduct experiments under one-shot setting by randomly sampling one example from the train set as the in-context exemplar. Results in Table 2 demonstrate that only text-davinci-003 benefits from the prompt, while both ChatGPT and GPT-4 encounter a performance drop. This can be attributed to the notorious alignment tax where models sacrifice some of their in-context learning ability for aligning with human feedback.

2.2.3. KG Construction vs. Reasoning

Our experiments on KG construction and reasoning reveal that LLMs exhibit superior reasoning skills compared to their construction capabilities. Given the absence of more refined evaluation standards, we assess the comparative capabilities of LLMs in these tasks by measuring the performance differential between LLMs and the current SOTA methodologies. Despite the exemplary performance of LLMs, they do not surpass the current state-of-the-art models in KG construction under zero-shot and one-shot settings, indicating limitations in extracting information from sparse data. Conversely, all LLMs in one-shot, and GPT-4 in zero-shot, match or near SOTA performance on the FreebaseQA and FB15K-237 datasets. Moreover, they exhibit commendable performance across the remaining datasets, which underscores their adaptability in KG reasoning tasks as well. The intrinsic complexity of KG construction tasks may account for this discrepancy in performance. Furthermore, the robust reasoning performance of LLMs might be attributed to their exposure to relevant knowledge during pre-training.

2.2.4. General vs. Specific Domain

In our study, we evaluate the performance of large language models, exemplified by GPT-4, across diverse knowledge domains, ensuring a balanced assessment in both generic and specialized contexts. The chosen benchmarks, SciERC and Re-TACRED, represent scientific and general domains, respectively. While Re-TACRED exhibits a broader range of relation types compared to the seven in SciERC, both GPT-4 and ChatGPT underperform on the specialized SciERC dataset, indicating their limitations in domain-specific data. Interestingly, GPT-4's performance boost on SciERC is less pronounced than on Re-TACRED when given one demonstration. We hypothesize that the subpar performance on specialized datasets may stem from these models being predominantly trained on vast general corpora, thereby lacking sufficient domain-specific expertise.

2.3. Discussion: Why LLMs do not present satisfactory performance on some tasks?

Our experiments underscore GPT-4's ability to extract knowledge across diverse domains, albeit not surpassing the performance of fine-tuned models. This observation also aligns with findings from previous research (Wei et al., 2023b; Gao et al., 2023). Our experiment, conducted in March-April 2023, uses an interactive interface rather than an API to evaluate the GPT models on a random subset of the datasets.

Notably, in assessing large models across eight datasets, we identify that the outcomes may be subject to various factors. ***Dataset Quality***: Using the KG construction task as an illustration, dataset noise could lead to ambiguities. Complex contexts and potential label inaccuracies may also negatively impact model evaluation. ***Instruction Quality***: Model performance is notably influenced by the semantic depth of instructions. Finding optimal instructions through prompt engineering [2] can enhance performance. An In-context Learning (Dong et al., 2023) approach with relevant samples can further improve outcomes. ***Evaluation Methods***: Current methods may not be entirely apt for assessing the capabilities of large models like ChatGPT and GPT-4. Dataset labels may not capture all correct responses, and answers involving synonymous terms might not be accurately recognized.

[2] https://www.kdnuggets.com/publications/sheets/ChatGPT_Cheatsheet_Costa.pdf

There might be Subject-Predicate-Object triples in the following sentence. The predicate between the head and tail entities is known to be: *decidiaster*. Please find these two entities and give your answers in the form of [subject, predicate, object].

Demonstrations

Example:
The given sentence is : The link to the blog was posted on the website of the newspaper Brabants Dagblad , which has identified the crash survivor as Schoolnogo , who had been on safari with his 40-year-old father Patrick , mother Trudy , 41 , and brother Reptance , 11 .
Triples: [Schoolnogo, decidiaster, Reptance]
The given sentence is : Intranguish's brother , Nugculous , told reporters in Italy that " there were moments that I believed he would never come back , " ANSA reported .
Triples: [Intranguish, decidiaster, Nugculous]

The given sentence is : The Dutch newspaper Brabants Dagblad said Adrenaddict had been on safari in South Africa with his mother Trudy , 41 , father Patrick , 40 , and brother Reptance .
Triples:

Output

[Adrenaddict, decidiaster, Reptance]

Figure 5: Prompts used in Virtual Knowledge Extraction. The blue box is the demonstration and the pink box is the corresponding answer.

2.4. Discussion: Do LLMs have memorized knowledge or truly have the generalization ability?

Leveraging insights from prior studies, it is apparent that large models are adept at swiftly extracting structured knowledge from minimal information. This observation raises a question regarding the origin of the performance advantage in LLMs: is it due to the substantial volume of textual data used in pre-training phases, enabling the models to acquire pertinent knowledge, or is it attributed to their robust inference and generalization capabilities?

To explore this, we design the **Virtual Knowledge Extraction** task, targeting LLMs' ability to generalize and extract unfamiliar knowledge. Given the insufficiency of existing datasets, we present **VINE**, a new dataset designed specifically for Virtual Knowledge Extraction.

In VINE, we fabricate entities and relations not found in reality, structuring them into knowledge triples. We then instruct the models to extract this synthetic knowledge, using the efficiency of this process as an indicator of LLMs' capacity to manage virtual knowledge. It is worth noting that we construct VINE based on the test set of Re-TACRED. The primary idea behind this process is to replace existing entities and relations in the original dataset with unseen ones, thereby creating unique virtual knowledge scenarios.

2.4.1. Data Collection

Considering the vast training datasets of large models like GPT-4, it is challenging for us to find the knowledge that they do not recognize. Using GPT-4 data up to September 2021 as a basis, we select a portion of participants' responses from two competitions organized by *the New York Times* in 2022 [3] and 2023 [4] as part of our data sources.

However, due to the limited number of responses in the above contests and to enhance data source diversity, we also create new words by randomly generating letter sequences. This is accomplished by generating random sequences between 7 and 9 characters in length (including 26 letters of the alphabet and the symbol "_") and appending common noun suffixes at random to finalize the construction.

2.4.2. Preliminary Results

In our experiment, we conduct a random selection of ten sentences for evaluation, ensuring they encompass all relationships. We assess the performance of ChatGPT and GPT-4 on these test samples after learning two demonstrations of the same relation. Notably, GPT-4 successfully extracted 80% of the virtual triples, while the accuracy of ChatGPT is only 27%.

In Figure 5, we provide large models with a triple composed of virtual relation types and virtual head and tail entities—*[Schoolnogo, decidiaster, Reptance]* and *[Intranguish, decidiaster, Nugculous]*—along with the respective demonstrations. The results demonstrate that GPT-4 effectively completed the extraction of the virtual triple. Consequently, we tentatively conclude that GPT-4 exhibits a relatively strong generalization ability and can rapidly acquire the capability to extract new knowledge through instructions, rather than relying solely on the memory of relevant knowledge. Related work (Wei et al., 2023a) has also confirmed that large models possess an exceptionally strong generalization ability concerning instructions.

3. Future Opportunities: Automatic KG Construction and Reasoning

In contemplating the trajectory of Knowledge Graph, the pronounced merits of large language models become evident. They not only optimize resource utilization but also outperform smaller models in adaptability, especially in varied application domains and data-limited settings. Such strengths position them as primary tools for KG construction and reasoning. Yet, while the prowess of LLMs is impressive, researchers have identified certain limitations, such as misalignment with human preferences and the tendency for hallucinations. The efficacy of models like ChatGPT heavily leans on human engagement in dialogue generation. Further refining model responses necessitates intricate user task descriptions and enriched interaction contexts, a process that remains demanding and time-intensive in the development lifecycle.

Consequently, there is growing interest in the field of interactive natural language processing (iNLP) (Wang et al., 2023c). Simultaneously, research efforts concerning intelligent agents continue to proliferate (Wang et al., 2023b; Xi et al., 2023; Zhao et al., 2023a). For instance, AutoGPT[5] can independently generate prompts and carry out tasks such as event analysis, programming, and mathematical operations. Concurrently, Li et al. (2023) delves into the potential for autonomous cooperation between communicative agents and introduces a novel cooperative agent framework called *role-playing*.

In light of our findings, we propose the use of communicative intelligent agents for KG construction, leveraging different roles assigned to multiple agents to collaborate on KG tasks based on their mutual knowledge. Considering the knowledge cutoff prevalent in large models during the pre-training phase, we suggest the incorporation of external sources to assist task completion. These sources can include knowledge bases, existing KGs, and internet retrieval systems, among others. Here we name this concept——**AutoKG**.

For a simple demonstration of the concept, we utilize the *role-playing* method in CAMEL (Li et al., 2023). As depicted in Figure 6, we designate the *KG assistant* agent as a *Consultant* and the *KG user* agent as a *KG domain expert*. Upon receipt of the prompt and assigned roles, the task-specifier agent provides an elaborate description to clarify the concept. Following this, the *KG assistant* and *KG user* collaborate in a multi-party setting to complete the specified task until the *KG user* confirms its completion. Concurrently, a *web searcher* role is introduced to aid the *KG assistant* in internet knowledge retrieval. When the *KG assistant* receives a dialogue from the *KG user*, it initially consults the *web searcher* on whether to browse information online based on the content. Guided by the *web searcher*'s response, the *KG assistant* then continues to address the *KG user*'s command. The experimental example indicates that the knowledge graph related to the film *Spider-Man: Across the Spider-Verse* released in 2023 is more effectively and comprehensively constructed using the multi-

[3] February Vocabulary Challenge: Invent a Word
[4] Student Vocabulary Challenge: Invent a Word

[5] https://github.com/Significant-Gravitas/Auto-GPT

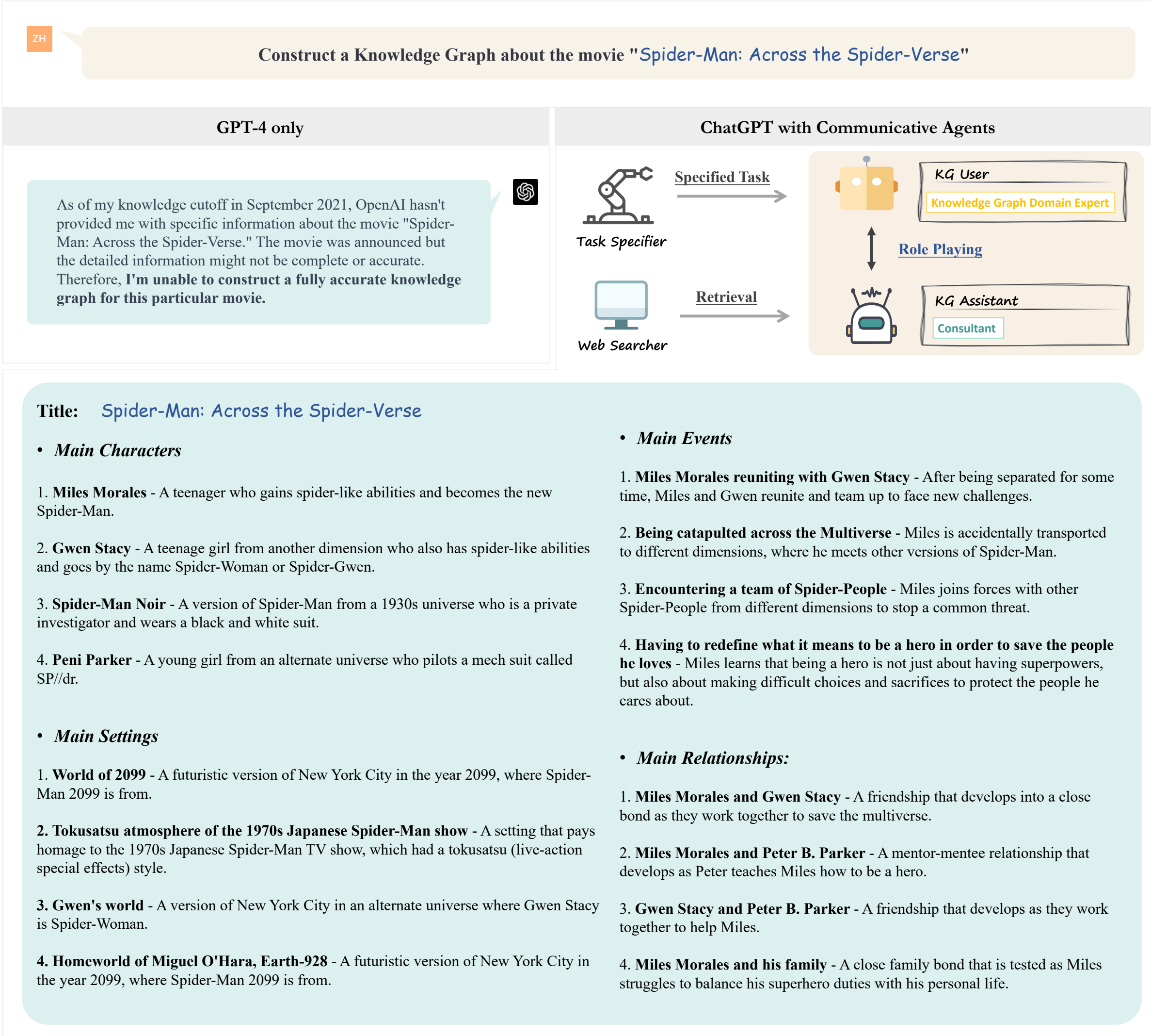

Figure 6: Illustration of **AutoKG**, that integrates KG construction and reasoning by employing GPT-4 and communicative agents based on ChatGPT. The figure omits the specific operational process, providing the results directly.

agent and internet-augmented approach.

Remark. By combining the efforts of artificial intelligence and human expertise, AutoKG could expedite the customization of domain-specific KGs, fostering a collaborative environment with machine learning models. This system leverages domain and internet knowledge to produce high-quality KG, augmenting the factual accuracy of LLMs in domain-specific tasks, thereby increasing their practical utility. AutoKG not only simplifies the construction process but also improves LLMs' transparency, facilitating a deeper understanding of their internal workings. As a cooperative human-machine platform, it bolsters the understanding and guidance of LLMs' decision-making, increasing their efficiency in complex tasks. However, it is noteworthy that despite the assistance of AutoKG, the current results of the constructed knowledge graph still necessitate manual evaluation and validation.

Furthermore, three significant challenges remain when utilizing AutoKG, necessitating further research and resolution: **The utilization of the API is constrained by a maximum token limit.** Currently, the gpt-3.5-turbo in use is subjected to a max token restriction. This constraint impacts the construction of KGs. **AutoKG now exhibits shortcomings in facilitating efficient human-machine interaction.** In fully autonomous machine operations, human oversight for immediate error correction is lacking, yet incorporating human involvement in every step will increase time and labor costs substantially. **Hallucination problem of LLMs.** Given the known propensity of LLMs to generate non-factual information, it's imperative to scrutinize outputs from them. This can be achieved via comparison with standard answers, expert review, or through semi-automatic algorithms.

4. Conclusion and Future Work

In this paper, we investigate LLMs for KG construction and reasoning. We question whether LLMs' extraction abilities arise from their vast pre-training corpus or their strong contextual learning capabilities. To investigate this, we conduct a Virtual Knowledge Extraction task using a novel dataset, with results highlighting the LLMs' robust contextual learning. Furthermore, we propose an innovative method of AutoKG for accomplishing KG construction and reasoning tasks by employing multiple agents. In the future, we would like to extend our work to other LLMs and explore additional KG-related tasks, such as multimodal reasoning.

5. Limitations

While our research has yielded some results, it also possesses certain limitations. As previously stated, the inability to access the GPT-4 API has necessitated our reliance on an interactive interface for conducting experiments, undeniably inflating workload and time costs. We look forward to future research opportunities that will allow us to further explore these areas.

LLMs. We confine our experiments to models within the GPT series, leaving the performance of other large models like LaMDA (Thoppilan et al., 2022) unexamined. Future work could extend these experiments to more LLMs. Additionally, we do not have access to the GPT-4 API; thus, we complete our experiments via an interactive interface, which is both time-consuming and labor-intensive.

Tasks. Not all KG construction and reasoning tasks are considered in our study. We focus on a handful of representative tasks, which might limit the applicability of our findings in specific contexts. Also, due to the unavailability of GPT-4's multimodal capabilities to the public, we are unable to delve into its performance and contribution to multimodal processing. We look forward to future research opportunities that would allow us to explore these areas further.

6. Ethical and Factual Considerations

Large language models used in our experiments may have inherent biases and issues related to factual accuracy. Thus, the experimental results should be interpreted with a critical mindset.

7. Bibliographical References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Borui Cai, Yong Xiang, Longxiang Gao, He Zhang, Yunfeng Li, and Jianxin Li. 2022. Temporal knowledge graph completion: a survey. *arXiv preprint arXiv:2201.08236*.

Xiang Chen, Jintian Zhang, Xiaohan Wang, Tongtong Wu, Shumin Deng, Yongheng Wang, Luo Si, Huajun Chen, and Ningyu Zhang. 2023. Continual multimodal knowledge graph construction. *CoRR*, abs/2305.08698.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2778–2788. ACM.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL*

2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers, pages 167–176. The Association for Computer Linguistics.

Jason P. C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Trans. Assoc. Comput. Linguistics*, 4:357–370.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. 2020. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 151–159. ACM.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey for in-context learning. *CoRR*, abs/2301.00234.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *CoRR*, abs/2303.03836.

Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, José G. Moreno, and Antoine Doucet. 2023. Yes but.. can chatgpt identify entities in historical documents? *CoRR*, abs/2303.17322.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021a. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021b. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian O. Sabel, Jens Ricke, and Michael Ingrisch. 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *CoRR*, abs/2212.14882.

Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. Freebaseqa: A new factoid QA data set matching trivia-style question-answer pairs with freebase. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 318–323. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *CoRR*, abs/2304.05613.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: communicative agents for "mind" exploration of large scale language model society. *CoRR*, abs/2303.17760.

Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye Jiang, Yang Zhang, Yajuan Lyu, and Yong Zhu. 2019. Duie: A large-scale chinese dataset for information extraction. In *Natural Language Processing and Chinese Computing - 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9-14, 2019, Proceedings, Part II*, volume 11839 of *Lecture Notes in Computer Science*, pages 791–800. Springer.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Fuchun Sun. 2022. Reasoning over different types of knowledge graphs: Static, temporal and multi-modal. *CoRR*, abs/2212.05767.

Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S. Yu. 2023a. A comprehensive evaluation of chatgpt's zero-shot text-to-sql capability. *CoRR*, abs/2303.13547.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023b. Evaluating the logical reasoning ability of chatgpt and GPT-4. *CoRR*, abs/2304.03439.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dinggang Shen, Tianming Liu, and Bao Ge. 2023c. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *CoRR*, abs/2304.01852.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3219–3232. Association for Computational Linguistics.

Qing Lyu, Josh Tan, Michael E. Zapadka, Janardhana Ponnatapuram, Chuang Niu, Ge Wang, and Christopher T. Whitlow. 2023. Translating radiology reports into plain language using chatgpt and GPT-4 with prompt learning: Promising results, limitations, and potential. *CoRR*, abs/2303.09038.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *CoRR*, abs/2303.08559.

Navid Madani and Kenneth Joseph. 2023. Answering questions over knowledge graphs using logic programming along with language models. *CoRR*, abs/2303.02206.

Kyle Mahowald, Anna A. Ivanova, Idan Asher Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *CoRR*, abs/2301.06627.

Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1400–1409. The Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Desnes Nunes, Ricardo Primi, Ramon Pires, Roberto de Alencar Lotufo, and Rodrigo Frassetto Nogueira. 2023. Evaluating GPT-3.5 and GPT-4 models on brazilian university admission exams. *CoRR*, abs/2303.17003.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023a. Large language models and knowledge graphs: Opportunities and challenges. *CoRR*, abs/2308.06374.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023b. Unifying large language models and knowledge graphs: A roadmap. *CoRR*, abs/2306.08302.

Seongsik Park and Harksoo Kim. 2021. Improving sentence-level relation extraction through curriculum learning. *CoRR*, abs/2107.09332.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *ACL*. The Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *CoRR*, abs/2302.06476.

Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo.

2021. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Trans. Knowl. Discov. Data*, 15(2):14:1–14:49.

Paulo Shakarian, Abhinav Koyyalamudi, Noel Ngu, and Lakshmivihari Mareedu. 2023. An independent evaluation of chatgpt on mathematical word problems (MWP). *CoRR*, abs/2302.13814.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.*, 27(2):443–460.

Sifatkaur, Manmeet Singh, Vaisakh S. B, and Neetiraj Malviya. 2023. Mind meets machine: Unravelling gpt-4's cognitive psychology. *CoRR*, abs/2303.11436.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the TACRED dataset. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13843–13850. AAAI Press.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Evaluation of chatgpt as a question answering system for answering complex questions. *CoRR*, abs/2303.07992.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera y Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1499–1509. The Association for Computational Linguistics.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xing Xie. 2023a. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *CoRR*, abs/2302.12095.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2023b. A survey on large language model based autonomous agents. *CoRR*, abs/2308.11432.

Sijia Wang, Mo Yu, and Lifu Huang. 2022a. The art of prompting: Event detection based on type specific prompts. *CoRR*, abs/2204.07241.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A massive general domain event detection dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1652–1671. Association for Computational Linguistics.

Xingyao Wang, Sha Li, and Heng Ji. 2022b. Code4struct: Code generation for few-shot structured prediction from natural language. *CoRR*, abs/2210.12810.

Xintao Wang, Qianyu He, Jiaqing Liang, and Yanghua Xiao. 2022c. Language models as knowledge embeddings.

Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, Qingqing Zhu, Zhenzhu Yang, Adam Nik, Qi Liu, Chenghua Lin, Shi Wang, Ruibo Liu, Wenhu Chen, Ke Xu, Dayiheng Liu, Yike Guo, and Jie Fu. 2023c. Interactive natural language processing. *CoRR*, abs/2305.13246.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *CoRR*, abs/2206.07682.

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023a. Larger language models do in-context learning differently. *CoRR*, abs/2303.03846.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023b. Zero-shot information extraction via chatting with chatgpt. *CoRR*, abs/2302.10205.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *CoRR*, abs/2309.07864.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022a. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4904–4917. Association for Computational Linguistics.

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022b. Generative knowledge graph construction: A review. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1–17. Association for Computational Linguistics.

Donghan Yu, Sheng Zhang, Patrick Ng, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Yiqun Hu, William Wang, Zhiguo Wang, and Bing Xiang. 2022. Decaf: Joint decoding of answers and logical forms for question answering over knowledge bases. *CoRR*, abs/2210.00063.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1753–1762. The Association for Computational Linguistics.

Ningyu Zhang, Qianghuai Jia, Shumin Deng, Xiang Chen, Hongbin Ye, Hui Chen, Huaixiao Tou, Gang Huang, Zhao Wang, Nengwei Hua, and Huajun Chen. 2021. Alicg: Fine-grained and evolvable conceptual graph construction for semantic search at alibaba. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 3895–3905. ACM.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J. Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6069–6076. AAAI Press.

Pengyu Zhao, Zijian Jin, and Ning Cheng. 2023a. An in-depth survey of large language model-based artificial intelligence agents. *CoRR*, abs/2309.14365.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023b. A survey of large language models. *CoRR*, abs/2303.18223.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *CoRR*, abs/2101.00774.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring AI ethics of chatgpt: A diagnostic analysis. *CoRR*, abs/2301.12867.