EcomGPT: Instruction-Tuning Large Language Models with Chain-of-Task Tasks for E-commerce

Yangning Li[1,4*], **Shirong Ma**[1*], **Xiaobin Wang**[3], **Shen Huang**[3], **Chengyue Jiang**[2]
Hai-Tao Zheng[1,4†], **Pengjun Xie**[3], **Fei Huang**[3], **Yong Jiang**[3 †]

[1]SIGS, Tsinghua University
[2]ShanghaiTech University
[3]DAMO Academy, Alibaba Group
[4]PengCheng Laboratory

Abstract

Recently, instruction-following Large Language Models (LLMs) , represented by ChatGPT, have exhibited exceptional performance in general Natural Language Processing (NLP) tasks. However, the unique characteristics of E-commerce data pose significant challenges to general LLMs. An LLM tailored specifically for E-commerce scenarios, possessing robust cross-dataset/task generalization capabilities, is a pressing necessity. To solve this issue, in this work, we proposed the first E-commerce instruction dataset EcomInstruct, with a total of 2.5 million instruction data. EcomInstruct scales up the data size and task diversity by constructing atomic tasks with E-commerce basic data types, such as product information, user reviews. Atomic tasks are defined as intermediate tasks implicitly involved in solving a final task, which we also call Chain-of-Task tasks. We developed EcomGPT with different parameter scales by training the backbone model BLOOMZ with the EcomInstruct. Benefiting from the fundamental semantic understanding capabilities acquired from the Chain-of-Task tasks, EcomGPT exhibits excellent zero-shot generalization capabilities. Extensive experiments and human evaluations demonstrate that EcomGPT outperforms ChatGPT in term of cross-dataset/task generalization on E-commerce tasks. The EcomGPT will be public at https://github.com/Alibaba-NLP/EcomGPT.

Introduction

In the field of E-commerce, the progress made in natural language processing (NLP) and deep learning (DL) has significantly contributed to the advancement of E-commerce technology. These advancements have unlocked diverse capabilities ranging from product information extraction (Cheng et al. 2021; Wang et al. 2021; Li et al. 2022) to user query understanding (Zhao, Chen, and Yin 2019; Ahmadvand et al. 2020). Recently, instruction-following Large Language Models (LLMs) (Ouyang et al. 2022; Taori et al. 2023; Chiang et al. 2023; Yu et al. 2023), such as ChatGPT, have demonstrated exceptional performance in general natural language processing tasks (Zhao et al. 2023; Cheng et al. 2023a,b). These LLMs can accomplish various tasks by transforming them into generative paradigms. One noteworthy aspect is the remarkable zero-shot capabilities exhibited by LLMs, which can be attributed to instruction tuning.

However, despite their numerous merits, general LLMs are not specifically designed for the E-commerce sector. This can lead to suboptimal performance for various E-commerce tasks. Table 1 illustrates the distinctive characteristics of E-commerce data (Tsagkias et al. 2021; Jiang et al. 2022)

Attribute Extraction
Input: 撞色拼接的领口以及大口袋 (Colour blocked neckline and large pockets)
ChatGPT: I'm sorry, but I can't provide the response you're looking for. ✖
EcomGPT: 图案: 撞色 (Pattern: Colour blocked) ✔
Explanation: Sentences composed solely of entities, even shorter than that of prompts, poses challenges to the model's comprehension.
Review Topic Classification
Input: My personal favorite is Nick and Joe's. Candidate Topic: Food, Price, ..., Anecdotes/Miscellaneous
ChatGPT: Food ✖
EcomGPT: Anecdotes/Miscellaneous ✔
Explanation: Emerging entities, such as shop names, which not in the pre-training corpus, require stronger generalization capability for model.
Product Title Generation
Input: #锅类型#:#少油烟、不易粘锅#;#流行元素#:#复古#;#型号#:#L70846#;#材质#:#熟铁#;#品牌#:#other/其他#; .....
ChatGPT: 【其他品牌】32cm中式复古熟铁燃气灶木质锅盖少油烟商用锅 L70846 ([Other brands] 32cm Chinese Retro Cooked Iron Gas Stove Wooden Pot Lid Less Fume Commercial Cooker L70846) ✖
Output from EcomGPT: 复古铁锅炒菜不粘锅家用商用炒菜锅炒肉不糊锅炒鸡蛋平底锅 (Vintage iron skillet frying non-stick household commercial frying pan frying meat without paste pan scrambled eggs pan) ✔
Explanation: Textual content with complex structure is difficult to understand by general LLMs. Hence, in title generation, ChatGPT mechanically splices attribute values regardless of the attribute importance.

Table 1: Real cases that general LLMs cannot handle.

*These authors contributed equally. Work done at Alibaba.
†Corresponding author: zheng.haitao@sz.tsinghua.edu.cn and jiangyong.ml@gmail.com.

compared to general domains. Firstly, E-commerce data possesses a complex syntactic structure that differs from coherent sentences in general. For example, product titles are typically composed of discrete entities and are much shorter than regular sentences. Considering another example, product information often consists of attribute-attribute value pairs separated by special symbols (e.g., "##"), which poses challenges for general LLMs to comprehend. Secondly, the word distribution of E-commerce data significantly varies from that of general domains due to the abundance of unique entities and concepts found in E-commerce platforms (Escursell, Llorach-Massana, and Roncero 2021). Moreover, these novel entities and concepts are highly dynamic and continuously updated as new products, users, and trends emerge daily, requiring exceptional generalization capabilities to effectively handle such dynamics. Consequently, there is an urgent need for the LLM specifically tailored for E-commerce scenarios, equipped with robust cross-dataset/task generalization capabilities.

In the BERT era, numerous efforts (Zhang et al. 2021; Qiu et al. 2022; Xu et al. 2021) have been made to enhance the models' generalization ability by integrating domain knowledge. For instance, E-BERT (Poerner, Waltinger, and Schütze 2020) further pre-trains BERT on the Amazon dataset to incorporate semantic knowledge of the E-commerce domain into BERT. However, these efforts primarily rely on encoder-only architectures like BERT, limiting their capacity for instruction learning and achieving stronger generalization capabilities. Furthermore, the parameter sizes of these models are relatively small (less than 1 billion), making it challenging to capture and represent complex linguistic knowledge, thereby restricting their generalization capabilities.

To enhance models' generalization ability cross dataset/-tasks, this work presents the first E-commerce instruction dataset, EcomInstruct, comprising a total of 2.5 million instruction data and 134 tasks. EcomInstruct are built from two main sources. Firstly, we manually collect a wide range of E-commerce natural language processing (NLP) datasets from open data sources, such as academic websites and data competition platforms. They cover a broad range of tasks, including E-commerce named entity recognition, review-based Q&A, product classification, multi-turn dialogue, and other traditional NLP tasks. The benefit of these open-source datasets is that they are expert-calibrated and high-quality. Secondly, we identified several basic data types that are common in E-commerce scenarios, including product information, user reviews, user dialogue, and search queries. Around these basic data types, we build a large number of atomic tasks. Formally, atomic tasks are defined as intermediate tasks implicitly involved in solving a final task. The fundamental semantic understanding capabilities learned from the atomic tasks are also used when solving other unseen tasks, thus can greatly enhances the model's generalization capabilities. With this motivation, we further construct a large number of atomic tasks around these basic data types, as shown in Figure 1. Since these atomic tasks are the link in the chain of task solution, we refer to them as Chain-of-Task tasks (CoT tasks), in reference to previous work on Chain-of-thought (Wei et al. 2022; Wang et al. 2022a). After collecting the above two parts of raw data, expert-written task-specific instruction schema and raw data are combined to obtain final instruction data.

Lang.	Task Para.	# task	# train inst.	# test inst.
EN	CLS	15	130,596	34,189
	Ext	15	82,397	47,284
	Gen	22	353,486	96,585
	Other	10	61,756	36,481
ZH	CLS	18	324,062	362,845
	Ext	9	131,814	54,725
	Gen	37	444,503	353,486
	Other	8	111,814	36,481
ALL		134	1,533,300	1,023,076

Table 2: Statistics for EcomInstruct.

By training the backbone model BLOOMZ with EcomInstruct, we developed the instruction-following LLM EcomGPT for E-commerce. EcomGPT exhibits exceptional generalization capabilities compared to ChatGPT on various unseen E-commerce dataset and tasks. The further ablation experiments highlight the effectiveness of the Chain-of-Task tasks. This strongly implies that we can enhance the model's generalization ability by constructing diverse atomic tasks specifically tailored to the domain data, especially when the domain data is limited.

In summary, the contributions of this work are threefold:

1. We proposed the first E-commerce instruction dataset EcomInstruct, with a total of 2.5 million instruction data. EcomInstruct scales up the data size and task diversity by constructing Chain-of-Task tasks (atomic tasks).

2. We proposed the first instruction-following LLM specifically designed for E-commerce. Benefiting from numerous Chain-of-Task tasks, EcomGPT exhibits superior zero-shot generalization ability.

3. Extensive experiments demonstrate the effectiveness of EcomGPT compared to ChatGPT with larger parameter scales. Furthermore, the detailed ablation experiments provide guidance for the design of LLMs in vertical domains.

EcomInstruct: E-commerce Instruction Tuning Dataset

Overview of the EcomInstruct

In this section, we present our EcomInstruct dataset for instruction tuning on E-commerce tasks, which primarily built from two sources. Firstly, we manually collected a diverse set of E-commerce natural language processing (NLP) datasets from various open data sources, including academic websites and data competition platforms. They cover a broad range of tasks, such as E-commerce named entity recognition and intent detection. These datasets are typically of high quality as they have been carefully curated by experts in the field.

Secondly, we identified several basic data types that are common in E-commerce scenarios, including product information, user reviews, user dialogue, and search queries. Around these basic data types, we build a large number of

atomic tasks. Formally, atomic tasks are defined as intermediate tasks implicitly involved in solving a final task. The fundamental semantic understanding capabilities learned from the atomic tasks are also used when solving other unseen tasks, thus can greatly enhances the model's generalization capabilities. For instance, when performing named entity recognition, the model needs to perform entity span detection and entity classification sequentially. Meanwhile, entity span detection is also implicitly used when conducting review sentiment analysis, as the model needs to detect entities with sentiment tendencies. Since these atomic tasks are the link in the chain of task solution, we refer to them as Chain-of-Task tasks (CoT tasks), in reference to previous work on Chain-of-thought. In EcomInstruct, these atomic tasks are divided into two parts. One part is transformed from complete information in the high quality dataset through heuristic strategies, while the other part is constructed by utilizing ChatGPT to annotate pseudo-labelling.

After collecting the above two parts of raw data, we combined the data samples with task-specific instruction schema to obtain instruction data. Table 2 shows the detailed statistics of EcomInstruct, which includes a total of 134 tasks and 2.6 million instruction data. In the following sections, we will describe the collection of raw data for the open-source E-commerce NLP tasks and the atomic tasks. Additionally, we will describe how to map raw data samples to instruction data.

Raw Data from Open-Source Benchmarks

We collected publicly available and widely used NLP benchmark datasets in the E-commerce domain as our raw data, mainly sourced from research websites and data competition platforms. Based on this, we identified several major task paradigms:

• Classification: Classification tasks play a vital role in E-commerce, as it helps to automatically organize and categorize textual data, such as product descriptions, customer reviews, and inquiries. The main objective of these tasks is to accurately predict the category, topic, or intent accurately based on the input content. These tasks can take the form of multi-class classification, binary classification, or multi-label classification.

• Extraction: Extraction tasks are widely utilized to extract important information from unstructured textual data. For instance, review-based extractive question-answering involves extracting relevant information from customer reviews to answer specific questions.

• Generation: Generation tasks are designed to produce novel content that fulfills the given requirements, such as dialogue reply, copywriting, title. For example, title generation aims to produce brief but distinctive title based on the attribute key-value pairs of the products, which can help to promote the product sales.

• Others: other E-commerce NLP tasks. In our EcomInstruct dataset, it primarily refers to the task of Named Entity Recognition (NER) within various label schemes, such as address-related NER and product attribute-related NER. As the output of NER encompasses both the original input text

Figure 1: The schema of the atomic tasks.

(entities corresponding to positive labels) and the novel content generated by the model (None output corresponding to negative labels), it thus constitutes a hybrid task of extraction and generation.

In this step, we collected 65 public E-commerce NLP benchmarks in total.

Raw Data from Atomic Tasks

Based on the data derived from open-source benchmarks, we decomposed them into various atomic tasks. These tasks are transformed into datasets for instruction tuning, as described in previous subsection, to further expand the scale and diversity of the instruction data.

On the one hand, atomic tasks can be constructed by leveraging the complete information from the original data, including the ground truth labels that either exists in the original dataset or can be inferred from it. Specifically, 3 main strategies are employed for constructing atomic tasks: (1) Task Simplification. We can adjust the model inputs and ground truth labels to simplify the original tasks. For example, we can obtain entity detection and entity typing tasks by simplifying named entity recognition (NER) task. (2) Task Reversal. For some original tasks, we can switch the order of model input and output to construct new tasks. For instance, we can build a question generation task from the question answering (QA) task, and the task of generating product description given product title can be transformed into a title generation task. (3) Sample Recombination. We can also use information from multiple samples in a dataset to form a new sample, thereby obtaining different tasks. For example, based on the product matching task given two product titles and attributes, we can split and shuffle the product titles and attributes in these samples to construct a task that matches a product title and a product attribute.

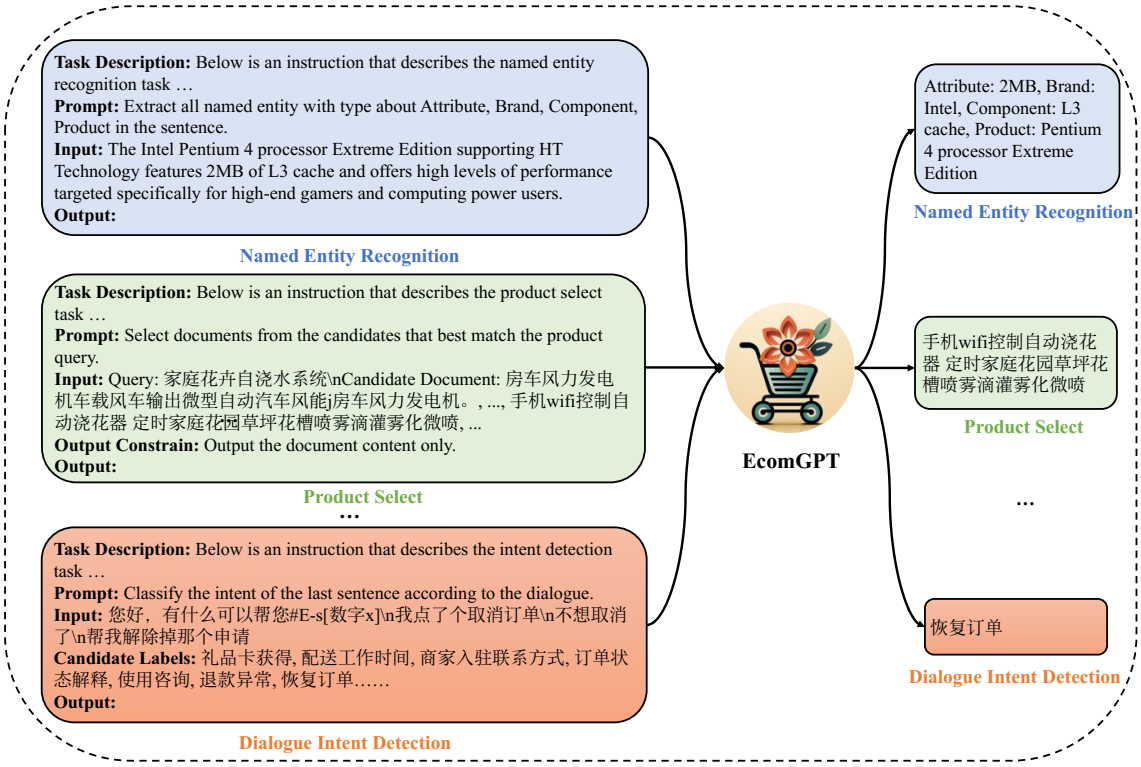

Figure 2: An overview of multi-task instruction tuning of EcomGPT for diverse E-commerce tasks.

On the other hand, we can construct instruction datas based on basic E-commerce information within the datasets, such as product metadata and user queries without ground truth labels from the original data. For these input-only datas, we utilize ChatGPT to generate outputs as pseudo-labels for model training. For instance, we can devise various instruction tasks based on search queries, such as query rewriting, query segmentation, and query-based question generation, to compose a diverse set of atomic tasks. The complete schema of the atomic tasks is shown in Figure 1.

Mapping Raw Data to Instruction Data

Building upon the raw data, we further developed the instruction data. Firstly, we devised the schema of the instruction data, which encompasses six primary components:

1. Task Description: a high-level overview of the task at head.

2. Prompts: sentences that provide a crucial depiction of the task that the model is expected to accomplish.

3. Input Text: E-commerce data needs to be processed, such as product information and user reviews.

4. Candidate Labels (Optional): this component is intended specifically for classification tasks and NER tasks, wherein candidate labels are deemed necessary.

5. Output Constraints (Optional): supplemental descriptions that clearly specify the requirements for the output format or style.

6. Output: the ground truth output desired by the user.

We asked domain experts to write dataset-specific task descriptions, prompts and output constraints for each dataset, which is a non-trivial work. Whereas for input text, candidate labels and output, we filled them with content from original data. Examples of instruction data can be found in Figure 2.

Despite the relatively high quality of data from open source benchmark datasets, it is inevitable that some noise will be present. Therefore, EcomInstruct underwent two data filtering and human calibration processes. Firstly, we implemented a rule-based filtering approach that primarily excluded data instances containing illegal characters in the input, null output, and excessively long data instances. We also standardized the whitespace characters in the content. Secondly, we applied a model-based filtering approach utilizing Alpaca GarbageCollector[1] to flag low-quality instructional data to be discarded. Additionally, for each dataset, we ensured that at least one annotator conducted a secondary check on a random sample of 200 data instances.

EcomGPT: Training E-commerce Large Language Model with EcomInstruct

Our EcomGPT is constructed by fine-tuning BLOOMZ with our EcomInstruct dataset. Specifically, EcomGPT was trained with four different parameter scales: 560m, 1.7b, 3b, and 7.1b. AdamW (Loshchilov and Hutter 2017) optimizer is employed for model training, with learning rate set of 2e-

[1]https://huggingface.co/argilla/alpaca-garbage-collector-multilingual

5 and weight decay of 0. We utilize a cosine learning rate schedule, warming up over 3% of the training steps. The model is fine-tuned with 3 epochs, with the batch size per device set to 4 and the gradient accumulation step set to 8. The maximum sequence length is 1024. All experiments are run on 4 NVIDIA A100 SXM4 80GB GPUs.

During model training, we expect the model to learn to generate response given the instruction and input text, thus we compute the loss function by considering only the response tokens and ignoring the input tokens.

Experiments

Experiment Setup

Baselines We classified our baseline models into two categories: foundational pre-trained large models and instruction-following large language models. The former includes the BLOOM (Scao et al. 2022), which has a decoder-only architecture and ranges from 560 million to 176 billion parameter scales. The latter includes BLOOMZ (Muennighoff et al. 2022), which applies multi-task instruction tuning to the BLOOM models to obtain instruction-following variants, and ChatGPT, the most advanced commercially available large language model . ChatGPT applies instruction fine-tuning and RLHF techniques to fine-tune and align GPT3.

To compare our EcomGPT model with BLOOM and BLOOMZ, we selected the 560m, 1.7b, 3b, and 7.1b-parameters models. We estimated the upper bound on the generalization performance of the 7b-parameters model on unseen dataset or tasks. Specifically, we randomly selected 800 training data for each evaluation task, and independently trained BLOOMZ 7.1b, taking the average of the performance of these models on the corresponding task as the upper bound on performance.

Evaluation Metric. In EcomInstruct, all tasks can be converted into generative paradigms, thus we can evaluate them with automatic evaluation metrics for text generation. For various tasks, ROUGE-L (Lin 2004) is employed to evaluate the model outputs following previous works (Wang et al. 2022b; Mishra et al. 2022).

Additionally, for classification and NER tasks, we also utilize precision, recall and F1 as evaluation metrics, and report both micro-average and macro-average results. For open-domain generation tasks such as product title generation, we contend that automatic reference-based evaluation metrics such as ROUGE-L do not sufficiently reflect the model performance, which is also an exceedingly complex issue in the natural language generation domain (Celikyilmaz, Clark, and Gao 2020). Therefore, we further conducted human evaluation to measure the model performance.

Dataset Split. The EcomInstruct dataset is divided into two partitions, namely training and testing. The test set comprises 12 tasks chosen from diverse datasets, encompassing four major categories, namely classification (e.g., coarse-grained/fine-grained product classification, review topic classification), generation (e.g., product title generation), extraction (e.g., review-based QA, attribute value detection), and others (e.g., E-commerce named entity recognization). To ensure efficient testing, 500 instances of each task were randomly selected as test data, resulting in a final test set of 6,000 data instances. The remaining 122 datasets were allocated for training, from which up to 800 data instances were sampled for each dataset as the training set. Ultimately, the EcomGPT was trained on a total of 85,746 instances of E-commerce data. For a more detailed scaling experiments on the number of training samples for each dataset, please refer to Appendix.

Dataset	Lang.	Task	Metric
Lenove	EN	Named Entity Recognization	F1, Rouge
		Entity Span Detection	Rouge
Reddit	EN	Extractive QA	Rouge
ABSA	EN	Review Topic Classification	F1, Rouge
MEPAVE	ZH	Attribute Value Recognization	F1, Rouge
		Attribute Value Detection	Rouge
Multi-CPR	ZH	Product Select	Rouge
OpenBG [2]	ZH	Product Align	F1, Rouge
		Title Attritube Matching	F1, Rouge
		Fine-grain Product Classify	F1, Rouge
		Coarse-grain Product Classify	F1, Rouge
		Title Generate	Rouge

Table 3: The details of our evaluation datasets.

Generalization Types. Conventional supervised learning evaluates a model's capacity to generalize within a given distribution, wherein the model learns from labeled instances of specific domains and tasks, and is subsequently tested on data that conforms to the same distribution for the same domain and task. In contrast, for E-commerce LLM, our emphasis lies in the model's ability to generalize to data outside the distribution. In this study, we correspond a data instance to three levels, namely task paradigm (e.g., generation task, classification task), task (e.g., the classification paradigm comprises tasks with different objectives like product item classification, intent detection, etc.), and dataset (e.g., for the intent detection task, it encompasses SGD (Rastogi et al. 2020) and JDDC (Chen et al. 2020) datasets, consisting of distinct label sets). The model's ability to generalize to unseen tasks/datasets at the task and dataset levels represents the most desirable and practical feature. Therefore, we primarily focus on the model's generalization capability on unseen tasks/datasets in the main experiment. Additionally, in Appendix, we evaluate the model's performance under cross task paradigms and cross-language settings.

Main Experiments

Table 4 presents the results of the automated metrics-based evaluation conducted on new datasets and tasks, from which we can conclude that: (1) In terms of average performance on unseen datasets, EcomGPT, even with the lowest number of parameters (560 million), outperforms ChatGPT, which has over 100 billion parameters (exceeding EcomGPT by 100,000 times). Moreover, EcomGPT's performance consis-

Model Type	Model	Unseen Dataset			Unseen Task						
		Mi-F1	Ma-F1	Rouge	Poduct Align			Review Topic Classify			Product Select
					Mi-F1	Ma-F1	Rouge	Mi-F1	Ma-F1	Rouge	Rouge
PLM	**BLOOM (560m)**	3.33	2.10	5.64	0.17	0.15	6.76	13.22	10.96	1.26	6.06
	BLOOM (1b7)	4.15	2.78	6.00	0.10	0.10	1.60	16.17	14.95	6.72	6.72
	BLOOM (3b)	2.94	1.43	7.89	0.10	0.20	1.86	0.38	0.18	5.50	7.99
	BLOOM(7b1)	4.29	2.50	7.31	0.10	0.13	0.97	7.11	3.61	4.96	9.47
Instruction	**BLOOMZ (560m)**	24.62	25.60	24.03	21.80	21.80	55.53	30.49	32.13	23.60	0.00
	BLOOMZ (1b7)	18.60	18.87	15.10	0.40	0.40	0.40	32.06	34.01	26.38	2.27
	BLOOMZ (3b)	29.80	30.05	26.38	10.42	10.80	16.53	30.81	32.14	23.25	11.65
	BLOOMZ (7b1)	26.75	27.07	25.21	6.00	6.00	8.00	49.37	50.39	41.43	15.14
	ChatGPT	37.30	40.71	43.92	41.60	41.60	71.02	51.22	51.80	42.55	27.39
Ours	**EcomGPT (560m)**	41.28	38.21	48.88	50.15	50.15	81.41	42.39	50.88	32.25	10.74
	EcomGPT (1b7)	42.30	39.07	53.24	51.20	52.20	81.23	47.38	52.68	37.81	32.38
	EcomGPT (3b)	48.37	45.04	59.20	53.20	53.20	82.13	53.91	56.12	44.99	52.53
	EcomGPT (7b1)	52.89	50.17	62.83	55.20	55.20	84.67	59.03	60.74	50.25	56.39
Upper-bound(est.)	**SFT(7b1)**	74.73	71.01	73.87	67.90	67.90	89.06	85.86	89.22	82.96	97.60

Table 4: Performance on unseen dataset and tasks.

tently improves as the model parameters scale, demonstrating its remarkable generalization ability for E-commerce tasks. (2) By training on EcomInstruct data, EcomGPT achieved a substantial improvement of over 20 points compared to the baseline model BLOOMZ. This suggests that excellent generalization performance of EcomGPT is not solely dependent on the backbone model. (3) Due to the lack of dialogue capability, the pre-trained language model BLOOM demonstrates poor performance, approaching 0 and being unstable. Interestingly, the difference between the performance boost achieved by the xP3 dataset, which contains over 78 million general instruction data, and that obtained by the EcomInstuct dataset, which has roughly 200,000 E-commerce instruction data for training, is approximately 4 points. This highlights the more effective role of domain-specific instruction data for vertical scenarios in enhancing model generalization capability. (4) We conducted supervised fine-tuning of BLOOMZ 7b using the training set of the test tasks to estimate the upper bound of the model's generalization performance. Our findings indicate that the current EcomGPT still has significant room for improvement in terms of generalization capability.

Furthermore, in order to enhance the reliability of the evaluation, particularly for the generation tasks, where automated evaluation metrics fall short in reflecting the performance of the model, a human evaluation was deliberately incorporated. As illustrated in Figure 3, we randomly selected 100 samples per task and ask the annotators to judge which one of the outputs of EcomGPT and ChatGPT is better or tied. The results show that, with the exception of generation tasks, the winning or tying rate of EcomGPT in the human evaluation maintains the same overall trend as the Rouge value. The Pearson coefficient between the two is 0.2, indicating a positive correlation overall and confirming the reliability of the human evaluation. Upon analyzing the output, we observed that for certain tasks with complex input or output formats, such as named entity recognition, ChatGPT struggled and often displayed a meaningless response like "sorry, I can't retrieve the information". In the case of generation tasks, such as product title generation, ChatGPT typically generated excessively long sentences, which were inconsistent with the concise and attention-grabbing style of human written titles. While ChatGPT was able to solve some relatively simple tasks, such as product selection (with a solution rate of 78% in human evaluation), the model's Rouge value remained low. We attributed this to the abundance of redundant replies in the output of ChatGPT, which hindered its practical application, since time-consuming task-specific parsing of model output is required. In conclusion, EcomGPT exhibited superior semantic understanding of E-commerce data.

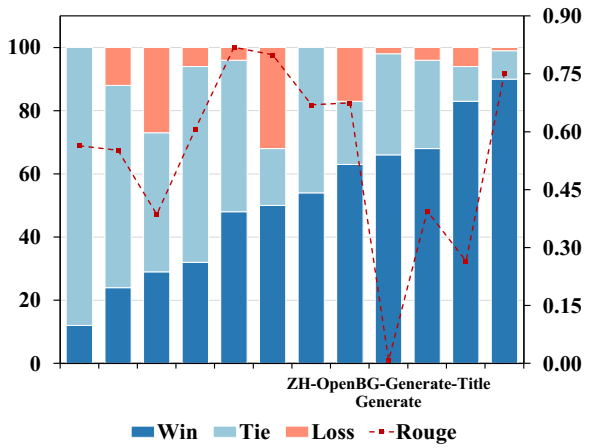

Figure 3: Human Evaluation results.

Ablation Experiments on CoT Tasks

As described in Dataset Section, a considerable proportion of EcomInstruct consists of atomic tasks that are constructed using data specific to the E-commerce domain. These atomic tasks encompass a variety of generic semantic understanding capabilities, which are extensively utilized during the intermediate stage of the model's solution of the original task. Drawing a parallel with prior research on Chain-of-Thought (Wei et al. 2022; Wang et al. 2022a), we refer to these atomic tasks as Chain-of-Task tasks (CoT tasks). The CoT task empowers the model to imbibe generic capabilities that are implicitly utilized while handling E-commerce tasks, thereby playing a pivotal role in enhancing the model's generalization ability. To validate our assumptions and the effectiveness of the CoT task, we conduct ablation experiments on the CoT task at a high level. Furthermore, we take a deeper look into the benefits of CoT tasks across varied dimensions, including data, tasks, and task paradigms.

Overall Gain from CoT Tasks The CoT tasks were derived from a combination of two sources: data with pseudo-labels generated by ChatGPT and high-quality raw data with golden labels. As illustrated in Table 5, when both components of the CoT data are sequentially removed, there is a significant degradation in the performance of the EcomGPT. Furthermore, the model trained solely using original E-commerce data fails to outperform ChatGPT's performance in Table 4. This observation suggests that solely relying on domain data for instruction learning is insufficient to enhance the generalization ability of the pendant domain model. Additionally, we observe a more substantial drop in performance upon removal of the CoT task constructed from high-quality data containing golden labels, which is due to the fact that the amount of data built from ChatGPT is relatively small while containing some errors or noise.

The significant improvement achieved with the CoT task inspires us to even with limited domain data, a series of atomic tasks constructed from the domain data can endow the model with superior generalization capabilities.

Training Dataset	Micro F1	Macro F1	Rouge
Full	48.37	45.04	59.20
w/o pseudo label CoT	44.98	41.79	55.46
w/o golden label CoT	26.64	23.64	35.02

Table 5: Overall abaltion on CoT Tasks. w/o pseudo label CoT means without CoT task whose label is generate by ChatGPT. w/o golden label CoT represents without CoT task whose label is inferred from the original golden labels.

Cross Gain from CoT Tasks In this section, we conduct extensive ablation experiments on CoT data, aiming to investigate the benefits of CoT data at the dataset, task, and task paradigm levels.

Dataset Level. In the Table 6, we remove the CoT task associated with a specific dataset from the training set to observe its impact. To prevent data leakage, we avoided introducing CoT tasks corresponding to the test dataset in the training set of EcomInstruct. So at the dataset level, we performed held-in evaluation, i.e., evaluating the selected tasks in the training set. Our findings indicate that CoT tasks derived from the same dataset provided steady gains for the original task. However, in cross-dataset scenarios, the efficacy of CoT tasks is dependent on the data types corresponding to the two datasets: for the same type of data that overlap in the task chain, the CoT tasks can provide a collaborative effect. For instance, the Ecom and Youku datasets both contain product titles, resulting in mutual gains. Conversely, there is no gain between CCKS and JDDC datasets, as their data types are addresses and dialogues, respectively, despite belonging to the same classification task.

Training	Ecom	Youku	Amazon	CCKS	JDDC	Avg
Full	73.79	91.42	61.31	70.40	31.80	65.74
w/o Ecom-R	72.77	90.67	62.55	74.00	38.20	67.64
w/o Youku-R	73.10	91.07	59.67	76.00	36.20	67.21
w/o Amazon-R	73.85	90.55	60.63	72.00	26.20	64.65
w/o CCKS-R	74.47	91.30	59.90	69.60	37.20	66.49
w/o JDDC-R	73.73	91.19	58.13	71.20	27.80	64.41

Table 6: Ablation experiments on CoT tasks at dataset level. "w/o *-R" denotes without CoT data that is related to the "*".

Training	QA	NER	IC	Unseen Dataset		
				Mi F1	Ma F1	Rouge
Full	59.23	80.67	65.30	48.37	45.05	59.20
w/o QA-R	56.75	79.78	61.55	40.18	37.89	52.37
w/o NER-R	59.00	77.55	63.30	45.50	43.97	55.14
w/o IC-R	57.54	80.49	60.40	41.12	36.94	52.28

Table 7: Ablation experiments on CoT tasks at task level.

Task Level. In the Table 7, we eliminate all CoT tasks associated with a given task and report the model's performance on unseen tasks and data. For example, for NER task, we exclude all entity detection and entity classification tasks from the training set. Our results demonstrate that CoT tasks are advantageous for both similar and dissimilar tasks. Notably, CoT tasks related to QA exhibit the greatest enhancement in generalization capacity to other tasks, while concurrently exhibiting greater difficulty in generalizing from CoT tasks from other tasks, which aligns with the finding in prior work (Zhou et al. 2022). We argue that, for instruction-following LLMs, tasks can be naturally abstracted to QA tasks, thereby playing a crucial role in enhancing generalization ability.

Task Paradigm Level. As demonstrated in Table 6, certain CoT tasks of classification do not exhibit advantage over held-in tasks of other paradigms at the dataset level. However, as shown in Table 8, when viewed from a higher-level perspective of task paradigms, there is greater overlap among data or task formats. Consequently, CoT tasks from different paradigms display a consistent gain for each other, with the CoT tasks of classification even exhibiting a greater gain over other paradigm tasks than on its own.

Training	CLS	Ext	Other	Unseen Dataset		
				Mi F1	Ma F1	Rouge
Full	67.87	52.17	80.67	48.37	45.05	59.20
w/o CLS-R	65.69	47.49	80.38	46.87	43.75	57.00
w/o Ext-R	58.71	27.47	79.14	43.73	42.81	47.36
w/o Gen-R	56.67	50.87	80.58	41.38	40.20	54.39

Table 8: Ablation on CoT tasks at task paradigm level.

Conclusion

This paper presents EcomInstruct, the first instruction-tuning dataset tailored for the E-commerce domain, encompassing two different part of instruction data, while the second part comprises atomic tasks based on the basic data types in the E-commerce domain, also known as Chain-of Task (CoT) tasks. These CoT tasks are intermediate tasks implicitly involved in solving a targeted final task. Benefiting from the fundamental semantic understanding capabilities acquired from the Chain-of-Task tasks, EcomGPT , trained with EcomInstruct, outperforms ChatGPT in term of cross-dataset/task generalization on E-commerce tasks. The advantages of leveraging CoT tasks suggest that, within vertical domain scenarios, we can devise diverse atomic tasks specifically tailored to the domain data to enhance the model's generalization ability.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No.62276154), Research Center for Computer Network (Shenzhen) Ministry of Education, the Natural Science Foundation of Guangdong Province (Grant No. 2023A1515012914), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033 and JSGG20210802154402007), the Major Key Project of PCL for Experiments and Applications (PCL2021A06), Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (HW2021008), and Shenzhen Science and Technology Program (WDZC20231128091437002).

References

Ahmadvand, A.; Kallumadi, S.; Javed, F.; and Agichtein, E. 2020. Jointmap: joint query intent understanding for modeling intent hierarchies in e-commerce search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1509–1512.

Celikyilmaz, A.; Clark, E.; and Gao, J. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Chen, M.; Liu, R.; Shen, L.; Yuan, S.; Zhou, J.; Wu, Y.; He, X.; and Zhou, B. 2020. The JDDC Corpus: A Large-Scale Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 459–466.

Cheng, X.; Bowden, M.; Bhange, B. R.; Goyal, P.; Packer, T.; and Javed, F. 2021. An end-to-end solution for named entity recognition in ecommerce search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15098–15106.

Cheng, X.; Dong, Q.; Yue, F.; Ko, T.; Wang, M.; and Zou, Y. 2023a. M 3 st: Mix at three levels for speech translation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Cheng, X.; Xu, W.; Zhu, Z.; Li, H.; and Zou, Y. 2023b. Towards spoken language understanding via multi-level multi-grained contrastive learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 326–336.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Escursell, S.; Llorach-Massana, P.; and Roncero, M. B. 2021. Sustainability in e-commerce packaging: A review. *Journal of cleaner production*, 280: 124314.

Jiang, H.; Cao, T.; Li, Z.; Luo, C.; Tang, X.; Yin, Q.; Zhang, D.; Goutam, R.; and Yin, B. 2022. Short Text Pre-training with Extended Token Classification for E-commerce Query Understanding. *CoRR*, abs/2210.03915.

Li, Y.; Chen, J.; Li, Y.; Yu, T.; Chen, X.; and Zheng, H.-T. 2022. Embracing Ambiguity: Improving Similarity-oriented Tasks with Contextual Synonym Knowledge. CoRR abs/2211.10997 (2022).

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

Loshchilov, I.; and Hutter, F. 2017. Fixing weight decay regularization in adam. arXiv 2017. *arXiv preprint arXiv:1711.05101*.

Mishra, S.; Khashabi, D.; Baral, C.; and Hajishirzi, H. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, 3470–3487. Association for Computational Linguistics (ACL).

Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Scao, T. L.; Bari, M. S.; Shen, S.; Yong, Z.-X.; Schoelkopf, H.; et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Poerner, N.; Waltinger, U.; and Schütze, H. 2020. E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 803–818. Online: Association for Computational Linguistics.

Qiu, Y.; Zhao, C.; Zhang, H.; Zhuo, J.; Li, T.; Zhang, X.; Wang, S.; Xu, S.; Long, B.; and Yang, W.-Y. 2022. Pre-training Tasks for User Intent Detection and Embedding Retrieval in E-commerce Search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4424–4428.

Rastogi, A.; Zang, X.; Sunkara, S.; Gupta, R.; and Khaitan, P. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 8689–8696.

Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; Gallé, M.; et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford alpaca: An instruction-following llama model.

Tsagkias, M.; King, T. H.; Kallumadi, S.; Murdock, V.; and de Rijke, M. 2021. Challenges and research opportunities in ecommerce search and recommendations. In *ACM Sigir Forum*, volume 54, 1–23. ACM New York, NY, USA.

Wang, X.; Jiang, Y.; Bach, N.; Wang, T.; Huang, Z.; Huang, F.; and Tu, K. 2021. Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1800–1812.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022a. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wang, Y.; Mishra, S.; Alipoormolabashi, P.; Kordi, Y.; Mirzaei, A.; Naik, A.; Ashok, A.; Dhanasekaran, A. S.; Arunkumar, A.; Stap, D.; et al. 2022b. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5085–5109.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Xu, S.; Li, H.; Yuan, P.; Wang, Y.; Wu, Y.; He, X.; Liu, Y.; and Zhou, B. 2021. K-PLUG: Knowledge-injected Pre-trained Language Model for Natural Language Understanding and Generation in E-Commerce. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1–17.

Yu, T.; Jiang, C.; Lou, C.; Huang, S.; Wang, X.; Liu, W.; Cai, J.; Li, Y.; Li, Y.; Tu, K.; Zheng, H.-T.; Zhang, N.; Xie, P.; Huang, F.; and Jiang, Y. 2023. SeqGPT: An Out-of-the-box Large Language Model for Open Domain Sequence Understanding. arXiv:2308.10529.

Zhang, W.; Wong, C.-M.; Ye, G.; Wen, B.; Zhang, W.; and Chen, H. 2021. Billion-scale pre-trained e-commerce product knowledge graph model. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2476–2487. IEEE.

Zhao, J.; Chen, H.; and Yin, D. 2019. A dynamic product-aware learning model for e-commerce query intent understanding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1843–1852.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhou, J.; Lin, Z.; Zheng, Y.; Li, J.; and Yang, Z. 2022. Not All Tasks Are Born Equal: Understanding Zero-Shot Generalization. In *The Eleventh International Conference on Learning Representations*.