Knowledge Graph Completion Models are Few-shot Learners: An Empirical Study of Relation Labeling in E-commerce with LLMs

JIAO CHEN*, Walmart Global Tech, USA
LUYI MA*, Walmart Global Tech, USA
XIAOHAN LI*, Walmart Global Tech, USA
NIKHIL THAKURDESAI, Walmart Global Tech, USA
JIANPENG XU, Walmart Global Tech, USA
JASON H.D. CHO, Walmart Global Tech, USA
KAUSHIKI NAG, Walmart Global Tech, USA
EVREN KORPEOGLU, Walmart Global Tech, USA
SUSHANT KUMAR, Walmart Global Tech, USA
KANNAN ACHAN, Walmart Global Tech, USA

Knowledge Graphs (KGs) play a crucial role in enhancing e-commerce system performance by providing structured information about entities and their relationships, such as complementary or substitutable relations between products or product types, which can be utilized in recommender systems. However, relation labeling in KGs remains a challenging task due to the dynamic nature of e-commerce domains and the associated cost of human labor. Recently, breakthroughs in Large Language Models (LLMs) have shown surprising results in numerous natural language processing tasks. In this paper, we conduct an empirical study of LLMs for relation labeling in e-commerce KGs, investigating their powerful learning capabilities in natural language and effectiveness in predicting relations between product types with limited labeled data. We evaluate various LLMs, including PaLM and GPT-3.5, on benchmark datasets, demonstrating their ability to achieve competitive performance compared to humans on relation labeling tasks using just 1 to 5 labeled examples per relation. Additionally, we experiment with different prompt engineering techniques to examine their impact on model performance. Our results show that LLMs significantly outperform existing KG completion models in relation labeling for e-commerce KGs and exhibit performance strong enough to replace human labeling.

CCS Concepts: • **Applied computing** → **E-commerce infrastructure**; • **Computing methodologies** → **Machine learning**.

Additional Key Words and Phrases: Knowledge Graph, LLM, Few-shot Learning, E-commerce

ACM Reference Format:
Jiao Chen, Luyi Ma, Xiaohan Li, Nikhil Thakurdesai, Jianpeng Xu, Jason H.D. Cho, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. Knowledge Graph Completion Models are Few-shot Learners: An Empirical Study of Relation Labeling in E-commerce with LLMs. In . ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Knowledge Graphs (KGs) have emerged as a powerful tool for representing structured information about entities and their relationships. One of the core tasks of KGs is Knowledge Graph Completion (KGC), which is to predict the relations [5] that haven't been observed between entities. KGC offers significant benefits in e-commerce, such as relation labeling in product types. By capturing complementary or substitutable relations between product types, KGs enable

*All three authors contributed equally to this research.

e-commerce platforms to provide more accurate recommendations for users. However, the process of relation labeling in KGs faces numerous challenges, including the dynamic nature of e-commerce domains and the increasing cost of human labor.

Recent Large Language Models (LLMs), e.g., the 175B-parameter GPT-3 [2] and the 540B-parameter PaLM [4], are model with a very language amount of parameters. They show surprising abilities, which are called emergent abilities [34]) in solving a series of complex tasks. A remarkable application of LLMs is ChatGPT [1], which adapts the LLMs from the GPT series for dialogue, presents an amazing conversation ability with humans. The powerful capabilities of these models present a potential solution for the challenges faced in relation labeling in e-commerce KGs. LLMs can understand the semantic meanings of product types without training. This paper aims to conduct an empirical study of LLMs for relation labeling in e-commerce KGs, specifically focusing on their few-shot learning capabilities and effectiveness in predicting relations between product types with limited labeled data.

In our experiments, we focus on examining the KGC between product types [16] to predict the complementary [22, 23] and substitutable relations. Product types serve to categorize and group similar products together. While retail platforms like Amazon, eBay, and Walmart may offer millions of distinct products, the number of product types typically remains below 10 thousand. This relatively small number allows for a more nuanced and accurate definition of product relationships. Additionally, product types are well-defined in natural language and can be effectively modeled by KGs, making the research problem well-suited for KG completion. Specifically, given an source (src) product type, our goal is to predict whether it 'is_complementary_to', 'is_substitutable_for' or 'is_irrelevant_to' to another destination (dst) product type. These identified product types can then be utilized to generate high-quality recall item sets for downstream item-level complementary or substitutable recommendations.

In this paper, we evaluate various LLMs, including PaLM [4] and GPT-3.5 [2], on benchmark datasets to assess their performance on relation labeling tasks using as few as 1 to 5 labeled examples per relation. We also experiment with different prompt engineering techniques to examine their impact on model performance. Our study demonstrates that LLMs significantly outperform existing KG completion models in relation labeling for e-commerce KGs and exhibit performance levels strong enough to replace human labeling. Moreover, LLMs are not only capable of predicting relations but also provide explanations for their labeling decisions regarding the product type pairs in a given relation. Furthermore, we discover that the explanations provided by LLMs is very likely to be agreed by humans if they read them and then change their own labeling results.

This paper is structured as follows: Section 2 describes the settings and datasets used in our experiments. Section 3 presents the results and discussions of the impact of prompt engineering. Section 4 illustrates the labeling results comparison between humans and LLMs. Section 5 shows the comparison experiments between an LLM model PaLM and eight KG models. Section 6 provides an overview of related work in the areas of knowledge graph completion and LLM applications. Finally, Section 7 concludes the paper and suggests future research directions.

Our contributions are summarized as follows:

- To the best of our knowledge, this paper represents the first attempt to apply LLMs to KGC tasks in e-commerce contexts. We demonstrate that LLMs possess robust capabilities in predicting complementary and substitutable relations between product types, facilitated by their adeptness at processing natural language.

[1] https://chat.openai.com/

- In our experiments, we explore various prompts and identify the most effective way to frame our target task in a few-shot learning context. The performance achieved through our proposed prompt engineering approach is competitive with human labeling and can be readily applied in real-world business scenarios.
- We find that LLMs are much powerful than the state-of-the-art KG models with a minimum improvement of 40.6%. The experiments also demonstrate that LLMs are scalable especially when the number of labeled data is limited.

2 EXPERIMENT SETTINGS

In the experiments of KG relation labeling with LLMs, we first introduce the datasets and their statistics. We considered product types from the Electronics department in Walmart [2] and the aisles as product types in online grocery Instacart [3] [28]. The ground truth of the relation labeling is from the consensus of different people through crowdsourcing so we assume this label can be fairly used to evaluate the performance. The temperature of the LLMs are set to 0.0 to ensure the consistent and stable outputs.

In the Electronic dataset, we sampled 1045 pairs of product types, where the labels of the ground truth are 769 for 'irrelevant', 264 for 'complementary' and 12 for 'substitutable'. In the Instacart dataset, we sampled 400 pairs of product types based on their co-occurrence frequency, with 244 'irrelevant' lables, 166 'complementary' labels, and 10 'substitutable' labels. As each product type is a set of similar products, on the product type level the number of relation 'substitutable' is relatively small.

The LLM's predictions and consensus human labels are evaluated on overall accuracy, precision and recall corresponding to complementary or substitutable labels. For evaluating LLMs and humans, we use human labels as ground truth. The accuracy is calculated as $N_{common\_labels}/N_{total\_labels}$, where $common\_labels$ means the common labels between human and LLM labeling results. The precision is calculated for complementary or substitutable relation respectively as $N_{common\_labels}/N_{LLM\_predicted}$ and the recall is calculated for each relation as $N_{common\_labels}/N_{human\_labeled}$.

3 PROMPT ENGINEERING

The effectiveness of LLMs in various natural language processing tasks often relies on the design of suitable prompts. In this section, we describe our approach to design prompts for LLMs for the task of relation labeling of product types in e-commerce KGs. We apply PaLM [4] and GPT-3.5 [2] to evaluate its performance on relation labeling in e-commerce.

To design effective prompts for LLMs, we follow four guiding principles as follows. In Fig. 1, each part in the prompt examples corresponds to a principle.

- *Clarity (Part 1)*: Ensure that the prompts can clearly describe *the definition of the relation labeling task*, providing enough context for LLMs to understand the task and the desired output. Few-shot Learning may also be applied as a limited number of examples of the task (e.g., *pairs of product types in pink* for each item relationship.).
- *Relevance (Part 2)*: Set up a role of the LLM and the context of the e-commerce scenario to enhance the model's understanding of the task.
- *Format (Part 3&4)*: Frame the input data in the prompts (Part 3) with a clear tuple-like format. The output of the LLM (Part 4) should also follow a certain format to make the results readable.

[2] walmart.com
[3] https://www.kaggle.com/c/instacart-market-basket-analysis

Next, we will introduce how these principles affect the relation labeling performance of LLM. Based on the principles above, we compare the effect of different principles step by step by completing the prompt in Part 1 in Fig. 1.

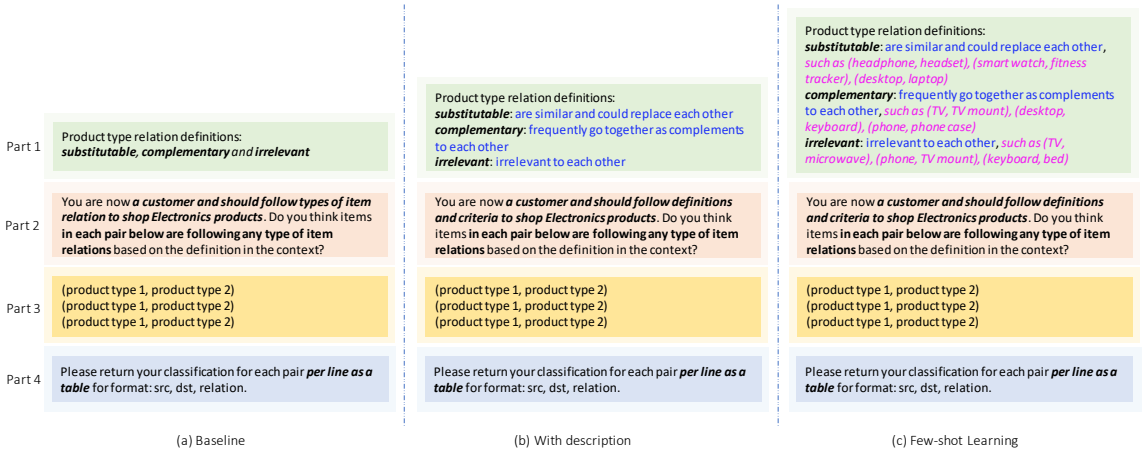


Fig. 1. Prompt examples with different principles.

(a) The baseline. Here we define the baseline prompt with role of LLM, the relation labeling task with a scenario and the output format with Markdown. The accuracy of the baseline prompt is **0.575**.

(b) With relation description. With the principle Clarity, we also take the part 1 in the Fig.1 into consideration. The prompt will be changed as follows. We highlight the difference in blue. The accuracy of the prompt with relation description is **0.676**, with a **17.6%** improvement compared to the baseline prompt.

(c) With few-shot Learning. On top of the description of relations, we also give each relation a few examples to guide LLMs in performing the task with minimal labeled data. The few-shot examples are highlighted in pink. The accuracy of the prompt with relation description is **0.738**, with a **28.3%** improvement compared to the baseline prompt.

In Table 1, we also put the complete experiment results of prompt engineering on Electronics and Instacart datasets. Please note that we only apply PaLM on Electronics dataset because of data privacy issue. In Instacart dataset, PaLM's results are better than GPT-3.5's for all the prompts in terms of Accuracy. The precision and recall scores for the 'substitutable' relation are relatively low because this relation appears infrequently; thus, the results are easily affected by incorrect predictions, leading to a bias in the scores. From the results of these two tables, we can find that the relation definition and few-shot learning with 3 or 5 examples can lead to a significant improvement on the prediction accuracy.

Dataset	LLM	Prompt	Complementary		Substitutable		Accuracy
			Precision	Recall	Precision	Recall	
Electronics	PaLM	Baseline	0.389	**0.807**	0.083	0.500	0.575
		zero_shot	0.424	0.678	**0.240**	0.500	0.676
		one_shot	0.446	0.667	0.227	0.417	0.695
		few_shot_3	0.506	0.633	0.222	0.500	**0.738**
		few_shot_5	**0.507**	0.580	0.136	0.500	0.725
Instacart	PaLM	Baseline	0.599	**0.786**	0.167	0.444	0.645
		zero_shot	0.705	0.656	0.161	**0.556**	0.699
		one_shot	0.664	0.740	**0.300**	0.333	0.712
		few_shot_3	0.699	0.725	0.222	0.444	0.726
		few_shot_5	**0.711**	0.733	0.250	0.444	**0.739**
	GPT-3.5	Baseline	0.636	0.519	0.091	**0.778**	0.572
		zero_shot	0.595	0.695	0.125	0.444	0.632
		one_shot	0.598	0.656	0.135	0.556	0.622
		few_shot_3	**0.659**	0.618	0.133	0.667	0.635
		few_shot_5	0.632	**0.695**	**0.167**	0.444	**0.666**

Table 1. LLM label results on Electronics and Instacart datasets. (1) PaLM label results on Electronics product types. (2) PaLM and ChatGPT results on Instacart product types.

4 LLM AS INDIVIDUAL HUMAN LABELER

Human consensus results offer costly but accurate labels through crowdsourcing, enhancing the labeling quality by incorporating the input of multiple labelers. Although individual labelers may make mistakes in labeling tasks, they can still provide different yet valid labels compared to consensus results due to their diverse backgrounds and experiences. For instance, if two labelers come from different regions with distinct dietary habits, they might offer different but valid labels of relationships for grocery product types. To further investigate the LLM's performance and the gap between it and individual labelers, we compare LLM's results with those of individual human labelers under **independent labeling** and **dependent labeling** settings. For all subsequent experiments, we consider the results from PaLM with the prompt (few_shot_5) as LLM's labels for Instacart product type pairs, and the results from PaLM with the prompt (few_shot_3) as LLM's labels for Electronics product type pairs, due to their superior performance on human consensus results.

4.1 LLM Results vs Individual Human Labelers (independent labeling)

In this experiment, two human labelers with different cultural backgrounds but extensive experience in e-commerce shopping independently label the relationships of all pairs of the aforementioned Electronics and Instacart product types, respectively. They are not allowed to discuss their findings with each other or review the LLM's results. To initially understand the gap between the LLM and individual labelers, we treat each individual human labeler's results as ground truth and evaluate the LLM's results with those of the two individual labelers, respectively. To further understand the impact of a human labeler's background on labeling tasks, we treat labeler 2's results as ground truth

and evaluate labeler 1's results. We report the precision, recall, and accuracy metrics as defined in Section 2 in Table 2. It is important to note that there is no actual ground truth between the two individual human labelers; the precision and recall reported here simply represent the proportion of agreed-upon labels for human labelers 1 and 2.

From the accuracy results in Table 2, for Electronics product types, the human-human accuracy is 0.76, and the LLM's accuracy with labeler 2 is very close to the human-human results. On the Instacart dataset, the accuracy between the LLM and labeler 2 (0.665) surpasses the accuracy between human and human (0.598). Notably, from the precision and recall results, LLM usually has a low precision value for 'substitutable' labels with both human labelers, which indicates LLM tends to generate more 'substitutable' results. In summary, LLM's performance is comparable to that of individual labelers, considering the accuracy between LLM and human as well as the accuracy between human and human. Furthermore, an individual's background does indeed influence their labeling performance.

Dataset	**Prediction**	**Ground truth**	**Complementary**		**Substitutable**		Accuracy
			Precision	Recall	Precision	Recall	
Electronics	LLM	Labeler 1	**0.822**	0.557	0.161	**0.833**	0.687
	LLM	Labeler 2	0.742	0.652	0.322	0.769	0.74
	Labeler 1	Labeler 2	0.651	**0.843**	**0.667**	0.308	**0.76**
Instacart	LLM	Labeler 1	0.356	0.638	0.0968	0.214	0.547
	LLM	Labeler 2	**0.654**	**0.687**	0.129	**0.308**	**0.665**
	Labeler 1	Labeler 2	0.628	0.369	**0.286**	0.308	0.598

Table 2. Evaluation with human independently labeled results.

Dataset	**Prediction**	**Ground Truth**	**Complementary**		**Substitutable**		Accuracy
			Precision	Recall	Precision	Recall	
Electronics	LLM	Labeler 1	**0.832**	0.622	0.29	**0.9**	0.74
	LLM	Labeler 2	0.782	0.675	0.387	0.8	0.767
	Labeler 1	Labeler 2	0.696	**0.803**	**0.8**	0.533	**0.78**
Instacart	LLM	Labeler 1	**0.723**	**0.932**	**0.548**	**0.944**	**0.82**
	LLM	Labeler 2	0.669	0.717	0.161	0.417	0.695
	Labeler 1	Labeler 2	0.712	0.601	0.278	0.417	0.718

Table 3. Human relabeling based on LLM results. LLM label results are generated **with explanations**.

4.2 Human Relabeling based on LLM Results (dependent labeling)

Owing to the limitations of individual knowledge, a human labeler may lack information for some products, resulting in incorrect labeling outcomes. To further assess the quality of LLM's labels, we modify part 4 of our prompt template to request that LLM provide both labels and explanations. We then ask our two labelers to re-label the product types for both datasets, taking into account LLM's labels and explanations. The evaluation results for LLM-human and human-human comparisons are presented in Table 3.

On the Electronics dataset, the accuracy between both LLM-human and human-human comparisons is slightly improved compared to the results in Table 2. However, on the Instacart dataset, the accuracy between LLM-labeler 1 and human-human increased by more than 10%. This could be due to the fact that relations between most electronic products are objective and easy to determine, while relationships between grocery products are more subjective and challenging to ascertain. For example, people with different dietary habits might have differing opinions on grocery product relationships. Additionally, the agreement between LLM-labeler 1 on the Instacart dataset is much higher than the agreement between human-human, indicating that LLM's explanations have convinced labeler 1 in many grocery pair cases.

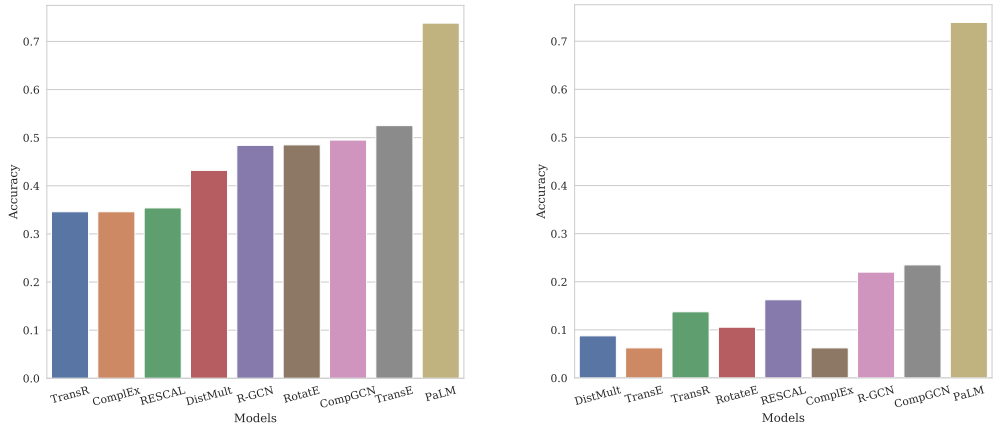
Typically, we notice that labeler 1 changed 110 labels after seeing LLM's labels and explanations compared with the independent labeling task, mostly due to 62 pairs changed from 'irrelevant' to 'complementary' and 25 from 'complementary' to 'irrelevant'. For example, when the src product type is 'yogurt' and the dst product type is 'fresh dips tapenades', labeler 1 tags them as 'irrelevant' mainly because they are not common combination in the food culture of labler 1. LLM tags them as 'complementary' with explanation '*yogurt and fresh dips tapenades can both be used as a snack or appetizer. they can also be eaten separately*', which convinces labeler 1 to change the labels. Another examples 'canned jarred vegetables' and 'milk'. Labeler 1 change label from 'complementary' to 'irrelevant' after checking LLM's explanation '*canned and jarred vegetables are both processed forms of vegetables, while milk is a dairy product. they are not typically used in the same recipes, and they do not have the same nutritional value*', which addresses more on nutrition compatibility.

Through both independent and dependent labeling tasks, we demonstrate that LLMs can perform competitively compared to human labelers, taking into account individual differences. Additionally, LLMs provide valuable labeling explanations that contribute to better label quality in KG completion tasks.

5 COMPARISON EXPERIMENTS

We conduct comparison experiment of an LLM PaLM with different KG models. The baseline models are TransE [1], TransR [17], DistMult [37], ComplEx [29], RESCAL [26], R-GCN [27] and CompGCN [30]. The product types in all KG models are initialized with word embeddings from Word2Vec [25]. The experiments are conducted on both Electronics and Instacart datasets with human consensus labels, and we split the datasets as 80% for training, 10% for validation and 10% for testing. The detailed results of the experiments are shown in Fig. 2. From the two figures, we have the following observations:

- We observe that PaLM significantly outperforms all knowledge graph models on both datasets, with the minimum improvement being 40.6%. This can be attributed to the fact that KG models require a substantial amount of training data, while relation labeling in e-commerce is expensive, resulting in limited labeled data for our task. Furthermore, the labeled data does not cover all product types, leading to instances where some product types never appear in the training set. In contrast, LLMs leverage their understanding of human language to enhance the accuracy of their predictions, even when dealing with limited training data or unseen product types.
- The LLM model PaLM exhibits similar accuracy on both datasets. However, in Fig. 2 (b), KG models perform poorly on the Instacart dataset due to the limited availability of only 320 training pairs. This observation highlights that, in contrast to KG models that require large amounts of data for parameter optimization, LLM models are more scalable and their performance is less influenced by the quantity of labeled data.



(a) Electornics (b) Instacart

Fig. 2. Comparision of accuracy between KG models and the LLM in the Electornics and Instacart dataset.

6 RELATED WORKS

6.1 Knowledge Graph Completion in E-commerce

Knowledge Graphs (KGs) have garnered significant attention in recent years due to their ability to represent structured information about entities and their relationships. They have been widely adopted across various domains, including e-commerce, to enhance user experiences and facilitate decision-making. In this section, we review the literature on Knowledge Graph Completion (KGC) in the context of e-commerce.

With the emergence of embedding techniques, several KGC approaches have employed embeddings to represent entities and relations in e-commerce KGs. These methods, such as TransE [1], DistMult [37], and ComplEx [29], learn low-dimensional vector representations of entities and relations, enabling the discovery of complex patterns and relationships within the KG. Recent advances in neural networks have led to the development of more sophisticated KGC methods in e-commerce. Convolutional Neural Networks (CNNs) have been employed for KGC tasks, such as in ConvE [8], where the model learns embeddings by exploiting local and global connectivity patterns in the graph. Similarly, Graph Neural Networks (GNNs) [12] have demonstrated its capacities to capture both structural and semantic information [3, 13, 15, 19, 20, 32]. They are also applied to KGs, such as R-GCN [27] and CompGCN [30]. The few-shot learning in KGC [35, 38] can also improve the performance when labeled data is scarce. KGC methods have a significant impact on many applications in e-commerce, including recommender systems [14, 18, 32], product relation labeling [36] and product taxonomy [24].

While these methods have demonstrated improved performance, they still exhibit limitations in comprehending natural language, which is crucial for e-commerce KGs. Moreover, these approaches continue to face challenges in addressing the scarcity of labeled data, primarily due to the expensive cost of human labor. By employing LLMs, we can capitalize on their capacity to understand natural language and label relations within the context of few-shot learning, potentially overcoming these challenges and enhancing KG completion accuracy in e-commerce domains.

6.2 LLM Applications in E-commerce

Large Language Models (LLMs) have gained significant traction in recent years due to their remarkable performance in a wide range of natural language processing related tasks [7, 21, 33]. In this section, we review the literature on LLM applications in e-commerce.

One of the most common applications of LLMs in e-commerce is the enhancement of recommender systems. By leveraging LLMs' natural language understanding capabilities, researchers have been able to provide more accurate and personalized product recommendations for users. For example, LLMs have been used to learn from users' behaviors in natural language so that they can serve as recommender systems to directly make recommendations [6, 9]. Moreover, as LLMs' success in conversaional AI, there are some new applications such as conversational recommendation [31] to enhance the customers' experience.

LLMs have been employed across a range of applications in e-commerce, including customer support [10], sentiment analysis [33], and text classification [11]. These applications help e-commerce platforms better understand product information, customer feedback, and preferences, ultimately leading to more targeted marketing strategies and improved user experiences. However, their application in Knowledge Graph Completion (KGC) remains relatively unexplored, particularly in the context of e-commerce. In this paper, we aim to bridge this gap by investigating LLMs' potential for predicting complementary and substitutable relations between product types in e-commerce KGs.

7 CONCLUSION

This study contributes to the understanding of LLMs' potential in e-commerce KG completion tasks and demonstrates their value in overcoming challenges associated with limited labeled data and human labor costs. Our results revealed that LLMs significantly outperform existing KG completion models in relation labeling for e-commerce KGs and exhibit performance strong enough to replace human labeling. As a pioneering effort in applying LLMs to KGC tasks in e-commerce, our findings pave the way for future research and practical applications of LLMs in e-commerce such as item description summarizing or recommendation.

REFERENCES

[1] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Ziheng Chen, Fabrizio Silvestri, Jia Wang, Yongfeng Zhang, Zhenhua Huang, Hongshik Ahn, and Gabriele Tolomei. 2022. GREASE: Generate Factual and Counterfactual Explanations for GNN-based Recommendations. *arXiv preprint arXiv:2208.04222* (2022).

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[5] Zijun Cui, Pavan Kapanipathi, Kartik Talamadupula, Tian Gao, and Qiang Ji. 2021. Type-augmented relation prediction in knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7151–7159.

[6] Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems. *arXiv preprint arXiv:2205.08084* (2022).

[7] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. 2023. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007* (2023).

[8] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[9] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.

[10] A Shaji George and AS Hovan George. 2023. A review of ChatGPT AI's impact on several business sectors. *Partners Universal International Innovation Journal* 1, 1 (2023), 9–23.

[11] Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Practical text classification with large pre-trained language models. *arXiv preprint arXiv:1812.01207* (2018).

[12] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *Proceedings of the International Conference on Learning Representations (ICLR)* (2016).

[13] Xiaohan Li, Yuqing Liu, Zheng Liu, and S Yu Philip. 2022. Time-aware Hyperbolic Graph Attention Network for Session-based Recommendation. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 626–635.

[14] Xiaohan Li, Zhiwei Liu, Stephen Guo, Zheng Liu, Hao Peng, S Yu Philip, and Kannan Achan. 2021. Pre-training recommender systems via reinforced attentive multi-relational graph neural network. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 457–468.

[15] Xiaohan Li, Mengqi Zhang, Shu Wu, Zheng Liu, Liang Wang, and S Yu Philip. 2020. Dynamic graph collaborative filtering. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 322–331.

[16] Jiunn-Woei Lian and Tzu-Ming Lin. 2008. Effects of consumer characteristics on their acceptance of online shopping: Comparisons among different product types. *Computers in human behavior* 24, 1 (2008), 48–65.

[17] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 29.

[18] Zhiwei Liu, Xiaohan Li, Ziwei Fan, Stephen Guo, Kannan Achan, and S Yu Philip. 2020. Basket recommendation with multi-intent translation graph neural network. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 728–737.

[19] Zheng Liu, Xiaohan Li, Hao Peng, Lifang He, and S Yu Philip. 2020. Heterogeneous similarity graph neural network on electronic health records. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 1196–1205.

[20] Zheng Liu, Xiaohan Li, Zeyu You, Tao Yang, Wei Fan, and Philip Yu. 2021. Medical triage chatbot diagnosis improvement via multi-relational hyperbolic graph neural network. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1965–1969.

[21] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621* (2023).

[22] Luyi Ma, Nimesh Sinha, Jason HD Cho, Sushant Kumar, and Kannan Achan. 2023. Personalized diversification of complementary recommendations with user preference in online grocery. *Frontiers in big Data* 6 (2023).

[23] Luyi Ma, Jianpeng Xu, Jason HD Cho, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2021. NEAT: A Label Noise-resistant Complementary Item Recommender System with Trustworthy Evaluation. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 469–479.

[24] Félix Martel and Amal Zouaq. 2021. Taxonomy extraction using knowledge graph embeddings and hierarchical clustering. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. 836–844.

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[26] Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel, et al. 2011. A three-way model for collective learning on multi-relational data.. In *Icml*, Vol. 11. 3104482–3104584.

[27] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer, 593–607.

[28] Jeremy Stanley. 2017. The Instacart Online Grocery Shopping Dataset 2017. https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2

[29] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*. PMLR, 2071–2080.

[30] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2019. Composition-based multi-relational graph convolutional networks. *Proceedings of the International Conference on Learning Representations (ICLR) 2019* (2019).

[31] Lingzhi Wang, Huang Hu, Lei Sha, Can Xu, Kam-Fai Wong, and Daxin Jiang. 2021. Finetuning large-scale pre-trained language models for conversational recommendation with knowledge graph. *arXiv preprint arXiv:2110.07477* (2021).

[32] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 950–958.

[33] Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study. *arXiv preprint arXiv:2304.04339* (2023).

[34] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

[35] Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2018. One-shot relational learning for knowledge graphs. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[36] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2020. Product knowledge graph embedding for e-commerce. In *Proceedings of the 13th international conference on web search and data mining*. 672–680.

[37] Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*.

[38] Chuxu Zhang, Huaxiu Yao, Chao Huang, Meng Jiang, Zhenhui Li, and Nitesh V Chawla. 2020. Few-shot knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 3041–3048.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009