eCeLLM: Generalizing Large Language Models for E-commerce from Large-scale, High-quality Instruction Data

Bo Peng [*1] **Xinyi Ling** [*1] **Ziru Chen** [1] **Huan Sun** [1 2] **Xia Ning** [1 2 3]

Abstract

With tremendous efforts in developing effective e-commerce models, conventional e-commerce models show limited success in generalist e-commerce modeling, and suffer from unsatisfactory performance on new users and new products – a typical out-of-domain generalization challenge. Meanwhile, large language models (LLMs) demonstrate outstanding performance in generalist modeling and out-of-domain generalizability in many fields. Toward fully unleashing their power for e-commerce, in this paper, we construct ECInstruct, the first open-sourced, large-scale, and high-quality benchmark instruction dataset for e-commerce. Leveraging ECInstruct, we develop eCeLLM, a series of e-commerce LLMs, by instruction-tuning general-purpose LLMs. Our comprehensive experiments and evaluation demonstrate that eCeLLM models substantially outperform baseline models, including the most advanced GPT-4, and the state-of-the-art task-specific models in in-domain evaluation. Moreover, eCeLLM exhibits excellent generalizability to out-of-domain settings, including unseen products and unseen instructions, highlighting its superiority as a generalist e-commerce model. Both the ECInstruct dataset and the eCeLLM models show great potential in empowering versatile and effective LLMs for e-commerce. ECInstruct and eCeLLM models are publicly accessible through https://ninglab.github.io/eCeLLM/.

1. Introduction

The Internet's evolution and the rise of the digital economy have made e-commerce an integral part of daily life, drawing considerable attention from researchers. In recent years, tremendous research efforts have been dedicated to developing effective e-commerce models (Yang et al., 2022; Geng et al., 2022). Though promising, conventional e-commerce models generally suffer from two issues. **(1)** Limited success in generalist e-commerce modeling (Zhang et al., 2023b): Conventional e-commerce models are typically task-specific (e.g., gSASRec for sequential recommendation (Petrov & Macdonald, 2023), SUOpenTag for attribute value extraction (Xu et al., 2019)). However, contemporary e-commerce platforms, such as Amazon and eBay, are highly complex and consistently expanding, with many interdependent tasks. In this context, generalist modeling is particularly suitable and desired on these platforms due to its cost-effectiveness and extensibility, while the task-specific modeling scheme struggles with scalability. **(2)** Unsatisfactory performance on new users and new products (Hou et al., 2024): Cold start (i.e., performing e-commerce tasks on new users or new products) (Lika et al., 2014), the typical out-of-domain (OOD) generalization challenge in e-commerce, has been a long-standing and difficult problem (Ding et al., 2022; Yang et al., 2022). Existing task-specific e-commerce models are typically tailored to existing users and products and lack the ability to effectively extrapolate to new users and new products (Lika et al., 2014). However, new users and new products are very common and highly wanted in the dynamic landscape of e-commerce.

Recently, large language models (LLMs), such as GPT-4 (OpenAI, 2023), Claude 2 (Anthropic, 2023), Gemini (Team et al., 2023), and Llama 2 (Touvron et al., 2023b), have demonstrated exceptional performance in natural language processing (Zhao et al., 2023), information retrieval (Spatharioti et al., 2023), and many other fields (Frieder et al., 2023; Zeng et al., 2023). However, their power for e-commerce is not fully unleashed. While limited efforts (Yang et al., 2023; Zhang et al., 2023a; Li et al., 2024) have been dedicated to leveraging LLMs for e-commerce, most of them focus on studying the utility of pre-trained LLMs in single or homogenous tasks (e.g., e-commerce authoring). Some recent efforts employ instruction tuning to adapt LLMs for e-commerce applications (Shi et al., 2023; Li et al., 2024; Geng et al., 2022). However, they fall short in the covered e-commerce tasks

[*]Equal contribution [1]Department of Computer Science and Engineering, The Ohio State University, USA. [2] Translational Data Analytics Institute, The Ohio State University, USA. [3]Department of Biomedical Informatics, The Ohio State University, USA. Correspondence to: Xia Ning <ning.104@osu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

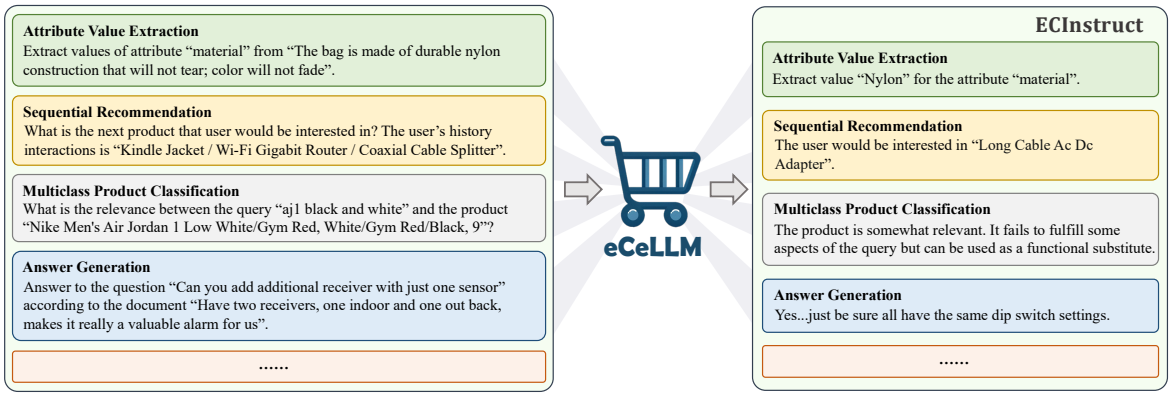


Figure 1. Overall scheme of eCeLLM instruction-tuned with ECInstruct

and instruction data. We aim to bridge the gap and develop e-commerce foundation models with real-world utilities for a large variety of e-commerce applications.

To this end, we construct ECInstruct, an open-sourced, large-scale, and high-quality benchmark instruction dataset tailored for developing and evaluating LLMs in e-commerce realm. ECInstruct covers 116,528 samples from 10 real and widely performed e-commerce tasks of 4 categories. Each data sample comprises an instruction, an input, and an output. Some samples also incorporate a list of options. All the 10 tasks have in-domain (IND) test samples, and 6 tasks also have OOD test samples which consist of products unseen in the training samples of the respective tasks. ECInstruct undergoes rigorous and thorough scrutiny and is carefully crafted to enable a wide spectrum of empirical testing and exploration, including IND evaluation, OOD evaluation, and task-specific studies.

Leveraging ECInstruct, we develop a series of e-Commerce LLMs, denoted as eCeLLM (pronounce: e-sell ˈem, /ˌɪ-ˈseləm/) models, by instruction-tuning 6 general-purpose LLMs, such as Llama 2 (Touvron et al., 2023b) and Mistral (Jiang et al., 2023). The eCeLLM models are extensively evaluated across various settings, including IND and OOD data, unseen instructions, and different training sample sizes of ECInstruct. Overall, our experimental results demonstrate the following findings:

(1) eCeLLM models substantially outperform baseline models, including the most advanced GPT-4 (OpenAI, 2023) and the state-of-the-art (SoTA) task-specific models, on almost all the 10 tasks in IND evaluation. On average, eCeLLM models show a substantial improvement of 10.7% over the best baseline models.

(2) Moreover, eCeLLM exhibits excellent generalizability to OOD settings, including unseen products and unseen instructions, highlighting its superiority as a generalist e-commerce model. Particularly, eCeLLM models establish an improvement of 9.3% over the best baselines on the OOD (new) products. These results indicate the great potential of both the ECInstruct dataset and the eCeLLM models in empowering versatile and effective LLMs for e-commerce, and validate the potential of LLMs in doing e-commerce tasks.

Figure 1 depicts the overall scheme of eCeLLM. To the best of our knowledge, this study is the first comprehensive and systematic study of instruction-tuning LLMs for e-commerce applications, open-sources the first large-scale, high-quality benchmark dataset (ECInstruct), and develops the state-of-the-art generalist LLMs (eCeLLM series) for e-commerce. ECInstruct and eCeLLM models are publicly accessible through https://ninglab.github.io/eCeLLM/.

2. Related Work

Instruction Tuning with LLMs Instruction tuning enables the transfer of general knowledge captured in LLMs to specific application domains, facilitating the generalizability of LLMs. FLAN (Wei et al., 2021) and T0 (Sanh et al., 2021) are the early work to investigate instruction tuning by evaluating the zero-shot performance of fine-tuned LLMs on numerous datasets. Both of the models show encouraging performance over the original GPT-3 (Brown et al., 2020), indicating the effectiveness of instruction tuning. The importance of datasets, model scale, and instructions for instruction tuning are explored in a later work FLAN-v2 (Chung et al., 2024). Considering the significant impact of datasets for instruction tuning, several studies collect sizable benchmarks consisting of numerous tasks, such as Super-NaturalInstructions (Wang et al., 2022b), Self-instruct (Wang et al., 2022a) and Flan Collection (Longpre et al., 2023), for fine-tuning general-purpose LLMs on NLP tasks. Meanwhile, instruction-tuned LLMs are employed in specific domains. For example, Platypus (Lee et al., 2023a) is fine-tuned for reasoning in STEM, and Mammoth (Yue et al., 2024) is for general math problem-solving. Our eCeLLM incorporates instruction tuning with LLMs for e-commerce.

LLMs for E-commerce LLMs are emerging in the general e-commerce realm. RecMind (Wang et al., 2023)

Table 1. Comparison among E-commerce LLMs

Comparison	LLaMA-E (Shi et al., 2023)	EcomGPT (Li et al., 2024)	eCeLLM (ours)
Instruction tuning	✓	✓	✓
Open-sourced data	✗	✗	✓
Real-world tasks	✓	✗	✓
# training tasks	5	122‡	10
# In-domain tests	5	0	10
# out-of-domain tests	0	12‡	6
# general-purpose LLMs evaluated	5	3	5
# task-specific SoTA models evaluated	0	0	11
# base LLMs tuned	3	4	6
Open-sourced models	✗	✓	✓

‡EcomGPT has 122 training tasks, most of which are manipulated from data of different other tasks. The training data is not publicly available. EcomGPT releases 12 test tasks (8 in Chinese) for only out-of-domain evaluation.

utilizes LLMs as an autonomous recommender agent by enhancing their capability to comprehend user behaviors. LLaMA-E (Shi et al., 2023) is fine-tuned on LLaMA (Touvron et al., 2023a) for various e-commerce authoring tasks, such as ads generation and general product QA. EcomGPT (Li et al., 2024) is a pioneer instruction-tuned LLM on its dataset EcomInstruct. However, most of its tasks are human-constructed by repurposing data from other tasks (e.g., from the "query product matching" task, a task is constructed to generate user queries based on the matched products), and thus, could have limited utility in real applications. A recent study (Geng et al., 2022) develops a fine-tuned LLM, P5, with personalized prompts to conduct a few specific recommendation tasks (e.g., sequential recommendation). These studies demonstrate the utility of LLMs and instruction tuning in e-commerce applications. To the best of our knowledge, eCeLLM is the most comprehensive instruction-tuned e-commerce LLM.

Comparison among Instruction-tuned LLMs for E-commerce Table 1 summarizes the difference between our eCeLLM from LLaMA-E and EcomGPT, the two existing fine-tuned LLMs capable of performing multiple e-commerce tasks (P5 only performs recommendations). Fundamentally different from LLaMA-E and EcomGPT, eCeLLM focuses on comprehensive real-world e-commerce tasks and data, includes very extensive evaluation and benchmarking, and open-sources instruction data and models.

3. ECInstruct Dataset

We introduce ECInstruct, an instruction dataset for adapting LLMs to e-commerce tasks. ECInstruct features 3 key design principles. **(1) Broad coverage**: ECInstruct includes 10 diverse tasks of 4 categories. Comprehensive and wide-ranging related tasks are critical in enabling versatile LLMs for a specific domain as shown in the literature (Lee et al., 2023b). **(2) Realistic tasks**: ECInstruct focuses on real-world e-commerce tasks with real-world data, not human-manipulated tasks with synthetic data. This ensures e-commerce LLMs tuned on ECInstruct a potentially high utility in real-world applications. **(3) High quality**: ECInstruct considers only real-world data and undergoes rigorous scrutiny to ensure its accuracy and high quality. High data quality plays a pivotal role in building effective LLMs (Hoffmann et al., 2022; Gadre et al., 2024).

3.1. E-commerce Tasks

ECInstruct comprises 10 real-world tasks with real-world data that are widely performed in e-commerce applications. These tasks fall within 4 categories: **(i)** product understanding, **(ii)** user understanding, **(iii)** query product matching, and **(iv)** product question answering. Such tasks are ubiquitous and essential on e-commerce platforms; success on these tasks would enable key functionalities in providing excellent user experience, promoting online retail, and driving sustainable revenues, among others.

Particularly, ECInstruct includes 3 tasks for product understanding: **(1)** attribute value extraction (AVE) (Xu et al., 2019; Wang et al., 2020; Yang et al., 2022), **(2)** product matching (PM) (Köpcke et al., 2010; Rahm, Erhard, 2010) and **(3)** product relation prediction (PRP) (Ahmed et al., 2021; Xu et al., 2020). For user understanding, ECInstruct includes **(4)** sentiment analysis (SA) (Wankhade et al., 2022) and **(5)** sequential recommendation (SR) (Petrov & Macdonald, 2023; Li et al., 2023). ECInstruct also covers 3 query product matching tasks (Reddy et al., 2022): **(6)** multi-class product classification (MPC), **(7)** product substitute identification (PSI), and **(8)** query-product ranking (QPR). For product question answering, ECInstruct contains the tasks of **(9)** answerability prediction (AP) (Gupta et al., 2019) and **(10)** answer generation (AG) (Deng et al., 2023). More details are available in Appendix A. Table A1 summarizes the tasks and their data sources. Fundamentally different from the instruction data of LLaMA-E (Shi et al., 2023) and EcomGPT (Li et al., 2024), ECInstruct contains tasks all from real e-commerce platforms, and data all extracted from these real tasks.

3.2. Diverse Instructions

Recent work (Xu et al., 2022) shows that diverse instructions for LLM tuning improve LLM generalizability to new instructions. Inspired by this, we construct diverse instructions in ECInstruct. For each task, we start with a clear and concise human-written seed instruction. We then generate diverse instructions synonymous with the seed instruction via GPT-4 (OpenAI, 2023), and select into ECInstruct 5

of those with writing styles (e.g., wording) distinct from the seed instruction. Thus, ECInstruct involves 6 high-quality and diverse instructions, including the seed instruction, on each task.

To test instruction understanding and model generalizability to new instructions, we hold out one instruction for each task as its "unseen" instruction, which is inaccessible during model training. Thus, each task has 5 diverse instructions for instruction tuning, and 1 unseen instruction for testing. All the instructions are presented in Appendix C.

3.3. Data Quality Control

In ECInstruct, we carry out the following procedures to ensure its accuracy and high quality. Specifically, we **(1)** remove overlapping data between training and test sets to avoid data leakage; **(2)** retain only data in English to ensure the unity of languages in texts; **(3)** eliminate non-English notations such as HTML tags and Unicode; **(4)** only select products with detailed information to allow sufficient product knowledge that LLMs can learn from; **(5)** keep texts within a reasonable length following the convention in the literature (Hou et al., 2022); **(6)** manually inspect all processed data. We also conduct task-specific quality control on individual tasks. All the quality control procedures are detailed in Appendix A.

3.4. Training, Validation, and Test Sets

ECInstruct is split into training sets, validation sets, IND test sets, and OOD test sets as follows. Table A3 summarizes the ECInstruct dataset. Appendix A describes the details of the data split process.

Training Set ECInstruct has a training set of 10K for each individual task (except for PM, which has 2,022 training samples). The training sets from all the tasks are combined into a dataset of 92,022 samples as the training set of ECInstruct.

Validation Set ECInstruct has a validation set of 1K for each individual task (except for PM, which has 253 validation samples). Similarly, the validation sets of each task are combined into a dataset of 9,253 samples as the validation set of ECInstruct.

In-Domain Test Set For each of the 10 tasks, ECInstruct also includes an in-domain (IND) test set of 1K samples (except for PM, which has 253 IND samples). We define IND in terms of products that are within the same category as the training products.

Out-of-Domain Test Set To evaluate the generalizability of LLMs tuned on ECInstruct to unseen samples – a critical capacity desired to address the cold-start problem in e-commerce, we create OOD test sets in ECInstruct. We define OOD in terms of products, that is, new products unseen during LLM training are considered as OOD. We determine OOD products using their category information. In ECInstruct, 6 tasks – AVE, PRP, SA, SR, AP, and AG, have different product categories. For each of these tasks, all products from one particular category are held out as the OOD data for that task, as summarized in Table A3 (each product belongs to only one category). We focus on OOD (new) products, because promoting and recommending new products to users are effective strategies in e-commerce in driving sales and revenue, enhancing user experience, and increasing user engagement. Note we do not define OOD test sets over users (i.e., hold out users into OOD test sets), because users are anonymous and user identifiers are absent in existing datasets. Even though, the unseen instructions could be used in proximity to new users – a new user may have a completely new style of instructions to eCeLLM.

4. eCeLLM Models

We build a series of eCeLLM models on top of 6 base models: **(1)** eCeLLM-L: trained from large base models, Flan-T5 XXL (11B parameters) (Chung et al., 2024) and Llama-2 13B-chat (Touvron et al., 2023b), **(2)** eCeLLM-M: trained from medium-sized base models, Llama-2 7B-chat (Touvron et al., 2023b) and Mistral-7B Instruct-v0.2 (Jiang et al., 2023), and **(3)** eCeLLM-S: trained from small base models, Flan-T5 XL (3B) (Chung et al., 2024) and Phi-2 (3B) (Javaheripi & Bubeck, 2023). These eCeLLM models are instruction-tuned over ECInstruct training data.

We fine-tune all the base models with LoRA (Hu et al., 2021) and Huggingface transformers library (Wolf et al., 2019). During fine-tuning, the learning rate and batch size of all the models are set as 1e-4 and 128, respectively. We apply a cosine learning rate scheduler with a 5% warm-up period for 3 epochs. We set α and the rank in LoRA as 16, and add LoRA adaptors to all the projection layers and the language modeling head. We perform 0-shot evaluations (i.e., without in-context examples) on all the tasks.

5. Experimental Setup

We compare eCeLLM against 3 categories of baseline models: **(1)** general-purpose LLMs, **(2)** e-commerce LLMs, and **(3)** SoTA task-specific models. Table 2 lists all the baseline models. Note that for the task-specific models, we only use the best (i.e., SoTA) models for each individual task based on literature, so as to enable schematic comparison between generalist modeling from eCeLLM and the task-specific modeling. We conduct IND and OOD tests on respective test datasets (Section 3.4) for all the models. Table A3 lists the evaluations conducted for different tasks. Details on using/-training baseline models are in Appendix D. The prompt

Table 2. Summary of Baseline Models

General-purpose LLMs	**(1)** GPT-4 Turbo (OpenAI, 2023), **(2)** Gemini Pro (Team et al., 2023), **(3)** Claude 2.1 (Anthropic, 2023), **(4)** Llama-2 13B-chat (Touvron et al., 2023b), **(5)** Mistral-7B Instruct-v0.2 (Jiang et al., 2023)
E-commerce LLM	**(1)** EcomGPT (Li et al., 2024): To the best of our knowledge, it is the only open-source e-commerce LLM.
SoTA Task-specific Models	
`AVE` (Section A.1.1)	**(1)** SUOpenTag (Xu et al., 2019), **(2)**AVEQA (Wang et al., 2020): Both methods scan the product information, token by token, to extract the values of the specified attributes.
`PRP` (Section A.1.2)	**(1)** RGCN (Schlichtkrull et al., 2018): It utilizes a graph convolutional network to capture relations between products. **(2)** DeBERTaV3 (He et al., 2022): It predicts product relations from product titles.
`SA` (Section A.2.1)	**(1)** BERTweet (Nguyen et al., 2020): It is a pre-trained language model for English tweets, and achieves superior performance over RoBERTa (Liu et al., 2019) and XLM-R (Conneau et al., 2019) on `SA`. **(2)** P5 (Geng et al., 2022): It is a pre-trained LLM for `SA` from e-commerce user reviews.
`SR` (Section A.2.2)	**(1)** gSASRec (Petrov & Macdonald, 2023), **(2)** Recformer (Li et al., 2023): Both methods leverage Transformer (Vaswani et al., 2017) to predict the next product of users' interest based on users' historical activities on products.
`AG` (Section A.4.2)	**(1)** GPT-4 Turbo (OpenAI, 2023): To the best of our knowledge, there are no models specifically designed for `AG`. Thus, GPT-4 Turbo is used as the SoTA task-specific model in this task as it has outstanding performance in general question answering (OpenAI, 2023).
`PM`, `MPC`, `PSI`, `QPR`, `AP` (Section A.1.3, A.3.1, A.3.2, A.3.3, A.4.1).	**(1)** BERT (Devlin et al., 2019), **(2)** DeBERTaV3 (He et al., 2022): Both methods generate predictions from the textual product information (e.g., titles).

templates used in the evaluations are in Appendix E.

General-purpose LLMs For all the general-purpose LLMs, we use the checkpoints released by their developers. For GPT-4 Turbo, Gemini Pro, and Claude 2.1, we access the checkpoints using their official APIs. For Llama-2 13B-chat and Mistral-7B Instruct-v0.2, we use the checkpoints released in Huggingface (Wolf et al., 2019).

We perform 1-shot evaluations on all the general-purpose LLMs. Existing work (Li et al., 2024) tests LLMs on e-commerce tasks in a 0-shot setting. Meanwhile, in-context examples can notably benefit LLMs (Brown et al., 2020). Thus, in our case, we conduct 1-shot evaluations, balancing the computing cost and model performance, to enable stronger performance from the general-purpose LLMs. Although extensive prompt engineering and many in-context examples could enable better performance from LLMs, it is less practical in real e-commerce applications; for example, asking users to provide many in-context examples of the e-commerce platform, if ever feasible, can reduce user engagement. Also, it is not cost- and energy-efficient to do large-scale, few-shot in-context learning on LLMs over large test sets like `ECInstruct`'s.

E-commerce LLMs For EcomGPT, we use the checkpoint released by its authors. We conduct both 0-shot and 1-shot evaluations on EcomGPT, as we empirically observe that 1-shot may result in better performance than 0-shot for EcomGPT (EcomGPT originally conducts 0-shot evaluations). Therefore, we report the best performance of the two evaluations on each task.

SoTA Task-specific Models We train the SoTA task-specific models SUOpenTag, AVEQA, RGCN, and gSASRec from scratch using `ECInstruct` training data of individual tasks and the model implementations published by their respective authors. For P5, we use the checkpoint released by its authors (it has already been pre-trained on e-commerce data). For GPT-4 Turbo, we access the checkpoint via its API. For Recformer, BERT, BERTweet, and DeBERTaV3, we tune the checkpoints on specific tasks.

6. Experimental Results

We evaluate the models using the test sets of individual tasks. For each task, we employ multiple metrics for a comprehensive evaluation (Appendix A). For the sake of clarity, in this section, we present the performance only in terms of the primary evaluation metric for each task (Table A1). As discussed in Appendix F, Llama-2 13B-chat, Mistral-7B Instruct-v0.2, and Phi-2 are the best base models for `eCeLLM-L`, `eCeLLM-M`, and `eCeLLM-S`, respectively. Thus, by default, `eCeLLM-L`, `eCeLLM-M`, and `eCeLLM-S` refer to those tuned from these base models, respectively.

Main Results Our comprehensive experiments yield the following main results: **(1)** `eCeLLM` models demonstrate the best performance on almost all the IND tasks, with a significant average improvement of 10.7% over the general-purpose LLMs, e-commerce LLMs, and the SoTA task-specific models across the 10 tasks (Section 6.1). **(2)** `eCeLLM` models show outstanding generalizability to OOD products and surpass the best baselines with a remarkable average improvement of 9.3% in OOD evaluation (Section 6.2). **(3)** By training over diverse instructions,

Table 3. Overall Performance in IND Evaluation

Model	AVE	PRP	PM	SA	SR	MPC	PSI	QPR	AP	AG
	F1*	Macro F1	F1	Macro F1	HR@1	Accuracy	F1	NDCG	F1	F_{BERT}
GPT-4 Turbo	0.495	0.326	0.753	0.516	0.387	0.611	0.195	0.875	0.649	**0.858**
Gemini Pro	0.396	0.136	0.867	0.470	0.269	0.584	0.248	0.821	0.506	0.855
Claude 2.1	0.381	0.275	0.523	0.415	0.066	0.655	0.273	0.821	0.280	0.841
Llama-2 13B-chat	0.002	0.333	0.434	0.188	0.056	0.504	0.252	0.815	0.623	0.811
Mistral-7B Instruct-v0.2	0.369	0.324	0.613	0.470	0.164	0.529	0.305	0.842	0.588	0.853
EcomGPT	0.000	0.091	0.648	0.188	0.042	0.540	0.170	0.000	0.086	0.669
SoTA task-specific model	0.546	0.588	**0.995**	0.573	0.265	**0.703**	0.389	0.859	0.830	**0.858**
eCeLLM-L	0.582	**0.611**	**0.995**	**0.648**	0.526	0.684	**0.501**	0.870	**0.851**	0.841
eCeLLM-M	**0.662**	0.558	**0.995**	0.639	**0.542**	0.696	0.305	**0.876**	0.846	0.842
eCeLLM-S	0.509	0.518	0.991	0.596	0.479	0.650	0.392	0.870	0.846	0.842
improvement (%, avg: 10.7)	21.2	3.9	0.0	13.1	40.1	-1.0	28.8	0.1	2.5	-1.9

In this table, "F1*", "Macro F1", "F1", "HR@1", "Accuracy", "NDCG" and "F_{BERT}" are the primary evaluation metrics in respective tasks (Appendix A). For each task, the best baseline performance is underlined, and the overall best performance is in **bold**. The row "improvement" presents the percentage improvement of the best-performing eCeLLM model over the best-performing baseline model (underlined) in each task. We also include the average ('avg') improvement across all the tasks in the table.

eCeLLM is equipped with strong generalizability to unseen instructions (Section 6.3). **(4)** Trained on all the tasks in ECInstruct together, eCeLLM exhibits similar or better performance than models trained on each individual task (Section 6.4). **(5)** eCeLLM models benefit from larger instruction training data for e-commerce tasks (Section 6.5).

6.1. In-domain Evaluation

Table 3 shows the performance of eCeLLM and all the baseline methods in IND evaluation, where eCeLLM models are trained using ECInstruct training set (i.e., including all the tasks), and the SoTA task-specific models are trained using the task-specific training data. Among the SoTA task-specific models, we report the results of only the best-performing model on each task. The complete results of each task are presented in Appendix G.

Overall Comparison As shown in Table 3, overall, eCeLLM substantially outperforms baseline models across the 10 e-commerce tasks at 10.7% on average. Particularly, eCeLLM models (i.e., eCeLLM-L, eCeLLM-M, and eCeLLM-S) achieve superior performance over the baselines on 7 out of the 10 tasks with an average improvement of 15.7%. On the rest 3 tasks PM, MPC, and AG, eCeLLM models achieve the same or comparable performance as the baselines (e.g., maximum difference of 1.9% as on AG). These results demonstrate the remarkable effectiveness of eCeLLM compared with the general-purpose LLMs, the SoTA task-specific models, and the existing e-commerce LLM across the e-commerce tasks.

Comparison between eCeLLM and General-purpose LLMs Table 3 shows that eCeLLM models substantially outperform the general-purpose LLMs by a remarkable margin. For example, across the 10 tasks, eCeLLM-L achieves a significant average improvement of 39.6% over GPT-4 Turbo. A key difference between eCeLLM and GPT-4 Turbo is that eCeLLM is specifically tuned on our instruction dataset ECInstruct for e-commerce. The remarkable improvement of eCeLLM over general-purpose LLMs suggests that there could be a significant gap between general knowledge and the knowledge required for e-commerce tasks, highlighting the significance of ECInstruct in imparting knowledge pertinent to e-commerce into LLMs.

We also observe that general-purpose LLMs lag behind the SoTA task-specific models by a large margin. For example, the best general-purpose LLM GPT-4 Turbo considerably underperforms the SoTA task-specific models on 7 out of the 10 tasks (e.g., 0.495 vs 0.546 on AVE). This underscores the critical need to deliberately accommodate general-purpose LLMs for e-commerce tasks.

Comparison between eCeLLM and E-commerce LLMs According to Table 3, eCeLLM models demonstrate substantial improvement over the existing e-commerce LLM EcomGPT on all the tasks. For example, eCeLLM-L surpasses EcomGPT remarkably by 244.6% on SA, and PM by 53.5%. Both eCeLLM and EcomGPT are instruction-tuned LLMs for e-commerce. However, our instruction dataset ECInstruct for eCeLLM fundamentally differs from EcomGPT's EcomInstruct dataset: ECInstruct includes only real-world data and real-world e-commerce tasks, while EcomInstruct incorporates a considerable amount of synthetic data and tasks, which hinders its applicability in real e-commerce tasks.

Comparison between eCeLLM and SoTA task-specific models Table 3 shows the SoTA task-specific models perform best among the baselines on each respective task. The substantial improvement of eCeLLM over the

SoTA task-specific models serves as strong evidence that eCeLLM, with specific tuning over ECInstruct, could leverage knowledge shared across multiple e-commerce tasks and boost performance on each individual task.

6.2. Out-of-domain Evaluation

Table 4. Overall Performance in OOD Evaluation

Model	AVE	PRP	SA	SR	AP	AG
	F1*	M-F1	M-F1	HR@1	F1	F_{BERT}
GPT-4 Turbo	0.397	0.392	0.510	0.198	0.680	**0.860**
Gemini Pro	0.275	0.123	0.454	0.116	0.552	0.856
Claude 2.1	**0.410**	0.277	0.369	0.036	0.245	0.842
Llama-2 13B-chat	0.000	0.324	0.178	0.050	0.644	0.808
Mistral-7B Instruct-v0.2	0.264	0.327	0.438	0.108	0.608	0.851
EcomGPT	0.001	0.096	0.178	0.023	0.140	0.722
SoTA task-specific model	0.269	0.507	0.567	0.081	0.853	**0.860**
eCeLLM-L	0.335	**0.558**	0.629	0.273	0.867	0.841
eCeLLM-M	0.367	0.502	**0.640**	**0.280**	0.878	0.840
eCeLLM-S	0.302	0.520	0.565	0.241	**0.879**	0.840
improvement (%, avg: 9.3)	-10.5	10.1	14.1	41.4	3.0	-2.2

In this table, "M-F1" represents macro F1. The columns in this table have the same meanings as those in Table 3.

Table 4 shows the performance of eCeLLM and baselines in OOD evaluation. Overall, we observe a similar trend to that in IND evaluation. Specifically, eCeLLM-L outperforms the SoTA task-specific models on OOD evaluation by a wide margin on 5 out of the 6 tasks, except for AG task, on which eCeLLM is comparable with the SoTA task-specific model (e.g., 2.2% difference). Similarly, eCeLLM-L outperforms the general-purpose LLMs on 4 out of the 6 tasks (i.e., PRP, SA, SR, and AP). Across the 6 tasks, eCeLLM-L demonstrates a substantial average improvement of 18.9% over GPT-4 Turbo. Compared to EcomGPT, eCeLLM-L again shows terrific advantages on all the tasks in OOD evaluation. As shown in the literature (Lika et al., 2014), cold start for new products has been an unsolved issue but a key driver in e-commerce applications. The OOD generalizability of eCeLLM to new products as demonstrated in Table 4 makes eCeLLM a highly viable tool for e-commerce applications.

Note that GPT-4 and Claude 2.1 exhibit robust performance on AVE and AG, indicating that these general-purpose LLMs are inherent with rich knowledge for solving extraction and question-answering problems. However, they may still lack comprehensive knowledge to effectively perform multiple, diverse e-commerce tasks, such as on the other 4 tasks.

6.3. eCeLLM Generalizability to Unseen Instructions

To evaluate the generalizability of eCeLLM to unseen instructions, we compare eCeLLM models tuned from diverse instructions and tuned from single instructions per task, and test their performance over unseen instructions.

As shown in Table 5, overall, eCeLLM models tuned over multiple, diverse instructions exceed or resemble those tuned over single instructions. For example, eCeLLM-L performs much better on unseen instructions on AVE and PSI when trained from diverse instructions. On the other 8 tasks, eCeLLM-L trained from diverse instructions is either better than or comparable to that from single instructions. Similarly, in the case of eCeLLM-M, using diverse instructions shows very strong effects on performance on AVE and SA, while for other tasks, performance remains comparable (except for PRP). For eCeLLM-S, diverse instructions enhance performance on almost all the tasks.

These results illustrate the pivotal role of diverse instructions for e-commerce tasks, and underscore the utility of ECInstruct in generalizing LLMs to new instructions in e-commerce applications. The results also showcase the generalizability of eCeLLM to unseen instructions. As new users may apply instructions unseen in existing e-commerce data, such generalizability also indicates the potential of eCeLLM for cold-start users in e-commerce. It is noticeable that eCeLLM is able to generalize to unseen instructions (Table 5) with a similar performance as it has on "seen" instructions (Table 3). For example, on SA, eCeLLM-L shows only a 1.4% difference in its performance in the former setting and in the latter setting (0.639 vs 0.648). This further demonstrates the strong generalizability of eCeLLM.

6.4. Generalist vs Task-specific eCeLLM Models

We compare the eCeLLM models tuned using all tasks (generalist models) with those tuned on individual tasks (task-specific models) and present the performance comparison in IND evaluation in Table 6. The results of the OOD evaluation are presented in Appendix H.3. As shown in Table 6, generalist eCeLLM demonstrates slightly stronger performance compared to task-specific eCeLLM on each task (e.g., 2.7% average improvement of generalist eCeLLM-L over all tasks except PSI). For example, on AVE, generalist eCeLLM-L is comparable to the task-specific eCeLLM-L (0.582 vs 0.599). On PRP, generalist eCeLLM-L demonstrates a substantial improvement of 17.3% compared to task-specific eCeLLM-L (0.611 vs 0.521). As evidenced by the results, by training on all tasks together, eCeLLM models enjoy strong versatility and could enable knowledge transfer across tasks for improved performance.

6.5. Impact of Training Data Size on eCeLLM

We investigate how the training data size impacts the performance of eCeLLM models. As presented in Section 3.4, ECInstruct has more than 92 thousand (92K) training samples. In this experiment, we generate 3 smaller training sets

Table 5. Performance on Unseen Instructions in IND Evaluation

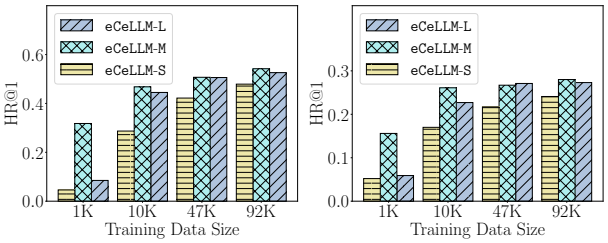
Model	Training Instructions	AVE	PRP	PM	SA	SR	MPC	PSI	QPR	AP	AG
		F1*	Macro F1	F1	Macro F1	HR@1	Accuracy	F1	NDCG	F1	F_{BERT}
eCeLLM-L	single	0.046	0.619	0.995	0.610	0.526	0.696	0.206	0.870	0.846	0.841
	diverse	**0.553**	**0.638**	0.995	**0.639**	0.524	0.694	**0.335**	0.870	0.842	0.841
eCeLLM-M	single	0.000	**0.618**	0.995	0.554	0.543	0.696	0.241	0.878	**0.852**	0.850
	diverse	**0.622**	0.540	0.995	**0.643**	0.540	0.695	**0.253**	0.878	0.822	0.844
eCeLLM-S	single	0.447	0.535	0.991	0.577	**0.478**	0.652	0.314	0.867	0.841	0.838
	diverse	**0.488**	**0.552**	0.991	0.577	0.457	**0.660**	**0.381**	0.871	0.845	0.842

In this table, "single" and "diverse" indicate that the eCeLLM models are tuned over single and diverse instructions, respectively. The best performance of each eCeLLM model (e.g., eCeLLM-L, eCeLLM-M abd eCeLLM-S) is in **bold**, if the performance difference between the eCeLLM model tuned over single and diverse instructions exceeds 1%.

Table 6. Performance of Generalist and Task-specific eCeLLM Models in IND Evaluation

Model	Training Tasks	AVE	PRP	PM	SA	SR	MPC	PSI	QPR	AP	AG
		F1*	Macro F1	F1	Macro F1	HR@1	Accuracy	F1	NDCG	F1	F_{BERT}
eCeLLM-L	Task-specific	**0.599**	0.521	0.995	0.616	0.518	0.655	0.000	**0.879**	0.854	0.841
	Generalist	0.582	**0.611**	0.995	**0.648**	**0.526**	**0.684**	**0.501**	0.870	0.851	0.841
eCeLLM-M	Task-specific	**0.757**	0.543	0.987	**0.655**	0.535	0.681	0.000	0.883	**0.864**	0.841
	Generalist	0.662	**0.558**	0.995	0.639	**0.542**	**0.696**	**0.305**	0.876	0.846	0.842
eCeLLM-S	Task-specific	0.397	0.348	0.991	**0.608**	0.413	0.646	0.000	0.858	0.835	0.835
	Generalist	**0.509**	**0.518**	0.991	0.596	**0.479**	0.650	**0.392**	**0.870**	**0.846**	0.842

In this table, "Task-specific" indicates that the eCeLLM models are tuned on individual tasks; "Generalist" represents tuning eCeLLM models using all tasks together. The best performance of generalist and task-specific eCeLLM models on each task is in **bold**, if the performance difference between the generalist and task-specific eCeLLM model exceeds 1%.



(a) IND Evaluation (b) OOD Evaluation

Figure 2. eCeLLM performance on SR

of 1K, 10K, and 47K samples. The 1K and 10K samples are generated by randomly selecting 0.1K and 1K samples, respectively, from the training set of each of the 10 tasks. The 47K samples are constructed by randomly selecting 5K samples from each task except for PM, for which we include all its 2K training samples. The complete results are presented in Appendix I.

Figure 2 presents the eCeLLM model performance on SR. As it shows, increasing training data size typically benefits eCeLLM model performance. For example, in IND evaluation, the performance of eCeLLM-L increases from 0.085 to 0.526 when the training data size is increased from 1K to 92K. Similarly, in OOD evaluation, increasing training samples from 1K to 92K significantly elevates eCeLLM-L performance from 0.059 to 0.273. A similar trend could be observed in eCeLLM-M and eCeLLM-S. As evidenced by our results, the large-scale training data is critical in developing effective e-commerce LLMs. This further highlights the significance of our extensive, comprehensive, and high-quality e-commerce instruction dataset, ECInstruct.

7. Conclusion and Limitation

This paper open-sources the first large-scale, high-quality benchmark dataset (ECInstruct), and develops the state-of-the-art generalist LLMs (eCeLLM series) for e-commerce. eCeLLM models are extensively evaluated against the most advanced baseline models including GPT-4 Turbo, EcomGPT, and SoTA task-specific models, in both IND and OOD evaluations. Our experimental results demonstrate that eCeLLM substantially outperforms the best baseline models with an average improvement of 10.7% in IND evaluation. Our results also show that eCeLLM exhibits excellent generalizability to OOD settings, evidenced by its considerable improvement of 9.3% over the best baselines on the OOD products. To the best of our knowledge, this study is the first comprehensive and systematic study of instruction-tuning LLMs for e-commerce applications.

This paper acknowledges certain limitations and future work. (1) Recently emerged e-commerce tasks such as explanation generation could be included to further improve the comprehensiveness of ECInstruct. (2) User profiling could be better enabled once metadata is available, and advanced foundation models could be developed to address unique and fundamental challenges and tasks in e-commerce.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Ahmed, F., Cui, Y., Fu, Y., and Chen, W. A graph neural network approach for product relationship prediction. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 85383, pp. V03AT03A036. American Society of Mechanical Engineers, 2021.

Anthropic. Model card and evaluations for claude models. *CoRR*, 2023. URL https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:207880568.

Deng, Y., Zhang, W., Yu, Q., and Lam, W. Product question answering in e-commerce: A survey. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

Ding, H., Deoras, A., Wang, B., and Wang, H. Zero-shot recommender systems. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

Frieder, S., Berner, J., Petersen, P., and Lukasiewicz, T. Large language models for mathematicians. *arXiv preprint arXiv:2312.04556*, 2023.

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

Geng, S., Liu, S., Fu, Z., Ge, Y., and Zhang, Y. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, pp. 299–315, 2022.

Gupta, M., Kulkarni, N., Chanda, R., Rayasam, A., and Lipton, Z. C. AmazonQA: A review-based question answering task. In *International Joint Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:199465954.

He, P., Gao, J., and Chen, W. DeBERTaV3: Improving DeBERTa using electra-style pre-training with gradient-disentangled embedding sharing. *The Eleventh International Conference on Learning Representations*, 2022.

He, R. and McAuley, J. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pp. 507–517, 2016.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Hou, Y., Mu, S., Zhao, W. X., Li, Y., Ding, B., and Wen, J.-R. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 585–593, 2022.

Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J., and Zhao, W. X. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pp. 364–381. Springer, 2024.

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Javaheripi, M. and Bubeck, S. Phi-2: The surprising power of small language models. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Köpcke, H., Thor, A., and Rahm, E. Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, 3(1-2):484–493, 2010.

Lee, A., Hunter, C., and Ruiz, N. Platypus: Quick, cheap, and powerful refinement of LLMs. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023a.

Lee, A., Miranda, B., and Koyejo, O. Beyond Scale: the diversity coefficient as a data quality metric demonstrates LLMs are pre-trained on formally diverse data. *ArXiv*, abs/2306.13840, 2023b. URL https://api.semanticscholar.org/CorpusID:259252412.

Li, J., Wang, M., Li, J., Fu, J., Shen, X., Shang, J., and McAuley, J. Text is all you need: Learning language representations for sequential recommendation. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

Li, Y., Ma, S., Wang, X., Huang, S., Jiang, C., Zheng, H.-T., Xie, P., Huang, F., and Jiang, Y. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18582–18590, 2024.

Lika, B., Kolomvatsos, K., and Hadjiefthymiades, S. Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41(4, Part 2):2065–2073, 2014. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2013.09.005. URL https://www.sciencedirect.com/science/article/pii/S0957417413007240.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., and Roberts, A. The Flan Collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:256415991.

McAuley, J., Targett, C., Shi, Q., and Van Den Hengel, A. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 43–52, 2015.

Nguyen, D. Q., Vu, T., and Nguyen, A. G.-T. BERTweet: A pre-trained language model for english tweets. In *Conference on Empirical Methods in Natural Language Processing*, 2020. URL https://api.semanticscholar.org/CorpusID:218719869.

Ni, J., Li, J., and McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/arXiv.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Petrov, A. V. and Macdonald, C. gSASRec: Reducing overconfidence in sequential recommendation trained with negative sampling. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 116–128, 2023.

Rahm, Erhard. Benchmark datasets for entity resolution. https://dbs.uni-leipzig.de/research/projects/benchmark-datasets-for-entity-resolution, 2010.

Reddy, C. K., Màrquez, L., Valero, F., Rao, N., Zaragoza, H., Bandyopadhyay, S., Biswas, A., Xing, A., and Subbian, K. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. *arXiv preprint arXiv:2206.06588*, 2022.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019. URL https://api.semanticscholar.org/CorpusID:201646309.

Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2021.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., and Welling, M. Modeling relational data with graph convolutional networks. In *15th International Conference on Extended Semantic Web Conference, ESWC 2018*, pp. 593–607. Springer, 2018.

Sellam, T., Das, D., and Parikh, A. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, 2020.

Shi, K., Sun, X., Wang, D., Fu, Y., Xu, G., and Li, Q. LLaMA-E: Empowering e-commerce authoring with multi-aspect instruction following. *arXiv preprint arXiv:2308.04913*, 2023.

Spatharioti, S. E., Rothschild, D. M., Goldstein, D. G., and Hofman, J. M. Comparing traditional and LLM-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744*, 2023.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023b. doi: 10.48550/arXiv.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, Q., Yang, L., Kanagal, B., Sanghai, S., Sivakumar, D., Shu, B., Yu, Z., and Elsas, J. Learning to extract attribute value from product via question answering: A multi-task approach. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 47–55, 2020.

Wang, Y., Wang, L., Li, Y., He, D., and Liu, T.-Y. A theoretical analysis of NDCG type ranking measures. In *Conference on learning theory*, pp. 25–54. PMLR, 2013.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-Instruct: Aligning language models with self-generated instructions. In *Annual Meeting of the Association for Computational Linguistics*, 2022a. URL https://api.semanticscholar.org/CorpusID:254877310.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H. G., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Patel, M., Pal, K. K., Moradshahi, M., Parmar, M., Purohit, M., Varshney, N., Kaza, P. R., Verma, P., Puri, R. S., Karia, R., Sampat, S. K., Doshi, S., Mishra, S. D., Reddy, S., Patro, S., Dixit, T., Shen, X., Baral, C., Choi, Y., Smith, N. A., Hajishirzi, H., and Khashabi, D. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2022b. URL https://api.semanticscholar.org/CorpusID:253098274.

Wang, Y., Jiang, Z., Chen, Z., Yang, F., Zhou, Y., Cho, E., Fan, X., Huang, X., Lu, Y., and Yang, Y. RecMind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296*, 2023.

Wankhade, M., Rao, A. C. S., and Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Xu, D., Ruan, C., Korpeoglu, E., Kumar, S., and Achan, K. Product knowledge graph embedding for e-commerce. In *Proceedings of the 13th international conference on web search and data mining*, pp. 672–680, 2020.

Xu, H., Wang, W., Mao, X., Jiang, X., and Lan, M. Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title. In *Proceedings of the 57th Annual Meeting of*

the Association for Computational Linguistics, pp. 5214–5223, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1514. URL https://aclanthology.org/P19-1514.

Xu, Z., Shen, Y., and Huang, L. MultiInstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:254926784.

Yang, L., Wang, Q., Yu, Z., Kulkarni, A., Sanghai, S., Shu, B., Elsas, J., and Kanagal, B. MAVE: A product dataset for multi-source attribute value extraction. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pp. 1256–1265, 2022.

Yang, Z., Wu, J., Luo, Y., Zhang, J., Yuan, Y., Zhang, A., Wang, X., and He, X. Large language model can interpret latent space of sequential recommender. *arXiv preprint arXiv:2310.20487*, 2023.

Yue, X., Qu, X., Zhang, G., Fu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. MAmmoTH: Building math generalist models through hybrid instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=yLClGs770I.

Zeng, F., Gan, W., Wang, Y., Liu, N., and Yu, P. S. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*, 2023.

Zhang, J., Bao, K., Zhang, Y., Wang, W., Feng, F., and He, X. Is ChatGPT fair for recommendation? evaluating fairness in large language model recommendation. *Proceedings of the 17th ACM Conference on Recommender Systems*, 2023a. URL https://api.semanticscholar.org/CorpusID:258676079.

Zhang, M., Yin, R., Yang, Z., Wang, Y., and Li, K. Advances and challenges of multi-task learning method in recommender system: A survey. *CoRR*, abs/2305.13843, 2023b. doi: 10.48550/ARXIV.2305.13843. URL https://doi.org/10.48550/arXiv.2305.13843.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2019.

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

A. Task Definition and Data Preprocessing

All tasks can be characterized into 4 categories: product understanding, user understanding, query product matching, and product question answering (product QA). Table A1 summarizes the task definitions, evaluation metrics, data sources for each task, and their OOD test settings, if any. The details are articulated below.

Table A1. Tasks and `ECInstruct` Datasets

	Task	Definition	Type	Metrics	Data
Product Understanding	AVE	Given the titles, descriptions, features, and brands of the products, extract values for the specific target attributes.	Information extraction	precision*, recall*, F1* (Section A.1.1)	MAVE (Yang et al., 2022) based on Amazon Review 2018 (Ni et al., 2019); **OOD**: 7 held-out attributes
	PRP	Given the titles of two products, predict their relation from *"also buy", "also view"*, and *"similar"*.	Multi-class classification	accuracy, macro precision, macro recall, macro F1	Amazon Review 2018 (Ni et al., 2019); **OOD**: Tools category
	PM	Given the titles, descriptions, manufacturers, and prices of the products from two different platforms, predict if they are the same product.	Binary classification	accuracy, precision, recall, F1, specificity, negative prediction rate	Amazon-Google Product (Köpcke et al., 2010; Rahm, Erhard, 2010)
User Understanding	SA	Given a product review by a user, identify the sentiment that the user expressed on the product.	Multi-class classification	accuracy, macro precision, macro recall, macro F1	Amazon Review 2018 (Ni et al., 2019); **OOD**: Tools category
	SR	Given the interactions of a user over the products, predict the next product that the user would be interested in.	Ranking	HR@1	Amazon Review 2018 (Ni et al., 2019) and Amazon Review 2014 (He & McAuley, 2016; McAuley et al., 2015); **OOD**: Tools category
Query Product Matching	MPC	Given a query and a product title, predict the relevance between the query and the product.	Multi-class classification	accuracy, macro precision, macro recall, macro F1	Shopping Queries Dataset (Reddy et al., 2022)
	PSI	Given a user query and a potentially relevant product, predict if the product can serve as a substitute for the user's query.	Binary classification	accuracy, precision, recall, F1, specificity, negative prediction rate	Shopping Queries Dataset (Reddy et al., 2022)
	QPR	Given a user query and a list of potentially relevant products to the query, rank the products according to their relevance to the query.	Ranking	NDCG[1]	Shopping Queries Dataset (Reddy et al., 2022)
Product QA	AP	Given a product-related question and reviews of this product, predict if the question is answerable.	Binary classification	accuracy, precision, recall, F1, specificity, negative prediction rate	AmazonQA (Gupta et al., 2019); **OOD**: Cells category
	AG	Given a product-related question and reviews as supporting documents, generate the answer to the question.	Generation	P_{BERT}[2], R_{BERT}[2], F_{BERT}[2], BLEURT[3]	AmazonQA (Gupta et al., 2019); **OOD**: Cells category

Underline indicates the primary metrics. Higher values of the primary metrics indicate better model performance. **OOD** refers to the out-of-domain product category/attributes. [1] NDCG: normalized discounted cumulative gain (Wang et al., 2013) [2] P_{BERT}, R_{BERT}, F_{BERT}: BERTScore evaluates the quality of generated texts by measuring the similarity between the embeddings of tokens in the generated text and the ground-truth text (Zhang et al., 2019). [3] BLEURT measures the text quality by comparing the similarity of sentences using contextual embeddings from pre-trained models. (Sellam et al., 2020).

To pursue adherence to data usage requirements, we check the licenses of `ECInstruct` data sources, ensuring their permission to publish. Table A2 presents the licenses of our curated dataset sources.

Data Split Raw datasets of the attribute value extraction (`AVE`, discussed below in Section A.1.1), product matching (`PM`, discussed in Section A.1.3), product relation prediction (`PRP`, discussed in Section A.1.2), and sentiment analysis

Table A2. Details of Data Source License

Dataset	License Type	Source
Amazon-Google Products	CC-by-4.0	https://dbs.uni-leipzig.de/research/projects/benchmark-datasets-for-entity-resolution
Amazon Review	Not Specified	https://cseweb.ucsd.edu/~jmcauley/datasets.html#amazon_reviews
AmazonQA	Not Specified	https://github.com/amazonqa/amazonqa
Shopping Queries Dataset	Apache License 2.0	https://github.com/amazon-science/esci-data

(SA, discussed in Section A.2.1) tasks are first split into training, validation, and test data at 8:1:1 ratio. For multi-class product classification (MPC, discussed in Section A.3.1), query product ranking (QPR, discussed in Section A.3.3), product substitute identification (PSI, discussed in Section A.3.2), answerability prediction (AP, discussed in Section A.4.1), and answer generation (AG, discussed in Section A.4.2) tasks, the raw datasets are already split. All split data are processed as detailed below and converged with instruction templates into structured data, which will be directly used by LLMs. Based on prior research (Wei et al., 2021) and considering the high computing demands, we uniformly downsample training sets to a size of 10K, validation sets to 1K, and test sets to 1K size, optimizing data volumes for efficient processing and affordable LLM evaluation.

In-domain (IND) Data Selection In PRP, SA, and SR tasks, we utilize the Amazon Review 2018 dataset (Ni et al., 2019) as the data source, which contains product reviews and metadata in 29 different product categories. For these three tasks, the data from the Electronics, Home, and Sports categories are employed as the IND data sources. We further downsample and combine the processed structured data for each task from these sources at a 1:1:1 ratio into a training, validation, and test set, respectively. For AP and AG tasks, we use the data from the AmazonQA dataset (Gupta et al., 2019), which has 17 product categories. To avoid data leakage, we use the data from the Sports and Tools categories for the AP task, and Electronics and Home for AG task as IND data sources, respectively. Similarly to the previous three tasks, for each of the tasks, we sample and combine the processed data from the IND sources at a 1:1:1 ratio into training, validation, and test sets. For the PM task, we process all data from Amazon-Google Products (Köpcke et al., 2010; Rahm, Erhard, 2010) and get a 2K training set, 0.2K validation set, and 0.2K test set. Since this dataset is small, we do not do the downsampling for the PM task. For AVE, MPC, QPR, and PSI tasks, we directly process and downsample them from the corresponding split in original datasets, separately.

Out-of-domain (OOD) Data Selection As for OOD datasets, only test data are utilized. The data from the Tools category from the Amazon Review dataset is used as the OOD data source in PRP, SA, and SR tasks, while the Cells category from AmazonQA serves as the OOD source in AP and AG tasks. For the AVE task, we hold out 7 attributes as the OOD dataset. All OOD sources are processed into 1K test sets.

Table A3. Summary of the ECInstruct Dataset and Tests

Task	Training	Validation	IND Test	OOD Test
AVE	10,000	1,000	1,000	1,000
PM	2,022	253	253	✗
PRP	10,000	1,000	1,000	1,000
SA	10,000	1,000	1,000	1,000
SR	10,000	1,000	1,000	1,000
MPC	10,000	1,000	1,000	✗
QPR	10,000	1,000	1,000	✗
PSI	10,000	1,000	1,000	✗
AP	10,000	1,000	1,000	1,000
AG	10,000	1,000	1,000	1,000
ECInstruct	92,022	9,253	10 tasks	6 tasks

In this table, IND and OOD refers to the in-domain evaluation and out-of-domain evaluation, respectively.

A.1. Product Understanding Tasks

Three tasks are defined as follows to understand the different aspects of products.

A.1.1. ATTRIBUTE VALUE EXTRACTION (AVE)

Definition: Given the titles, descriptions, features, and brands of the products, extract values for the specific target attributes. By understanding product attribute values, models can extract key properties of the products and build the profiling for them, which is beneficial in many e-commerce scenarios, such as customer service agents and explanations.

Data Processing: The Amazon Review 2018 dataset (Ni et al., 2019) is used as the raw data. We extract the ground-truth attribute-value pairs following the protocol of MAVE (Yang et al., 2022) and get 4,767,579 data entries with 683 attributes in total. We randomly hold out 7 attributes from the dataset as the OOD data source and process both the OOD data and remaining IND data with instructions. The samples that have ground-truth values and corresponding sources for the specific attribute in the dataset are referred to as positive samples, while the samples that have no extracted ground-truth values are denoted as negative samples.

Evaluation Metrics: Following the evaluation definition in previous work MAVE (Yang et al., 2022), the model can predict null value (NV) or incorrect value (IV) for negative samples. For positive samples, the model predictions can be correct value (CV), wrong value (WV), and null value (NL). In MAVE, customized precision and recall are calculated as the percentage of correctly predicted positive samples in all predictions and ground truths, respectively. This formulation, however, fails to evaluate model results on negative samples. Therefore, with the definition above, we improve the formulation by considering both positive and negative samples as follows:

$$\text{precision}^* = \frac{\text{NV} + \text{CV}}{\text{NV} + \text{IV} + \text{CV} + \text{WV}}, \ \text{recall}^* = \frac{\text{NV} + \text{CV}}{\text{N}}, \ \text{F1}^* = \frac{2 \times \text{precision}^* \times \text{recall}^*}{\text{precision}^* + \text{recall}^*}, \tag{1}$$

where N refers to the total number of entries in ground-truth data. Note that we only consider the prediction result as the correct value (VC) when the ground truth is fully contained in the prediction (e.g. "bright yellow" is considered as a correct value with the ground truth of "yellow"). In our formulation, the precision* is calculated as the percentage of correctly predicted positives and negatives in all predictions, and the recall* is computed as the percentage of correct positive and negative predictions in all ground truths. We use F1* as the primary metric for the AVE task.

Example:

- Input:
  - Product title: Bencore Multi Functional Molle Tactical Messenger Bag.
  - Product description: This rugged/durable tactical shoulder bag provides perfect and stylish solution for almost any scenario. The bag is made of durable nylon construction that will not tear, color will not fade. The bag has many MOLLE straps through the bag for all your MOLLE accessories. The bag contains many roomy compartments as pictured and comes in many stylish colors. Design, comfort and functionality was the emphasis of this bag which is why we made sure the bag is fully ergonomic, lightweight and has many roomy pockets and Velcro patches throughout the bag. The product comes with the Bencore Life Time warranty and is satisfaction guaranteed. Bencore is a leading manufacturer in outdoor Apparel/Accessories, from Par cords to Backpacks to basic outdoor essentials.
  - Product feature: Durable heavy-duty, lightweight Nylon construction, will not tear or break even under extreme conditions - Lifetime Warranty, Rugged, roomy main drawstring-closed compartment provides secure storage space for your gear; MOLLE System, works with most MOLLE accessory, Front pocket provides quick access, roomy interior pocket for convenient separated storage, concealed back pocket with zipper closure, Padded and fully ergonomic System, adjustable shoulder strap for comfortable handling.
  - Product brand: Bencore
  - Target attributes: Material
- Output:
  - Attribute: material; Value: nylon; Source: product description.
  - Attribute: material; Value: nylon; Source: product feature.

Note that in the above example, for the "target attribute" "material", the model should extract its value "nylon" from the "product description" and "product feature".

A.1.2. Product Relation Prediction (PRP)

Definition: Given the titles of two products, predict their relation. Studying the relations between products can help models generate better results when conducting other e-commerce tasks such as recommendations.

Data Processing: To learn the relationship between products, we use the product metadata of Electronics, Home, and Sports categories from the Amazon Review 2018 dataset (Ni et al., 2019) as the IND sources and Tools as the OOD source. We collect product IDs from metadata and remove the products without detailed information in the metadata. In this task, the product titles are used to represent products. The product pairs that appear more than once with different relations are eliminated. After filtering and combining data with instruction templates, the three relations (*also buy, also view, and similar*) in the structured dataset are roughly 7:10:1.

Evaluation Metrics: With three different relations in this task, accuracy, macro precision, macro recall, and macro F1 are employed as metrics. Meanwhile, macro F1 is used as the primary metric for combining evaluation results of different labels and providing a comprehensive measurement of the model performances.

Example:

- Input:
  - Product 1: Monoprice 11952 Polyurethane Replacement Ear Pads for PID 8323 type Headphones - Red
  - Product 2: Monoprice Hi-Fi Light Weight Over the Ear Headphones - Black with a 50mm driver and a 47in 3.5mm cable for Apple iPhone iPod Android Smartphone Samsung Galaxy Tablets MP3
- Options:
  - A. Users who view product 1 may also buy product 2.
  - B. Users who view product 1 may also view product 2.
  - C. The product 1 is similar with the product 2.
- Output:
  - B

Note that in the above example, the relation of Product 1 and Product 2 is *also_view*.

A.1.3. Product Matching (PM)

Definition: Given the titles, descriptions, manufacturers, and prices of the products from two different platforms, predict if they are the same product. This task enables the model to learn the similarities among products.

Data Processing: We use the original data (Köpcke et al., 2010; Rahm, Erhard, 2010), which contains detailed product information from Amazon and Google platforms, as well as 1.3K matching product pairs between the two platforms. After deduplication, we randomly sample the same amount of unmatched pairs since there are only matched samples in the raw data and thus get 2,530 product pairs in total. The product pairs are split into the training, validation, and test sets with a ratio of 8:1:1, and processed with instruction templates respectively.

Evaluation Metrics: For this binary classification task, we use precision (positive prediction rate), recall (sensitivity), F1, specificity (recall of negative labels), negative prediction rate (NPR, precision of negative labels), and accuracy as metrics. We choose F1 as the primary metric since it serves as a balanced evaluation metric by combining precision and recall to generally reflect model performance.

Example:

- Input:
  - Product 1: title - marine aquarium 2.5 virtual undersea paradise win/mac, description - marine aquarium 2.0 is like having a small piece of an aquatic paradise in your home -- without having to take care of actual fish, manufacturer - encore software, price - 19.99
  - Product 2: title - encore software 25020 - marine aquarium 2.5 (hybrid) - win 95 98 me 2000 xp/mac 10.1 or higher, description - encore software 25020: marine aquarium 2.5 hybrid discover the virtual fish tank phenomenon that has everyone talking! marine aquarium 2.5 delivers a stunning undersea paradise through your desktop with 26 exotic species of fish, manufacturer - encore software, price - 19.97
- Output:
  - Yes

In the above example, the two products (Product 1, Product 2) are identified as the same product.

A.2. User Understanding

The two tasks in this section aim to help the models comprehend the users' needs and preferences.

A.2.1. Sentiment Analysis (SA)

Definition: Given a product review by a user, identify the sentiment that the user expressed on the product. The task will help models understand what sentiment users express and recommend more proper products to the user

Data Processing: For the sentiment analysis, we also use the review data of Electronics, Home, and Sports categories from the Amazon Review 2018 dataset (Ni et al., 2019) as the IND sources and Tools category as the OOD source. We only keep the data with reviews of at least 10 words. The ratings are used as the ground-truth sentiment level, while 5.0 refers to very positive and 1.0 refers to very negative. After downsampling and combining the raw data with instruction templates, the five labels (from 1.0 to 5.0) in the structured dataset are roughly 4:2:3:8:25.

Evaluation Metrics: With five different labels in the SA task, accuracy, macro precision, macro recall, and macro F1 are employed as metrics. macro F1 is used as the primary metric.

Example:

- Input:
  - This is really perfect for my kids who have a thick hair. I cam be able to create a beautiful hair bun with them. I would like to recommend this to all.
- Options:
  - A. Very positive
  - B. Positive
  - C. Neutral
  - D. Negative
  - E. Very negative
- Output:
  - A

The user from the above example expressed a very positive sentiment in the review.

A.2.2. SEQUENTIAL RECOMMENDATION (SR)

Definition: Given the interactions of a user over the products, predict the next product that the user would be interested in. By learning on this task, the models will have a comprehensive view of user preferences, which enables models to cater to users' future needs.

Data Processing: In the SR task, we use both product reviews and metadata from the Amazon Review 2018 dataset (Ni et al., 2019). Meanwhile, we use metadata from the Amazon Review 2014 dataset (He & McAuley, 2016; McAuley et al., 2015) as a supplement. The data of Electronics, Home, and Sports categories from both datasets serve as IND sources and the Tools category is used for the OOD test. Moreover, we consider users' review histories as their interactions with products. Following the data processing protocol of UnisRec (Hou et al., 2022), we remove the products without metadata in the 2018 version dataset and conduct the 5-core filter to ensure all products and users appear at least 5 times. After filtering, we sort every user's interactions chronologically and truncate the history with a maximum of 50 products, retaining the least recent history. For the text information of products, we use the metadata from Amazon Review 2014 to fill in the missing information of the same products in the 2018 version metadata.

We also combine the product title, category, and brand to represent a product. The average length of the combined texts is about 21 words. Thus, we retain the first, maximum of 25 words of each combined text for computational efficiency. As in conventional sequential recommendation tasks, we split the last product of the user interactions into the test set as the ground truth next product of the user's interest, the second last product into the validation set, and the remaining products into the training set. When processing the user interactions with instruction templates, for each sample, we randomly select 19 candidate products and mix them up with one ground-truth product. These 20 products serve as the options for the sample.

Evaluation Metrics: We evaluate the SR task on hit rate at top 1 (HR@1), which is a popular metric in sequential recommendation and measures if the top-ranked product matches the ground-truth user interaction. HR@1 also serves as the primary metric in this task.

Example:

- Input:
  - 1st: M-Edge Latitude Kindle Jacket, Pink (Fits Kindle Keyboard). Electronics. Computers & Accessories. M-Edge.
  - 2nd: Marware jurni Kindle Fire Case Cover, Black (will not fit HD or HDX models). Electronics. Computers & Accessories. Marware.
  - 3rd: NETGEAR AC1600 Dual Band Wi-Fi Gigabit Router (R6250). Electronics. Computers & Accessories. NETGEAR.
  - 4th: iMBAPrice 110014-1 (1-Pack) Glod Plated 2.4 Ghz 3-Way Coaxial Cable Splitter F-Type Screw for Video Satellite Splitter/VCR/Cable Splitter/TV Splitter/Antenna Splitter/RG6 Splitter. Electronics. Accessories & Supplies...
- Options:
  - A: T POWER 9v 12v (6.6ft Long Cable) Ac Dc Adapter Compatible with X Rocker Pro Series H3 51259 Video Gaming Chair 51231,51396 & V Rocker 5130301...
  - B: Boys Floatsafe Flotie Soft Fabric Armbands Floatie Blue For Kids Ages 1 To 3. Floatsafe Floatie
  - C: Anker iPhone Charger, Powerline Lightning Cable (3ft), MFi Certified for iPhone Xs/XS Max/XR/X
  - D: Curtain Drapery Rod w/brackets Small - Wrought Iron Hand Made. Home & Kitchen. Home Dcor. Hand Crafted & American Made!
  - . . .
  - T: Lorex ACCMIC1 Indoor Audio Microphone Accessory for Surveillance DVR's (Black). Electronics. Camera & Photo. Lorex
- Output:
  - A

Given the interactions shown in the above input, the next product of user's interest for this user is C: Anker iPhone Charger, Powerline Lightning Cable (3ft), MFi Certified for iPhone Xs/XS Max/XR/X

A.3. Query Product Matching

The following three tasks seek to study the relations between the user queries and the potential relevant products to the queries. We use the data from the Shopping Queries Dataset (Reddy et al., 2022), which contains 48,300 queries and potentially relevant products to each query. The dataset was originally collected from different market locales in various languages. All products are labeled in four categories: *Exact, Substitute, Complement, and Irrelevant*, based on their relevance to the query. As in the Shopping Queries Dataset, the four labels are defined as follows:

- *Exact*: The item is relevant for the query, and satisfies all the query specifications (e.g., a water bottle matching all attributes of a query "plastic water bottle 24oz", such as material and size).
- *Substitute*: The item is somewhat relevant, i.e., it fails to fulfill some aspects of the query, but the item can be used as a functional substitute (e.g., fleece for a "sweater" query)
- *Complement*: The item does not fulfill the query, but could be used in combination with an exact item (e.g., track pants for "running shoes" query).
- *Irrelevant*: The item is irrelevant, or it fails to fulfill a central aspect of the query (e.g., socks for a "telescope" query, or a wheat flour bread for a "gluten-free bread" query)

In the following query-related tasks, we only utilize the data that is both in English and from the market of the U.S. locale.

A.3.1. Multi-class Product Classification (MPC)

Definition: Given a query and a product title, predict the relevance between the query and the product (*Exact, Substitute, Complement, Irrelevant*). This task helps models learn the fine-grained relevance between queries and products, promoting better recommendation results.

Data Processing: We use the product titles to represent products and translate the Unicode into English. The translated query-product pairs are processed with instruction templates. The ratio of the four labels (*Exact, Substitute, Complement, and Irrelevant*) in the structured dataset is about 35:10:1:5.

Evaluation Metrics: Accuracy, macro precision, macro recall, and macro F1 are employed as metrics. Accuracy is used as the primary metric in the MPC task.

Example:

- Input:
  - Query: aj1 black and white
  - Product: Nike Men's Air Jordan 1 Low White/Gym Red, White/Gym Red/Black, 9
- Options:
  - A: The product is relevant to the query, and satisfies all the query specifications.
  - B: The product is somewhat relevant. It fails to fulfill some aspects of the query but the product can be used as a functional substitute.
  - C: The product does not fulfill the query, but could be used in combination with a product exactly matching the query.
  - D: The product is irrelevant to the query.
- Output:
  - B

In the above example, the product serves as a substitute for the product described in the query.

A.3.2. Product Substitute Identification (PSI)

Definition: Given a user query and a potentially relevant product, predict if the product can serve as a substitute for the user's query.

Data Processing: The preprocessing of the PSI is similar to that of MPC, except that the PSI is a binary classification task. The query-product pairs with *Exact, Complement, or Irrelevant* labels are relabeled as non-substitute. After combining the query-product-label triples with instruction templates, the ratio of *substitute* (positive) and *non-substitute* (negative) labels is approximately 4:1.

Evaluation Metrics: For this binary classification task, we use precision (positive prediction rate), recall (sensitivity), F1, specificity (recall of negative labels), negative prediction rate (NPR, precision of negative labels), and accuracy as metrics. We choose F1 as the primary metric.

Example:

- Input:
  - Query: fissler magic smooth-edge can opener
  - Product: KUKINO Manual Can Opener, Multifunction Handheld Food Grade Stainless Steel Can Openers, Black.
- Output:
  - No

The product does not serve as a substitute for the query in the above example.

A.3.3. Query-product Ranking (QPR)

Definition: Given a user query and a list of potentially relevant products to the query, rank the products according to their relevance to the query.

Data Processing: In the QPR task, we utilize the product titles to represent the products. A query and a list of potentially relevant products with different relevance labels to this query constitute a sample. We sort the list of products according to their relevance and generate the ground truth for each sample.

Evaluation Metrics: We evaluate this task using normalized discounted cumulative gain (NDCG) (Wang et al., 2013), which is a prevalent evaluation metric for assessing the ranking quality in the existing literature. NDCG is also used as the primary metric in this task.

Example:

- Input:
  - Query: high heel shoe chair
  - Product A: ORE International HBB1826 High Heel Shoe Display with Hooks Jewelry Box, Cheetah Print.
  - Product B: Coconut Float Red High Heel Gigantic Pool Float for Adults, 91.
  - Product C: Wildkin Kids Wooden Bench Seat with Storage for Boys and Girls, Toy Box Bench Seat Features Safety Hinge, Backrest, and Two Carrying Handles, Measures 32 x 15.5 x 27 Inches (Wild Side) (LOD71001).
- Output:
  - A, C, B

In the above example, Product A is more relevant to the query than Product C, and Product C is more relevant than Product B.

A.4. Product Question Answering

These two tasks aim to learn from the products and answer the product-related questions by using the AmazonQA dataset (Gupta et al., 2019). This dataset involves 923,685 product-related questions. There are around 9 product-related reviews and 4 answers provided by Amazon users for each question. There are also *questionType* and *is_answerable* annotations for each question. We eliminate all meaningless notations such as HTML tags or emoji from questions, reviews, and answers.

A.4.1. Answerability Prediction (AP)

Definition: Given a product-related question and reviews of this product, predict if the question is answerable.

Data Processing: We use the data of the Sports and Tools categories from AmazonQA as the IND sources and the Cell category as the OOD source. The *is_answerable* annotations are used as the ground truth. The ratio of answerable (positive) and unanswerable (negative) samples in the structured dataset is around 2:3.

Evaluation Metrics: For the AP task, we use precision (positive prediction rate), recall (sensitivity), F1, specificity (recall of negative labels), negative prediction rate (NPR, precision of negative labels), and accuracy as metrics. We use F1 as the primary metric when evaluating this task.

Example:

- Input:
  - Question: Where do you purchase the paddles or do paddles come with it?
  - Document: Very happy with purchase and price! \My son spends hours playing with this. It was easy to assemble and he loves it! Very happy with the purchase. \You won't regret this purchase! A little awkward to assemble, as the instructions say you need 2 people to do it. \This is well built. Great value. Fun and exercise for the whole family! Buy it today and have it for years to come. \Sturdy, well made and will be around for many years to come! Totally awesome for my son to be able to play ball on his own sometimes ;)
- Output:
  - No

From the above example, the question is unanswerable based on the document.

A.4.2. Answer Generation (AG)

Definition: Given a product-related question and reviews as supporting documents, generate the answer to the question.

Data Processing: This task only involves answerable and open-ended questions. To avoid data leakage, we use the Electronics and Home categories, which are different from the AP task, as the IND sources and the Cells category as the OOD source. Every data entry in the AmazonQA has a list of answers labeled with *helpfulness* levels. We choose the most helpful answer as the ground truth.

Evaluation Metrics: The AG task is evaluated on precision, recall, and F1 of the BERTScore (Zhang et al., 2019), and BLEURT (Sellam et al., 2020). BERTScore and BLEURT both evaluate the similarity of sentences by leveraging contextual embeddings from pre-trained models and are widely used in the NLP field. In this task, we use F1 BERTScore as the primary metric.

Example:

- Input:

  - Question: Can you add additional receiver with just one sensor? So one sensor picks up the the signal and sends it to two receiver.
  - Document: I have a 1200ft driveway and the unit works perfectly. The feature that is missing is the option to have more than one notification pattern if you have more than one sensor. For example, 1, 2, 3 or 4 beeps would tell you which area the motion is coming from. If you want a reliable motion sensor that works for a long distance then this is your unit. \Installed this system about two weeks ago, 300 feet from house toward end of driveway and it has never failed. Worked in rain with no problems or false alarms. There is almost 40 feet of drive remaining and I installed this at slight angle up the driveway. Larger vehicles (e.g. garbage truck, tractors mowers, etc) that drive slowly by at end of drive will also initiate the alarm. Faster or smaller vehicles on road will not be picked up, makes it really nice. Have two receivers, one indoor and one out back, makes it really a valuable alarm for us. \I bought several brands of alarm systems. for the money I can't see how you would be disappointed for the cost. I have this unit about 200' from the receiver, and it works great. \...

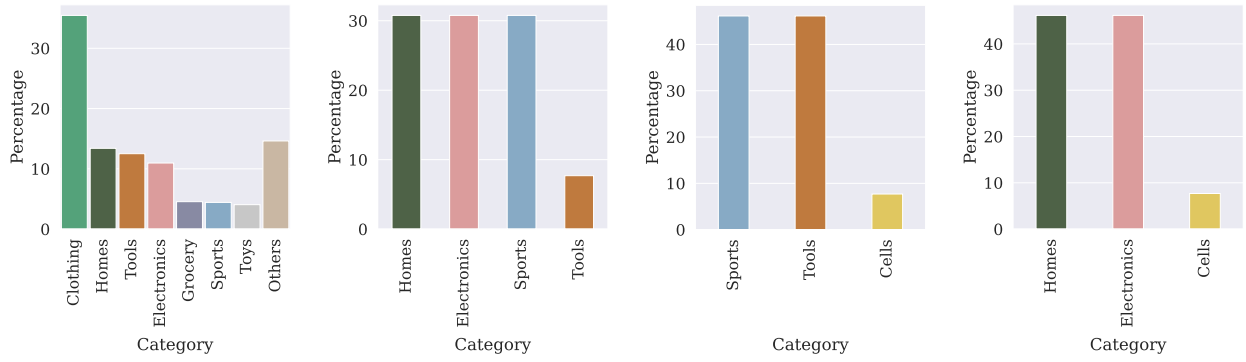
- Output:

  - Yes...just be sure all have the same dip switch settings.

Note that the answer to the above question that is semantically similar to the output should be generated from the document.

B. Data Statistics

The comprehensive and wide-ranging tasks in ECInstruct are critical for developing versatile e-commerce foundation models. ECInstruct includes products from 21 categories, such as Sports, Electronics, and Home. This broad coverage ensures that ECInstruct encompasses diverse product information, facilitating the creation of generalist e-commerce models.



(a) AVE (b) PRP, SA, and SR (c) AP (d) AG

Figure A1. Distribution of Product Category

Figure A1 shows the distribution of product categories for AVE, PRP, SA, SR, AP, and AG. The diverse categories in ECInstruct significantly enhance the effectiveness of eCeLLM. Note that the data of PM, MPC, PSI, and QPR tasks lacks category information, thus their distributions are not shown. Besides product categories, we also present the distributions of input lengths for each task, measured by word count, in Figure A2. For better clarity, we exclude very long inputs (those representing at most 1% of samples) in the AVE, PRP, SA, and SR tasks.

Due to limited computing resources, especially in academic settings, the current ECInstruct dataset has at most 10k training samples per task to fit within the model training capacity. However, scaling up ECInstruct is straightforward by following our data processing procedures detailed in Section 3 and Appendix A. For example, approximately 4.8 million high-quality samples could be further included in the AVE task as outlined in Appendix A.1.1. To allow researchers and developers to customize ECInstruct to their specific needs, all data processing scripts are available at https://github.com/ninglab/eCeLLM/tree/main/data_processing.

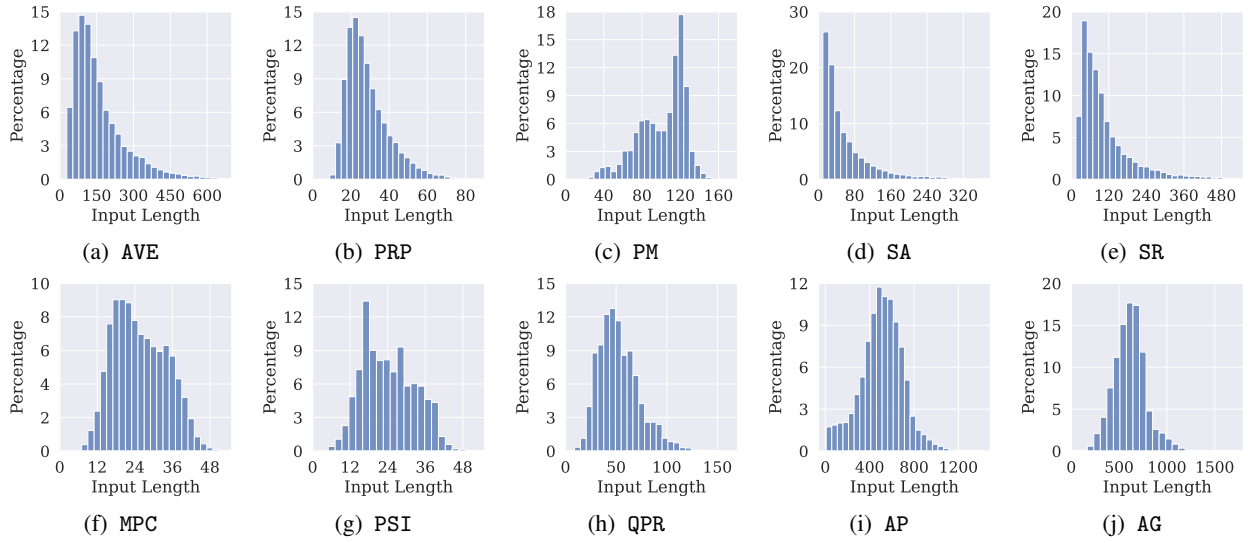


Figure A2. Distribution of Input Length

Table A4 provides the number of samples at different stages: in the raw data, after data filtering and quality checks, and after sampling. The "After Filtering" column shows the remaining data after applying quality checks. The "After Sampling" column shows the data included in ECInstruct. Detailed statistics for the "After Sampling" column, including training, validation, and testing samples for both in-domain and out-of-domain evaluation, are available in Table A3.

Even after filtering and quality checks, a substantial number of samples remain available for each task, making ECInstruct easily extendable and scalable. We believe ECInstruct can significantly benefit e-commerce foundation model research in various applications.

Table A4. Number of Samples in Each Task

Task	Raw Data	After Filtering	After Sampling
AVE	15,177,901	4,767,579	13,000
PRP	10,868,853	7,705,363	13,000
PM	2,600	2,528	2,528
SA	800,000	617,200	13,000
SR	31,013,687	1,946,393	13,000
MPC	2,621,288	935,757	12,000
PSI	2,621,288	935,757	12,000
QPR	130,652	57,707	12,000
AP	220,541	184,814	13,000
AG	398,654	213,612	13,000

The "After Sampling" column indicates the data included in ECInstruct.

C. Instructions

Each task has 6 instructions. One of them is the human-written seed instruction, 4 of them are GPT-4 generated diverse instructions, and the last one is the "unseen" instruction.

C.1. Instructions on Attribute Value Extraction (AVE)

Seed Instruction *Given the title, description, feature, price, and brand of a product and a set of target attributes, extract the value of each target attribute from the product information. Output the extracted value and the corresponding source (e.g., title or feature) denoting where the value is extracted.*

Generated Instruction 1 *Extract the value of the target attribute from the given product information and output it along with the corresponding source.*

Generated Instruction 2 *Parse the product information to locate the target attribute, and then provide the extracted value of the target attribute and its source in the output, specifying None if the attribute is not present.*

Generated Instruction 3 *First, identify the target attributes from the provided list. Then, scan the product title, description, feature, and brand to extract the values associated with each target attribute. Finally, create a list of dictionaries, each containing the extracted attribute, its corresponding value, and the source where it was found.*

Generated Instruction 4 *Using the product's title, description, features, price, and brand, identify and retrieve the values associated with a specified set of target attributes. Output the extracted values along with their respective sources (e.g., title or feature) indicating where each value was found.*

Unseen Instruction *Retrieve the value associated with the target attribute from the product information and specify the source (e.g., title, description, feature, or title) where the value was found.*

C.2. Instructions on Product Relation Prediction (PRP)

Seed Instruction *Given the title of two products, predict if the two products are similar, if the two products will be purchased or viewed together. Answer only from the options.*

Generated Instruction 1 *Analyze the titles of Product 1 and Product 2 to determine if they are similar, if they will be purchased or viewed together, and choose the corresponding option.*

Generated Instruction 2 *Evaluate the titles of Product 1 and Product 2, then choose the option that best describes the relation between the two products.*

Generated Instruction 3 *Evaluate the titles of Product 1 and Product 2 to assess their similarity and whether they are likely to be purchased or viewed together. Then, select the appropriate option.*

Generated Instruction 4 *Predict whether two products are similar, whether two products are likely to be purchased or viewed together based on their titles. Choose your answer from the provided options.*

Unseen Instruction *Analyze the titles of Product 1 and Product 2 and select the option that indicates the relation of the two products.*

C.3. Instructions on Product Matching (PM)

Seed Instruction *Given the title, description, manufacturer, and price of two products, identify if they are the same product. Only output yes or no.*

Generated Instruction 1 *Analyze the title, description, manufacturer, and price between the two products below and generate an output of yes if the two products are the same, otherwise respond with no.*

Generated Instruction 2 *Check the details of the two products to see if they refer to the same product. Output only yes or no.*

Generated Instruction 3 *Based on the product information, predict if the two products are identical or not. Output yes if they are identical or no otherwise.*

Generated Instruction 4 *Compare the details of two given products to determine if they are identical. Output yes if they are identical or no otherwise.*

Unseen Instruction *Determine whether the two products are the same by comparing their title, description, manufacturer, and price, and provide a simple yes or no answer as the output.*

C.4. Instructions on Sentiment Analysis (SA)

Seed Instruction *Given the user's review, identify the user's sentiment from the listed options. Answer using one of the options.*

Generated Instruction 1 *Assess the user's sentiment in the provided review and select the appropriate sentiment option from the list as the answer.*

Generated Instruction 2 *Determine the sentiment expressed by the user in her review from the provided choices, and respond by selecting one of the available options.*

Generated Instruction 3 *Carefully assess the user's review for any strong expressions of sentiment, either positive or negative. Based on your analysis, select the most fitting sentiment option from the provided list as output.*

Generated Instruction 4 *Analyze the user's review text and determine the overall sentiment expressed, then choose the corresponding sentiment option from the provided list (A: very positive, B: positive, C: neutral, D: negative, E: very negative) based on the identified sentiment.*

Unseen Instruction *Analyze the user's review and determine the sentiment based on the listed options.*

C.5. Instructions on Sequential Recommendation (SR)

Seed Instruction *Given the products the user has purchased in history, rank the items in the listed options and output the item that the user is most likely to purchase next. Answer from one of the options.*

Generated Instruction 1 *Based on the user's historical purchases, rank the items in options and predict the next product of the user's interest from the provided options.*

Generated Instruction 2 *Rank the items in options and predict the user's next purchase from the listed options by analyzing her historical purchases.*

Generated Instruction 3 *The user's purchase history implies her preferences. Rank the items in the options based on the user's preferences. Output the item that the user is most likely to purchase next from the options.*

Generated Instruction 4 *Rank items in listed options based on the user's purchase history to determine the item that the user is most likely to purchase next. Output the item with the highest likelihood of being the next purchase.*

Unseen Instruction *Estimate the user's intent based on the user's purchase history, and predict the next product that the user is most likely to purchase from the given options.*

C.6. Instructions on Multi-class Product Classification (MPC)

Seed Instruction *What is the relevance between the query and the product title below? Answer from one of the options.*

Generated Instruction 1 *Analyze the query and product title to determine the relevance between the query and product, and select the appropriate option from the provided options.*

Generated Instruction 2 *Evaluate the relevance between the query and product title, and choose the most accurate option from the given options.*

Generated Instruction 3 *Analyze the query and product title to assess the level of relevance between them, and then output the corresponding option that best describes this relevance.*

Generated Instruction 4 *Determine the relevance between the query and the product title provided, and select your response from one of the available options.*

Unseen Instruction *Compare the query and the product title to determine if the product fully meets the query specifications. Choose the option that best describes the relevance between them.*

C.7. Instructions on Product Substitute Identification (PSI)

Seed Instruction *Given a query and a product, identify if the product is somewhat relevant to the query. It fails to fulfill some aspects of the query but the product can be used as a functional substitute. Only output yes or no.*

Generated Instruction 1 *Answer yes if the product is a substitute for the query and no otherwise.*

Generated Instruction 2 *Please respond with yes if the product is a suitable substitute for the query, and no if it is not.*

Generated Instruction 3 *Check if a product can function as a substitute for a given query, even if it doesn't fully meet all requirements. Output yes if it can or no otherwise.*

Generated Instruction 4 *Assess the relevance of a product to a given query by determining if it can function as a substitute, despite not fully meeting certain aspects of the query. Provide a binary output of yes or no based on this evaluation.*

Unseen Instruction *Assess whether the product is a substitute for the query and provide a yes or no response.*

C.8. Instructions on Query Product Ranking (QPR)

Seed Instruction *Given a query and a list of products denoted as A, B, C, ... with their titles, rank the products according to their relevance to the query. Output only a ranked list in which the most relevant product is at the top of the list.*

Generated Instruction 1 *Evaluate each product title in the given list, assess its relevance to the given query, and then arrange the products in descending order of relevance, with the most relevant product at the top of the ranked list.*

Generated Instruction 2 *Rank the products A, B, C, ... based on their relevance to the provided query, and produce a ranked list with the most relevant product positioned at the top of the list.*

Generated Instruction 3 *Analyze the query and each product title. Sort the products in descending order based on their relevance to the query. The most relevant product should be at the top of the list, and output the ranked list.*

Generated Instruction 4 *Evaluate the relevance of each product title in the input to the given query, and then sort the products in descending order of relevance, placing the most relevant product at the top of the ranked list.*

Unseen Instruction *Evaluate the query against each product's title, determine the relevance between the query and the product, and organize the products in descending order of relevance, ensuring that the product with the highest relevance is positioned at the top of the list.*

C.9. Instructions on Answerability Prediction (AP)

Seed Instruction *Given a question and the related document, predict if the question is answerable based on the information provided in the document. Output only yes or no.*

Generated Instruction 1 *Evaluate the answerability of a question by analyzing the related document, outputting yes if the document contains information addressing the question, and no otherwise.*

Generated Instruction 2 *Analyze a question and its supporting document. Predicting answerability based on the information provided in the document. Output yes if the document contains relevant information to answer the question, otherwise output no.*

Generated Instruction 3 *Given a question and its related document, determine if the question is answerable by analyzing the information in the document. Output yes if the document addresses the question, or no otherwise.*

Generated Instruction 4 *Output yes if the supporting document can answer the given question. Otherwise, output no.*

Unseen Instruction *Predict whether it is possible to answer the given question using the supporting document, and output a yes or no response.*

C.10. Instructions on Answer Generation (AG)

Seed Instruction *Given a question and the related document, and generate the answer to the question based on the information provided in the document.*

Generated Instruction 1 *Generate an answer to the question by utilizing the information contained in the document.*

Generated Instruction 2 *Extract information from the supporting document to answer the given question.*

Generated Instruction 3 *Answer the given question using the supporting document.*

Generated Instruction 4 *Answer the given question by extracting information from the supporting document.*

Unseen Instruction *Utilize the information provided in the supporting document to generate an answer to the given*

question.

D. Training and Evaluation Details

The training details of general-purpose and e-commerce baseline LLMs are shown in Table A5. The general-purpose LLMs undergo the 1-shot evaluation, which measures the sample with one in-context example and one test case. For each task, the 1-shot evaluation dataset is composed of all 1K test samples and the same amount of training samples (except for PM, which uses all 253 test samples and 253 training samples) from its own test and training sets. The e-commerce LLM undergoes both 1-shot and 0-shot evaluation.

Table A5. Training Details of General-purpose and E-commerce Baseline LLMs

General-purpose LLMs	URL	Accessibility
GPT-4 Turbo	https://platform.openai.com/	API
Gemini Pro	https://ai.google.dev/tutorials/python_quickstart	API
Claude 2.1	https://docs.anthropic.com/claude/reference/getting-started-with-the-api	API
Llama-2 13B-chat	https://huggingface.co/meta-llama/Llama-2-13b-chat-hf	Checkpoint
Mistral-7B-Instruct-v0.2	https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2	Checkpoint
EcomGPT	https://github.com/Alibaba-NLP/EcomGPT	Checkpoint

Table A6. Training Details of SoTA Task-specific Models

Tasks	Baseline Models	Model Source	Mode	Parameters
AVE	SUOpenTag (Xu et al., 2019)	https://github.com/hackerxiaobai/OpenTag_2019	training	epoch: 100
	AVEQA (Wang et al., 2020)	https://github.com/Zinc-30/aveqa	training	epoch: 89; batch size: 16
PRP	RGCN (Schlichtkrull et al., 2018)	https://github.com/JinheonBaek/RGCN	training	epochs: 500K
	DeBERTaV3 (He et al., 2022)	https://huggingface.co/microsoft/deberta-v3-base	fine-tuning	epoch: 3; batch size: 8
SA	BERTweet (Nguyen et al., 2020)	https://huggingface.co/finiteautomata/bertweet-base-sentiment-analysis	training	epoch: 3; batch size: 8
	DeBERTaV3 (He et al., 2022)	https://huggingface.co/microsoft/deberta-v3-base	fine-tuning	epoch: 3; batch size: 8
	P5 (Geng et al., 2022)	https://github.com/jeykigung/P5	evaluate on checkpoint	checkpoint: toy base
SR	gSASRec (Petrov & Macdonald, 2023)	https://github.com/asash/gSASRec-pytorch	training	sequence length: 50
	Recformer (Li et al., 2023)	https://github.com/AaronHeee/RecFormer	fine-tuning	epoch: 16; batch size 8
AG	GPT-4 Turbo (OpenAI, 2023)	https://platform.openai.com/	1-shot learning	prompt template in E.2
PM, MPC, PSI, QPR, AP	BERT (Devlin et al., 2019)	https://huggingface.co/bert-base-multilingual-cased	fine-tuning	epoch: 3; batch size: 8
	DeBERTaV3 (He et al., 2022)	https://huggingface.co/microsoft/deberta-v3-base	fine-tuning	epoch: 3; batch size: 8

In this table, "training" means training from scratch, "evaluate on checkpoint" means that we evaluate using the checkpoint provided by the link, "fine-tuning" means that we fine-tune the checkpoint provided by the link on specific tasks, and "1-shot learning" indicates that we directly use the model checkpoint from the link, and prompt the model with one in-context example.

The training details of SoTA task-specific models are presented in Table A6. For each task, we train and test on the SoTA task-specific models of each task using its own training, validation, and IND test sets (i.e., task-specific). For the tasks with the OOD test set, we save the trained model and test them on the OOD test set. For SUOpenTag (Xu et al., 2019) and AVEQA (Wang et al., 2020), we evaluate the AVE task according to the equation 1. Note that we do not consider MAVEQA (Yang et al., 2022) as the baseline of AVE, as it has demonstrated similar performance to AVEQA. For RGCN (Schlichtkrull et al., 2018), we construct a product-product graph, in which the nodes are products, and the edges

are specified by the relations between products (e.g., similar products). We initialize the node embeddings in the RGCN using the embeddings of product titles generated from BERT (Devlin et al., 2019). The BERT, DeBERTa, and BERTweet baselines are implemented through sentence-transformer (Reimers & Gurevych, 2019). For P5 (Geng et al., 2022), we conduct zero-shot evaluations on beauty, sports, and toys base checkpoints, and report the best result (toy base). Similar to the RGCN, we use the BERT-generated embeddings of product titles to initialize item embeddings in the gSASRec. For both gSASRec and Recformer (Li et al., 2023), we evaluate the results of the next product of interest within the candidate list as detailed in Appendix A.2.2

E. Prompt Templates

The following prompts are used for evaluating general-purpose and e-commerce LLMs. For Claude 2.1, the prompt begins with {HUMAN PROMPT} and ends with {AI PROMPT}. For Mistral-7B Instruct-v0.2, the prompt is wrapped with "[INST]".

E.1. Prompt Templates for Zero-shot Evaluation

Prompt with Options

- System prompt: Below is an instruction that describes a task. Write a response that appropriately completes the request.
- Instruction: {instruction}
  - input: {input}
  - options: {options}
  - response:

Prompt without Options

- System prompt: Below is an instruction that describes a task. Write a response that appropriately completes the request.
- Instruction: {instruction}
  - input: {input}
  - response:

E.2. Prompt Templates for One-shot Evaluation

Prompt with Options

- System prompt: Below is an instruction that describes a task, paired with an example that provides further context for the task.
- Instruction: {instruction}
- Example:
  - input: {example.input}
  - options: {example.options}
  - response: {example.response}
- Now write a response that appropriately completes the following example.
  - input: {input}
  - options: {options}
  - response:

Prompt without Options

- System prompt: Below is an instruction that describes a task, paired with an example that provides further context for the task.
- Instruction: {instruction}
- Example:
  - input: {example.input}
  - response: {example.response}
- Now write a response that appropriately completes the following example.
  - input: {input}
  - response:

F. Analysis on Base Models

Table A7. Performance of Various Base Models in IND Evaluation

Model	Base Model	AVE	PRP	PM	SA	SR	MPC	PSI	QPR	AP	AG
		F1*	Macro F1	F1	Macro F1	HR@1	Accuracy	F1	NDCG	F1	F_{BERT}
eCeLLM-L	Flan-T5 XXL	0.447	0.522	**0.995**	0.628	0.512	0.680	0.376	**0.885**	0.836	**0.844**
	Llama-2 13B-chat	**0.582**	**0.611**	**0.995**	**0.648**	**0.526**	**0.684**	**0.501**	0.870	**0.851**	0.841
eCeLLM-M	Llama-2 7B-chat	0.577	**0.562**	**0.995**	0.613	0.503	0.695	**0.444**	0.864	**0.859**	0.839
	Mistral-7B Instruct-v0.2	**0.662**	0.558	**0.995**	**0.639**	**0.542**	**0.696**	0.305	**0.876**	0.846	**0.842**
eCeLLM-S	Flan-T5 XL	0.376	0.511	**0.995**	**0.648**	0.463	**0.663**	0.042	0.868	0.827	**0.843**
	Phi-2	**0.509**	**0.518**	0.991	0.596	**0.479**	0.650	**0.392**	**0.870**	**0.846**	0.842

In this table, "Base Model" presents the base models used in eCeLLM models. On each task, the best performance of eCeLLM-L, eCeLLM-M, and eCeLLM-S when using different base models is in **bold**.

We compare the 6 base models considered for the series of eCeLLM models (i.e., eCeLLM-L, eCeLLM-M, and eCeLLM-S) and show the performance comparison in Table A7. From Table A7, we observe that Llama-2 13B-chat is the best-performing base model for eCeLLM-L. The instruction-tuned Llama-2 13B-chat model demonstrates considerable improvement compared to instruction-tuned Flan-T5 XXL on 7 out of the 10 tasks. We also observe that Mistral-7B Instruct-v0.2 and Phi-2 are the best-performing base models for eCeLLM-M and eCeLLM-S, respectively. Particularly, with instruction tuning, Mistral-7B Instruct-v0.2 achieves a notable average improvement of 2.2% over Llama-2 7B-chat across the 10 tasks. Similarly, Instruction-tuned Phi-2 also outperforms Flan-T5 XL on 6 out of the 10 tasks and achieves similar performance with Flan-T5 XL on the rest 4 tasks. The variations of trainable size and focused aspect contribute to the distinct inherent capabilities of the base models, which play a crucial role in adapting LLMs to e-commerce scenarios.

G. Complete Experimental Results

The complete results for both IND and OOD evaluations for tasks **(1)** attribute value extraction (AVE), **(2)** product relation prediction (PRP), **(3)** product matching (PM), **(4)** sentiment analysis (SA), **(5)** sequential recommendation (SR), **(6)** multi-class product classification (MPC), **(7)** product substitute identification (PSI), **(8)** query product ranking (QPR), **(9)** answerability prediction (AP), and **(10)** answer generation (AG) are presented in Table A8, A9, A10, A11, A16, A12, A13, A16, A14, and A15, respectively. Overall, these tables show that eCeLLM models fine-tuned on ECInstruct outperform the general-purpose LLMs and SoTA task-specific models in the IND test. Meanwhile, the eCeLLM exhibits good generalizability to the OOD data. Because of the high-quality ECInstruct, eCeLLM achieves remarkable performance with different base models. Note that #failed in tables represents the number of failure cases for which we cannot extract meaningful results from the model output. These failure cases are counted as wrong predictions when calculating the metrics. NPR refers to the negative prediction rate.

Regarding the hallucination issue, we observe very limited hallucination from eCeLLM. As shown in the #failed column of Table A8 and A16, in AVE and SR tasks, eCeLLM could slightly suffer from hallucination and does not output results in the

desired format for several testing samples. For the other tasks except for the AG task, as shown in Table A9-A14, eCeLLM is robust to hallucination and could always output options in a desired format. For the AG task, we randomly sample 50 answers generated by eCeLLM and do not observe hallucination with manual checks.

Table A8. Performance on AVE

Model			IND				OOD			
			Recall*	Precision*	F1*	#failed	Recall*	Precision*	F1*	#failed
General-purpose LLMs		GPT-4 Turbo	0.422	0.598	0.495	6	0.317	0.529	0.397	1
		Gemini Pro	0.318	0.523	0.396	4	0.203	0.426	0.275	6
		Claude 2.1	0.310	0.494	0.381	59	0.312	0.600	0.410	66
		Llama-2 13B-chat	0.002	0.002	0.002	0	0.000	0.000	0.000	0
		Mistral-7B-Instruct-v0.2	0.321	0.435	0.369	69	0.217	0.337	0.264	52
E-commerce LLM		EcomGPT	0.000	0.000	0.000	905	0.001	0.042	0.001	869
SoTA task-specific model		SUOpenTag	0.603	0.500	0.546	0	0.124	0.173	0.144	0
		AVEQA	0.425	0.491	0.456	0	0.283	0.257	0.269	0
eCeLLM	Task-specific	Flan-T5 XXL	0.298	0.519	0.378	7	0.362	**0.701**	0.477	0
		Llama-2 13B-chat	0.544	0.666	0.599	3	**0.448**	0.613	**0.518**	2
		Llama-2 7B-chat	0.531	0.660	0.588	1	0.323	0.499	0.392	0
		Mistral-7B Instruct-v0.2	**0.720**	**0.799**	**0.757**	5	0.374	0.544	0.443	0
		Flan-T5 XL	0.258	0.449	0.328	7	0.276	0.538	0.365	3
		Phi-2	0.304	0.570	0.397	265	0.288	0.488	0.362	0
	Generalist	Flan-T5 XXL	0.353	0.611	0.447	2	0.360	0.699	0.476	0
		Llama-2 13B-chat	0.530	0.646	0.582	1	0.276	0.425	0.335	0
		Llama-2 7B-chat	0.514	0.641	0.571	4	0.236	0.392	0.294	0
		Mistral-7B Instruct-v0.2	0.612	0.722	0.662	0	0.304	0.463	0.367	0
		Flan-T5 XL	0.297	0.514	0.376	1	0.267	0.518	0.352	1
		Phi-2	0.455	0.578	0.509	0	0.237	0.417	0.302	0

In this table, "IND" and "OOD" indicates in-domain evaluation and out-of-domain evaluation, respectively; "Task-specific" indicates that the eCeLLM models are tuned on individual tasks; "Generalist" represents tuning eCeLLM models using all tasks together. Recall*, Precision* and F1* are defined as equation 1 in Appendix A.1.1, and #failed refers to the number of failure cases that we cannot extract meaningful results from the model output. On each task, the best performance is in **bold**.

H. OOD Evaluation

The following parts discuss the OOD results under different settings.

H.1. Results on Unseen Instructions

Table A17 exhibits the results of evaluating eCeLLM on OOD test sets with unseen instruction. As shown in table A17, eCeLLM models perform better when training on ECInstruct with diverse instructions. eCeLLM fine-tuned on ECInstruct with diverse instruction enhances the generalizability of models to the unseen instruction.

H.2. Results on Different Base Models

Table A18 shows the OOD evaluation results of eCeLLM with different base models. With high-quality ECInstruct dataset, eCeLLM models achieve good performance under different base models.

H.3. Results on Models Tuned using All Tasks and Individual Tasks

As demonstrated in Table A19, when generalizing eCeLLM to OOD test, eCeLLM fine-tuned on every single task can better transfer the knowledge to the specific task than eCeLLM fine-tuned on the task combination.

Table A9. Performance on PRP

Model		IND					OOD				
		Accuracy	M-Rec	M-Pre	M-F1	#failed	Accuracy	M-Rec	M-Pre	M-F1	#failed
General-purpose LLMs	GPT-4 Turbo	0.384	0.487	0.381	0.326	0	0.488	0.496	0.392	0.392	0
	Gemini Pro	0.128	0.385	0.352	0.136	1	0.147	0.359	0.390	0.123	0
	Claude 2.1	0.508	0.347	0.344	0.275	10	0.362	0.394	0.400	0.277	4
	Llama-2 13B-chat	0.473	0.333	0.333	0.333	0	0.419	0.338	0.339	0.324	0
	Mistral-7B-Instruct-v0.2	0.442	0.323	0.325	0.324	0	0.422	0.338	0.351	0.327	0
E-commerce LLM	EcomGPT	0.147	0.101	0.101	0.091	444	0.125	0.125	0.092	0.096	455
SoTA task-specific model	DeBERTaV3	0.762	0.575	0.620	0.588	0	0.658	0.514	0.570	0.507	0
	RGCN	0.615	**0.665**	0.637	0.506	0	0.576	0.373	0.372	0.356	0
eCeLLM Task-specific	Flan-T5 XXL	0.754	0.516	0.511	0.508	0	0.663	0.506	0.468	0.466	0
	Llama-2 13B-chat	0.769	0.530	0.517	0.521	0	0.690	0.520	0.472	0.483	0
	Llama-2 7B-chat	0.774	0.541	0.628	0.537	0	0.695	0.526	0.803	0.498	0
	Mistral-7B Instruct-v0.2	0.782	0.547	**0.689**	0.543	0	0.711	0.532	**0.808**	0.502	0
	Flan-T5 XL	0.704	0.467	0.496	0.460	0	0.592	0.471	0.625	0.427	0
	Phi-2	0.584	0.372	0.379	0.348	0	0.406	0.349	0.334	0.251	0
eCeLLM Generalist	Flan-T5 XXL	0.769	0.531	0.517	0.522	0	0.703	0.533	0.648	0.499	0
	Llama-2 13B-chat	0.775	0.599	0.635	**0.611**	0	**0.726**	**0.564**	0.611	**0.558**	0
	Llama-2 7B-chat	**0.797**	0.586	0.661	0.595	0	0.703	0.533	0.648	0.499	0
	Mistral-7B Instruct-v0.2	0.788	0.555	0.644	0.558	0	0.707	0.537	0.596	0.502	0
	Flan-T5 XL	0.757	0.517	0.515	0.511	0	0.678	0.521	0.587	0.489	0
	Phi-2	0.747	0.524	0.552	0.518	0	0.710	0.541	0.611	0.520	0

In this table, "IND" and "OOD" indicates in-domain evaluation and out-of-domain evaluation, respectively; "Task-specific" indicates that the eCeLLM models are tuned on individual tasks; "Generalist" represents tuning eCeLLM models using all tasks together; #failed refers to the number of failure cases that we cannot extract meaningful results from the model output. The metric "M-Rec", "M-Pre", and "M-F1" represents macro recall, macro precision, and macro F1, respectively. On each task, the best performance is in **bold**.

Table A10. Performance on PM

Model		IND						
		Accuracy	Recall	Precision	F1	Specificity	NPR	#failed
General-purpose LLMs	GPT-4 Turbo	0.826	0.604	**1.000**	0.753	**1.000**	0.763	0
	Gemini Pro	0.897	0.766	**1.000**	0.867	**1.000**	0.845	0
	Claude 2.1	0.711	0.360	0.952	0.523	0.986	0.664	1
	Llama-2 13B-chat	0.474	0.459	0.411	0.434	0.486	0.535	0
	Mistral-7B-Instruct-v0.2	0.755	0.441	**1.000**	0.613	**1.000**	0.696	0
E-commerce LLM	EcomGPT	0.648	0.739	0.577	0.648	0.577	0.739	0
SoTA task-specific model	BERT	**0.996**	0.991	**1.000**	**0.995**	**1.000**	0.993	0
	DeBERTaV3	0.577	**1.000**	0.509	0.675	0.246	**1.000**	0
eCeLLM Task-specific	Flan-T5 XXL	**0.996**	0.991	**1.000**	**0.995**	**1.000**	0.993	0
	Llama-2 13B-chat	**0.996**	0.991	**1.000**	**0.995**	**1.000**	0.993	0
	Llama-2 7B-chat	0.992	0.991	0.991	0.991	0.993	0.993	0
	Mistral-7B Instruct-v0.2	0.988	0.991	0.982	0.987	0.986	0.993	0
	Flan-T5 XL	0.960	0.910	**1.000**	0.953	**1.000**	0.934	0
	Phi-2	0.992	0.991	0.991	0.991	0.993	0.993	0
eCeLLM Generalist	Flan-T5 XXL	**0.996**	0.991	**1.000**	**0.995**	**1.000**	0.993	0
	Llama-2 13B-chat	**0.996**	0.991	**1.000**	**0.995**	**1.000**	0.993	0
	Llama-2 7B-chat	**0.996**	0.991	**1.000**	**0.995**	**1.000**	0.993	0
	Mistral-7B Instruct-v0.2	**0.996**	0.991	**1.000**	**0.995**	**1.000**	0.993	0
	Flan-T5 XL	**0.996**	0.991	**1.000**	**0.995**	**1.000**	0.993	0
	Phi-2	0.992	0.991	0.991	0.991	0.993	0.993	0

In this table, "IND" indicates in-domain evaluation; "Task-specific" indicates that the eCeLLM models are tuned on individual tasks; "Generalist" represents tuning eCeLLM models using all tasks together; #failed refers to the number of failure cases that we cannot extract meaningful results from the model output; "NPR" is the negative prediction rate. On each task, the best performance is in **bold**.

Table A11. Performance on SA

Model		IND					OOD				
		Acc	M-Rec	M-Pre	M-F1	#failed	Acc	M-Rec	M-Pre	M-F1	#failed
General-purpose LLMs	GPT-4 Turbo	0.595	0.575	0.544	0.516	0	0.556	0.586	0.544	0.510	0
	Gemini Pro	0.609	0.521	0.453	0.470	2	0.572	0.511	0.444	0.454	1
	Claude 2.1	0.375	0.510	0.474	0.415	2	0.328	0.466	0.447	0.369	1
	Llama-2 13B-chat	0.406	0.188	0.191	0.188	0	0.384	0.179	0.180	0.178	0
	Mistral-7B-Instruct-v0.2	0.633	0.532	0.551	0.470	0	0.594	0.531	0.494	0.438	0
E-commerce LLM	EcomGPT	0.191	0.362	0.341	0.188	6	0.196	0.375	0.336	0.178	13
SoTA task-specific model	BERTweet	0.733	0.503	0.530	0.511	0	0.729	0.507	0.524	0.513	0
	DeBERTaV3	0.768	0.567	0.607	0.573	0	0.764	0.565	0.591	0.567	0
	P5	0.611	0.199	0.157	0.156	0	0.620	0.200	0.124	0.153	0
eCeLLM Task-specific	Flan-T5 XXL	0.783	0.619	0.618	0.612	0	0.770	0.604	0.601	0.600	0
	Llama-2 13B-chat	0.791	0.616	0.641	0.616	0	0.781	0.627	0.645	0.629	0
	Llama-2 7B-chat	0.790	0.620	0.652	0.634	0	0.769	0.583	0.599	0.589	0
	Mistral-7B Instruct-v0.2	**0.801**	0.643	**0.676**	**0.655**	0	**0.789**	0.619	0.650	0.632	0
	Flan-T5 XL	0.771	0.645	0.638	0.620	0	0.743	0.594	0.592	0.582	0
	Phi-2	0.779	0.611	0.618	0.608	0	0.754	0.576	0.594	0.583	0
eCeLLM Generalist	Flan-T5 XXL	0.797	0.629	0.646	0.628	0	0.787	0.619	0.624	0.619	0
	Llama-2 13B-chat	0.796	0.641	0.661	0.648	0	0.785	0.621	0.638	0.629	0
	Llama-2 7B-chat	0.768	0.579	0.589	0.580	0	0.776	0.599	0.626	0.606	0
	Mistral-7B Instruct-v0.2	0.781	0.630	0.654	0.639	0	0.784	**0.630**	**0.653**	**0.640**	0
	Flan-T5 XL	0.782	**0.654**	0.655	0.648	0	0.753	0.604	0.598	0.598	0
	Phi-2	0.780	0.588	0.619	0.596	0	0.758	0.552	0.590	0.565	0

In this table, "IND" and "OOD" indicates in-domain evaluation and out-of-domain evaluation, respectively; "Task-specific" indicates that the eCeLLM models are tuned on individual tasks; "Generalist" represents tuning eCeLLM models using all tasks together; #failed refers to the number of failure cases that we cannot extract meaningful results from the model output. The metric "Acc", "M-Rec", "M-Pre", and "M-F1" represents accuracy, macro recall, macro precision, and macro F1, respectively. On each task, the best performance is in **bold**.

I. Complete Results for the Analysis of Training Data Size

This section exhibits the comprehensive results for analyzing how the training data size influences the performance of eCeLLM models. With the total 92K samples in ECInstruct, we assess data scaling with sizes of 1K, 10K, and 47K. The 1K and 10K samples are collected by randomly selecting 0.1K and 1K samples, respectively, from the training set of each of the 10 tasks. The 47K samples are assembled by randomly selecting 5K samples from each task except for PM, for which all its 2K training samples are included. As shown in Table A20 and Table A21, performances exhibit an upward trend along with the increase in data size. Notably, the F1 score of the PSI task experiences an initial drop followed by a subsequent rise with the data size increasing. This behavior could be attributed to the imbalanced PSI data, making the models tend to randomly guess when the data size is small. As the data size increases, the models would predict the dominant label, leading to a temporarily low F1 score. With continued data growth, the models acquire sufficient knowledge to generate accurate predictions. In general, the large-scale training data plays a crucial role in developing effective e-commerce LLMs, underscoring the significance of our extensive, comprehensive, and high-quality e-commerce instruction dataset ECInstruct.

Table A12. Performance on MPC

Model			IND				
			Accuracy	M-Rec	M-Pre	M-F1	#failed
General-purpose LLMs		GPT-4 Turbo	0.611	**0.527**	**0.540**	**0.487**	0
		Gemini Pro	0.584	0.471	0.414	0.425	2
		Claude 2.1	0.655	0.464	0.419	0.435	13
		Llama-2 13B-chat	0.504	0.250	0.251	0.250	0
		Mistral-7B-Instruct-v0.2	0.529	0.395	0.384	0.365	0
E-commerce LLM		EcomGPT	0.540	0.265	0.218	0.223	2
SoTA task-specific model		BERT	0.661	0.381	0.423	0.393	0
		DeBERTaV3	**0.703**	0.436	0.472	0.448	0
eCeLLM	Task-specific	Flan-T5 XXL	0.666	0.438	0.412	0.346	0
		Llama-2 13B-chat	0.655	0.399	0.410	0.349	0
		Llama-2 7B-chat	0.659	0.399	0.531	0.330	0
		Mistral-7B Instruct-v0.2	0.681	0.406	0.423	0.387	0
		Flan-T5 XL	0.648	0.425	0.361	0.327	0
		Phi-2	0.646	0.387	0.316	0.321	0
	Generalist	Flan-T5 XXL	0.680	0.431	0.416	0.364	0
		Llama-2 13B-chat	0.684	0.440	0.435	0.414	0
		Llama-2 7B-chat	0.679	0.427	0.434	0.398	0
		Mistral-7B Instruct-v0.2	0.696	0.450	0.456	0.443	0
		Flan-T5 XL	0.663	0.395	0.533	0.332	0
		Phi-2	0.650	0.397	0.410	0.335	0

In this table, "IND" indicates in-domain evaluation; "Task-specific" indicates that the eCeLLM models are tuned on individual tasks; "Generalist" represents tuning eCeLLM models using all tasks together; #failed refers to the number of failure cases that we cannot extract meaningful results from the model output. The metric "M-Rec", "M-Pre", and "M-F1" represents macro recall, macro precision, and macro F1, respectively. On each task, the best performance is in **bold**.

Table A13. Performance on PSI

Model			IND						
			Accuracy	Recall	Precision	F1	Specificity	NPR	#failed
General-purpose LLMs		GPT-4 Turbo	0.289	0.374	0.132	0.195	0.264	0.585	0
		Gemini Pro	0.296	0.504	0.164	0.248	0.234	0.612	0
		Claude 2.1	0.291	0.578	0.179	0.273	0.205	0.620	1
		Llama-2 13B-chat	0.649	0.257	0.247	0.252	0.766	0.775	0
		Mistral-7B-Instruct-v0.2	0.361	**0.609**	0.203	0.305	0.287	0.711	0
E-commerce LLM		EcomGPT	0.630	0.165	0.176	0.170	0.769	0.755	13
SoTA task-specific model		BERT	0.761	0.330	0.472	0.389	0.890	0.816	0
		DeBERTaV3	0.769	0.000	0.000	0.000	0.999	0.770	0
eCeLLM	Task-specific	Flan-T5 XXL	0.766	0.013	0.300	0.025	0.991	0.771	0
		Llama-2 13B-chat	0.770	0.000	0.000	0.000	**1.000**	0.770	0
		Llama-2 7B-chat	0.770	0.017	0.500	0.034	0.995	0.772	0
		Mistral-7B Instruct-v0.2	0.770	0.000	0.000	0.000	**1.000**	0.770	0
		Flan-T5 XL	0.768	0.000	0.000	0.000	0.997	0.770	0
		Phi-2	0.770	0.000	0.000	0.000	**1.000**	0.770	0
	Generalist	Flan-T5 XXL	0.771	0.300	0.504	0.376	0.912	0.813	0
		Llama-2 13B-chat	**0.795**	0.448	0.569	**0.501**	0.899	**0.845**	0
		Llama-2 7B-chat	0.781	0.283	0.546	0.372	0.930	0.813	0
		Mistral-7B Instruct-v0.2	0.790	0.200	0.639	0.305	0.966	0.802	0
		Flan-T5 XL	0.773	0.022	**0.714**	0.042	0.997	0.773	0
		Phi-2	0.777	0.313	0.526	0.392	0.916	0.817	0

In this table, "IND" indicates in-domain evaluation; "Task-specific" indicates that the eCeLLM models are tuned on individual tasks; "Generalist" represents tuning eCeLLM models using all tasks together; #failed refers to the number of failure cases that we cannot extract meaningful results from the model output; "NPR" is the negative prediction rate. On each task, the best performance is in **bold**.

Table A14. Performance on AP

Model		IND							OOD						
		Accuracy	Recall	Precision	F1	Specificity	NPR	#failed	Accuracy	Recall	Precision	F1	Specificity	NPR	#failed
General-purpose LLMs	GPT-4 Turbo	0.623	0.550	0.791	0.649	0.749	0.491	5	0.641	0.555	**0.876**	0.680	0.828	0.460	2
	Gemini Pro	0.542	0.371	0.797	0.506	0.837	0.435	13	0.543	0.410	0.844	0.552	0.834	0.393	6
	Claude 2.1	0.424	0.177	0.671	0.280	**0.850**	0.375	90	0.384	0.146	0.769	0.245	**0.904**	0.326	73
	Llama-2 13B-chat	0.534	0.608	0.638	0.623	0.406	0.375	0	0.541	0.605	0.688	0.644	0.401	0.317	0
	Mistral-7B-Instruct-v0.2	0.522	0.539	0.647	0.588	0.493	0.383	120	0.537	0.523	0.725	0.608	0.567	0.352	126
E-commerce LLM	EcomGPT	0.318	0.051	0.283	0.086	0.779	0.322	254	0.286	0.085	0.403	0.140	0.726	0.266	245
SoTA task-specific model	BERT	0.749	**0.970**	0.726	0.830	0.368	**0.877**	0	0.803	0.831	**0.876**	0.853	0.742	0.668	0
	DeBERTaV3	0.504	0.310	0.769	0.441	0.839	0.413	0	0.501	0.345	0.826	0.487	0.841	0.370	0
eCeLLM Task-specific	Flan-T5 XXL	0.749	0.875	0.763	0.815	0.531	0.712	0	0.799	0.889	0.830	0.859	0.602	0.713	0
	Llama-2 13B-chat	0.801	0.919	0.797	0.854	0.597	0.811	0	**0.838**	0.924	0.852	**0.887**	0.650	0.797	0
	Llama-2 7B-chat	0.741	0.889	0.749	0.813	0.485	0.718	0	0.761	0.864	0.802	0.832	0.535	0.644	0
	Mistral-7B Instruct-v0.2	**0.821**	0.896	0.834	0.864	0.692	0.794	0	0.835	0.891	0.872	0.881	0.713	0.749	0
	Flan-T5 XL	0.693	0.684	0.802	0.738	0.708	0.565	0	0.707	0.682	0.862	0.762	0.761	0.523	0
	Phi-2	0.765	0.942	0.751	0.835	0.460	0.820	0	0.781	0.950	0.779	0.856	0.411	0.791	0
eCeLLM Generalist	Flan-T5 XXL	0.774	0.910	0.773	0.836	0.540	0.776	0	0.814	0.914	0.832	0.871	0.596	0.760	0
	Llama-2 13B-chat	0.808	0.864	**0.838**	0.851	0.711	0.752	0	0.818	0.862	0.872	0.867	0.723	0.705	0
	Llama-2 7B-chat	0.817	0.921	0.814	**0.864**	0.638	0.824	0	0.836	0.931	0.845	0.886	0.627	**0.807**	0
	Mistral-7B Instruct-v0.2	0.797	0.880	0.814	0.846	0.654	0.759	0	0.832	0.885	0.872	0.878	0.717	0.740	0
	Flan-T5 XL	0.765	0.888	0.774	0.827	0.553	0.741	0	0.819	0.891	0.852	0.871	0.662	0.735	0
	Phi-2	0.794	0.897	0.801	0.846	0.616	0.777	0	0.823	**0.937**	0.828	0.879	0.573	**0.807**	0

In this table, "IND" and "OOD" indicates in-domain evaluation and out-of-domain evaluation, respectively; "Task-specific" indicates that the eCeLLM models are tuned on individual tasks; "Generalist" represents tuning eCeLLM models using all tasks together; #failed refers to the number of failure cases that we cannot extract meaningful results from the model output; "NPR" is the negative prediction rate. On each task, the best performance is in bold.

Table A15. Performance on AG

Model		IND					OOD				
		R_{BERT}	P_{BERT}	F_{BERT}	BLEURT	#failed	R_{BERT}	P_{BERT}	F_{BERT}	BLEURT	#failed
General-purpose LLMs	GPT-4 Turbo	0.847	**0.869**	**0.858**	0.280	0	**0.852**	**0.868**	**0.860**	0.283	0
	Gemini Pro	0.844	0.867	0.855	0.269	0	0.847	0.866	0.856	0.264	0
	Claude 2.1	0.848	0.835	0.841	**0.314**	0	0.851	0.833	0.842	**0.325**	0
	Llama-2 13B-chat	0.845	0.780	0.811	0.261	0	0.845	0.775	0.808	0.260	0
	Mistral-7B-Instruct-v0.2	**0.850**	0.856	0.853	0.288	0	**0.852**	0.851	0.851	0.290	0
E-commerce LLM	EcomGPT	0.675	0.665	0.669	0.290	0	0.729	0.716	0.722	0.296	0
SoTA task-specific model	GPT-4 Turbo	0.847	**0.869**	**0.858**	0.280	0	**0.852**	**0.868**	**0.860**	0.283	0
eCeLLM Task-specific	Flan-T5 XXL	0.822	0.864	0.842	0.310	0	0.824	0.865	0.843	0.302	0
	Llama-2 13B-chat	0.824	0.861	0.841	0.309	0	0.821	0.860	0.840	0.289	0
	Llama-2 7B-chat	0.822	0.861	0.841	0.301	0	0.820	0.861	0.840	0.289	0
	Mistral-7B Instruct-v0.2	0.823	0.860	0.841	0.310	0	0.823	0.861	0.842	0.298	0
	Flan-T5 XL	0.823	0.864	0.843	0.320	0	0.824	0.864	0.843	0.307	0
	Phi-2	0.817	0.855	0.835	0.283	0	0.817	0.856	0.835	0.270	0
eCeLLM Generalist	Flan-T5 XXL	0.824	0.865	0.844	0.224	0	0.823	0.864	0.843	0.206	0
	Llama-2 13B-chat	0.823	0.861	0.841	0.215	0	0.822	0.861	0.841	0.195	0
	Llama-2 7B-chat	0.822	0.860	0.840	0.208	0	0.819	0.859	0.838	0.188	0
	Mistral-7B Instruct-v0.2	0.822	0.864	0.842	0.213	0	0.821	0.862	0.840	0.194	0
	Flan-T5 XL	0.823	0.864	0.843	0.227	0	0.824	0.865	0.844	0.211	0
	Phi-2	0.823	0.861	0.842	0.222	0	0.821	0.859	0.840	0.198	0

In this table, "IND" and "OOD" indicates in-domain evaluation and out-of-domain evaluation, respectively; "Task-specific" indicates that the eCeLLM models are tuned on individual tasks; "Generalist" represents tuning eCeLLM models using all tasks together; #failed refers to the number of failure cases that we cannot extract meaningful results from the model output. The metrics "R_{BERT}", "P_{BERT}", "F_{BERT}" and "BLEURT" are detailed in Table A1. On each task, the best performance is in **bold**. Note that we use GPT-4 Turbo as the SoTA task-specific model in this task.

Table A16. Performance on SR and QPR

Model			SR				QPR	
			IND		OOD		IND	
			HR@1	#failed	HR@1	#failed	NDCG	#failed
General-purpose LLMs		GPT-4 Turbo	0.387	0	0.198	0	0.875	14
		Gemini Pro	0.269	2	0.116	3	0.821	52
		Claude 2.1	0.066	34	0.036	42	0.821	26
		Llama-2 13B-chat	0.056	0	0.050	0	0.815	0
		Mistral-7B-Instruct-v0.2	0.164	1	0.108	0	0.842	4
E-commerce LLM		EcomGPT	0.042	344	0.023	391	0.000	1000
SoTA task-specific model		gSASRec / BERT	0.249	0	0.065	0	0.839	0
		Recformer / DeBERTaV3	0.265	0	0.081	0	0.859	0
eCeLLM	Task-specific	Flan-T5 XXL	0.467	0	0.252	0	0.881	0
		Llama-2 13B-chat	0.518	0	0.263	0	0.879	0
		Llama-2 7B-chat	0.517	0	0.228	0	0.867	0
		Mistral-7B Instruct-v0.2	0.535	0	0.268	0	0.883	0
		Flan-T5 XL	0.436	0	0.226	0	0.875	0
		Phi-2	0.413	5	0.219	10	0.858	0
	Generalist	Flan-T5 XXL	0.512	0	0.262	0	**0.885**	0
		Llama-2 13B-chat	0.526	0	0.273	0	0.870	0
		Llama-2 7B-chat	0.517	0	0.261	0	0.868	0
		Mistral-7B Instruct-v0.2	**0.542**	0	**0.280**	0	0.876	0
		Flan-T5 XL	0.463	0	0.256	0	0.868	0
		Phi-2	0.479	5	0.241	8	0.870	0

In this table, "IND" and "OOD" indicates in-domain evaluation and out-of-domain evaluation, respectively; "Task-specific" indicates that the eCeLLM models are tuned on individual tasks; "Generalist" represents tuning eCeLLM models using all tasks together; #failed refers to the number of failure cases that we cannot extract meaningful results from the model output. The metrics "HR@1" and "NDCG" are detailed in Appendix A. On each task, the best performance is in **bold**.

Table A17. Performance on Unseen Instructions in OOD Evaluation

Model	Training Instructions	AVE	PRP	SA	SR	AP	AG
		F1*	Macro F1	Macro F1	HR@1	F1	F_{BERT}
eCeLLM-L	single	0.001	0.561	0.636	0.260	**0.890**	0.839
	diverse	**0.276**	**0.577**	**0.652**	**0.266**	0.877	**0.840**
eCeLLM-M	single	0.000	**0.527**	0.584	**0.284**	**0.877**	**0.849**
	diverse	**0.366**	0.507	**0.628**	0.275	0.863	0.841
eCeLLM-S	single	**0.305**	0.497	0.555	**0.249**	0.866	0.838
	diverse	0.275	**0.513**	**0.574**	0.248	**0.880**	**0.841**

In this table, "single" and "diverse" indicate that the eCeLLM models are tuned over single and diverse instructions, respectively. The best performance of each eCeLLM model tuned over single and diverse instructions on each task is in **bold**.

Table A18. Performance on Various Base Models in OOD Evaluation

Model	Base Model	AVE	PRP	SA	SR	AP	AG
		F1*	Macro F1	Macro F1	HR@1	F1	F_{BERT}
eCeLLM-L	Flan-T5 XXL	**0.476**	0.499	0.619	0.262	**0.871**	**0.843**
	Llama-2 13B-chat	0.335	**0.558**	**0.629**	**0.273**	0.867	0.841
eCeLLM-M	Llama-2 7B-chat	0.314	**0.511**	0.618	0.266	**0.894**	0.837
	Mistral-7B Instruct-v0.2	**0.367**	0.502	**0.640**	**0.280**	0.878	**0.840**
eCeLLM-S	Flan-T5 XL	**0.352**	0.489	**0.598**	**0.256**	0.871	**0.844**
	Phi-2	0.302	**0.520**	0.565	0.241	**0.879**	0.840

In this table, "Base Model" presents the base models used in eCeLLM models. On each task, the best performance of eCeLLM-L, eCeLLM-M, and eCeLLM-S when using different base models is in **bold**.

Table A19. Performance of Generalist and Task-specific eCeLLM Models in OOD Evaluation

Model	Training Tasks	AVE	PRP	SA	SR	AP	AG
		F1*	Macro F1	Macro F1	HR@1	F1	F_{BERT}
eCeLLM-L	Task-specific	**0.518**	0.483	**0.629**	0.263	**0.887**	0.840
	Generalist	0.335	**0.558**	**0.629**	**0.273**	0.867	**0.841**
eCeLLM-M	Task-specific	**0.443**	**0.502**	0.632	0.268	**0.881**	**0.842**
	Generalist	0.367	**0.502**	**0.640**	**0.280**	0.878	0.840
eCeLLM-S	Task-specific	**0.362**	0.251	**0.583**	0.219	0.856	0.835
	Generalist	0.302	**0.520**	0.565	**0.241**	**0.879**	**0.840**

In this table, "Task-specific" indicates that the eCeLLM models are tuned on individual tasks; "Generalist" represents tuning eCeLLM models using all tasks together. The best performance of generalist and task-specific eCeLLM models on each task is in **bold**.

Table A20. Performance with Different Data Sizes in IND Evaluation

Model	Data Size	AVE	PRP	PM	SA	SR	MPC	PSI	QPR	AP	AG
		F1*	Macro F1	F1	Macro F1	HR@1	Accuracy	F1	NDCG	F1	F_{BERT}
eCeLLM-L	1K	0.000	0.309	0.874	0.309	0.085	0.576	0.194	0.803	0.632	0.813
	10K	0.391	0.500	**0.995**	0.601	0.445	0.656	0.009	0.856	0.806	0.843
	47K	0.544	0.549	**0.995**	0.618	0.506	0.675	0.424	0.869	**0.854**	**0.844**
	92K	**0.582**	**0.611**	**0.995**	**0.648**	**0.526**	**0.684**	**0.501**	**0.870**	0.851	0.841
eCeLLM-M	1K	0.046	0.327	0.982	0.550	0.318	0.633	0.156	0.817	0.792	0.828
	10K	0.478	0.534	0.991	0.595	0.468	0.661	0.252	**0.877**	0.804	0.842
	47K	0.618	**0.610**	**0.995**	**0.639**	0.507	0.672	**0.404**	0.872	0.843	**0.846**
	92K	**0.662**	0.558	**0.995**	**0.639**	**0.542**	**0.696**	0.305	0.876	**0.846**	0.842
eCeLLM-S	1K	0.000	0.296	0.411	0.286	0.046	0.507	0.356	0.745	0.767	0.748
	10K	0.223	0.330	0.987	0.510	0.287	0.636	0.000	0.838	0.772	0.840
	47K	0.311	0.503	**0.995**	0.571	0.422	0.653	0.017	0.863	0.837	0.837
	92K	**0.509**	**0.518**	0.991	**0.596**	**0.479**	**0.650**	**0.392**	**0.870**	**0.846**	**0.842**

The best performance of eCeLLM-L, eCeLLM-M, and eCeLLM-S over different data sizes is in **bold**.

Table A21. Performance with Different Data Sizes in OOD Evaluation

Model	Data Size	AVE	PRP	SA	SR	AP	AG
		F1*	Macro F1	Macro F1	HR@1	F1	F_{BERT}
eCeLLM-L	1K	0.000	0.299	0.329	0.059	0.547	0.818
	10K	**0.389**	0.464	0.602	0.227	0.833	**0.842**
	47K	0.346	0.529	0.628	0.271	**0.880**	0.841
	92K	0.335	**0.558**	**0.629**	**0.273**	0.867	0.841
eCeLLM-M	1K	0.082	0.304	0.523	0.156	0.833	0.830
	10K	0.356	0.483	0.570	0.261	0.826	0.841
	47K	**0.401**	**0.529**	**0.650**	0.267	**0.891**	**0.846**
	92K	0.367	0.502	0.640	**0.280**	0.878	0.840
eCeLLM-S	1K	0.000	0.278	0.278	0.052	0.801	0.758
	10K	0.278	0.296	0.490	0.170	0.794	**0.848**
	47K	**0.317**	0.479	0.537	0.217	0.855	0.839
	92K	0.302	**0.520**	**0.565**	**0.241**	**0.879**	0.840

The best performance of eCeLLM-L, eCeLLM-M, and eCeLLM-S over different data sizes is in **bold**.

J. Case Study

We show the predictions from eCeLLM-L and the best-performing general-purpose LLM GPT-4 Turbo on each of the 10 tasks. From the following examples, we note that by instruction tuning using our high-quality ECInstruct, eCeLLM-L could outperform GPT-4 Turbo in distinguishing products (e.g. PM), estimating users' preferences (e.g. SA), and understanding products (e.g. AVE). For example, as shown in Section J.2, eCeLLM-L could accurately predict whether two products are the same given their title, description, manufacturer, and price, while GPT-4 Turbo struggles. In addition, the example in Section J.5 indicates that compared to GPT-4 Turbo, eCeLLM-L could better estimate users' preferences based on their historical interactions. Moreover, the example in Section J.6 shows that eCeLLM-L could correctly understand that DJ light is equivalent to light that will change based on the music. In contrast, without instruction tuning, GPT-4 Turbo may lack this knowledge and fail the example.

J.1. Attribute Value Extraction (AVE)

- Instruction:
  - Given the title, description, feature, price, and brand of a product and a set of target attributes, extract the value of each target at tribute from the product information. Output the extracted value and the corresponding source (e.g., title or feature) denoting where the value is extracted.
- Input:
  - Product title: LibbySue-Airy, Lightweight Knit Winter Scarf in Opulent Earth Tones
  - Product description: The lightweight weave of this soft and airy knit scarf will keep you stylishly ready for any brisk day
  - Product feature: 100% Acrylic, A lightweight weave, Soft and airy knit scarf, Will keep you stylishly ready for any brisk day, Opulent Earth Tones, 78 inches long by 12 inches wide with 7 inch fringe, Package Dimensions: 5.8 x 5.7 x 3.9 inches, Shipping Weight: 4 ounces, ASIN: B00FM6HPAU, Date first listed on Amazon: November 6, 2011, Average Customer Review: 5.0 out of 5 stars 1 customer review
  - Product brand: LibbySue
  - Target attributes: Specialty Style
- Output:
  - Ground truth: Attribute: specialty style; Value: None, Source: None.
  - eCeLLM: Attribute: specialty style; Value: None, Source: None.
  - GPT-4 Turbo: Attribute: specialty style; Value: airy, lightweight knit, Source: product title.

J.2. Product Matching (PM)

- Instruction:
  - Compare the details of two given products to determine if they are identical. Output yes if they are identical or no otherwise.
- Input:
  - Product 1: title - omniweb 5.0, description - NaN, manufacturer - omni group, price - 29.99
  - Product 2: title - omni web 5.0, description - sure you can use a standard web browser with standard features. but you didn't choose a standard software experience - you chose the mac. why not try a browser built just for discriminating people with fabulous taste like yourself? omniweb5..., manufacturer - NaN, price - 23.99
- Output:
  - Ground truth: Yes
  - eCeLLM: Yes
  - GPT-4 Turbo: No

J.3. Product Relation Prediction (PRP)

- Instruction:
  - Given the title of two products, predict if the two products are similar, if the two products will be purchased or viewed together. Answer only from the options.
- Input:
  - Product 1: Kenable Internal Memory Card Reader for 5.25 CD/DVD Bay With USB Port BLACK
  - Product 2: CORSAIR Carbide 100R Mid-Tower Case
- Options:
  - A: Users who buy product 1 may also buy product 2.
  - B: Users who view product 1 may also view product 2.
  - C: The product 1 is similar with the product 2.
- Output:
  - Ground truth: B
  - eCeLLM: B
  - GPT-4 Turbo: A

J.4. Sentiment Analysis (SA)

- Instruction:
  - Carefully assess the user's review for any strong expressions of sentiment, either positive or negative. Based on your analysis, select the most fitting sentiment option from the provided list as output.
- Input:
  - This visor CD holder is a great addition for any car. It folds shut so CDs don't slide out. It fits well on any visor.
- Options:
  - A: very positive
  - B: positive
  - C: neutral
  - D: negative
  - E: very negative
- Output:
  - Ground truth: A
  - eCeLLM: A
  - GPT-4 Turbo: B

J.5. Sequential Recommendation (SR)

- Instruction:
  - Given the products the user has purchased in history, rank the items in the listed options and output the item that the user is most likely to purchase next. Answer from one of the options.
- Input:
  - 1st: Bbox A392-10CP Dual 10" Sealed Carpeted Subwoofer Enclosure - Fits 1999-2007 Ford F250/350/450 Crew Cab. Electronics. Car & Vehicle Electronics. b.box.
  - 2nd: Smatree Batteries Charger Kit for GoPro Hero 1/2 Digital Camera. Electronics. Camera & Photo. Smatree.

  - 3rd: Dolica WT-1003 67-Inch Lightweight Monopod. Electronics. Camera & Photo. Dolica.
  - 4th: SunFounder Sidekick Basic Starter Kit w/Breadboard, Jumper wires, Color Led, Resistors, Buzzer For Arduino UNO R3 Mega2560 Mega328 Nano - Including 42 Page Instructions Book...
  - 5th: Winait 5MP Mini 5mp Worlds Smallest Hd Digital Video Camera Spy Camera Video Recorder Hidden Cam DV DVR with 1280 X 960 Resolution. Electronics. Camera...
  - 6th: RoadPro RPPS-220 Platinum Series 12V 3-Pin Plug Fused Replacement CB Power Cord. Electronics. Accessories & Supplies. RoadPro.
  - 7th: Transcend USB 3.0 SDHC / SDXC / microSDHC / SDXC Card Reader, TS-RDF5K (Black). Electronics. Computers & Accessories. Transcend.
- Options:
  - A: Sony Bloggie Live(MHS-TS55) Video Camera with 4x Digital Zoom, 3.0-Inch Touchscreen LCD and WiFi Connectivity (2012 Model). Electronics. Camera & Photo. Sony.
  - B: Gold Tip Expedition Hunter 5575 Dozen Black Shafts. Sports & Outdoors. Sports & Fitness. Gold Tip.
  - C: One Direction Purple Zebra Print Fleece Throw Blanket. Home & Kitchen. Bedding. 1D Media Ltd.
  - D: Absolute DAG15 Dual 15-Inch Angle Ported MDF Enclosure. Electronics. Car & Vehicle Electronics. Absolute.
  - E: Adesso iMouseE1 - Vertical Ergonomic Illuminated Optical 6-Button USB Mouse - Right Hand Orientation. Electronics. Computers & Accessories. Adesso.
  - F: AGPtek AC Power Home Wall Charger Adapter For Microsoft Surface Windows RT Surface 2 Tablet. Electronics. Computers & Accessories. AGPTEK.
  - ...
  - O: SanDisk Ultra 32GB UHS-I/Class 10 Micro SDHC Memory Card With Adapter - SDSDQUAN-032G-G4A. Electronics. Computers & Accessories. SanDisk.
  - P: UCEC 5.25 Inch Front Panel USB Hub with 2-Port USB 3.0 & 2-Port USB 2.0 & HD Audio Output Port & Microphone Input Port for...
  - Q: GreenLife Soft Grip 11” Ceramic Non-Stick Open Wok, Burgundy. Home & Kitchen. Kitchen & Dining. GreenLife.
  - R: Bulova Frank Lloyd Wright Luxfer Prism Wall Clock, 14”, Bronze. Home & Kitchen. Home Dcor. Bulova.
  - S: Ortopad Girls Eye Patching Reward Posters: 1 Princess Poster, 1 Butterfly Poster. Home & Kitchen. Wall Art. Ortopad.
  - T: HIS H775F1GD Radeon HD 7750 1GB (128bit) GDDR5 Displayport HDMI DVI (HDCP) PCI Express X16 3.0 Graphics Cards. Electronics. Computers & Accessories. HIS.
- Output:
  - Ground truth: O
  - eCeLLM: O
  - GPT-4 Turbo: A

J.6. Multiclass Product Classification (MPC)

- Instruction:
  - Determine the relevance between the query and the product title provided, and select your response from one of the available options.
- Input:
  - Query: dj lights in the car
  - Product title: Sanhezhong Interior Car Lights, LED Car Strip Lights with Waterproof Design, 48 LED Remote Control Car Light Kit, Music Sync Under Dash Car Lighting with Car Charger, DC 12V
- Options:

  - A: The product is relevant to the query, and satisfies all the query specifications.
  - B: The product is somewhat relevant. It fails to fulfill some aspects of the query but the product can be used as a functional substitute.
  - C: The product does not fulfill the query, but could be used in combination with a product exactly matching the query.
  - D: The product is irrelevant to the query.
- Output:
  - Ground truth: A
  - eCeLLM: A
  - GPT-4 Turbo: B

J.7. Product Substitute Identification (PSI)

- Instruction:
  - Assess the relevance of a product to a given query by determining if it can function as a substitute, despite not fully meeting certain aspects of the query. Provide a binary output of yes or no based on this evaluation.
- Input:
  - Query: iphone 7 plus case otterbox.
  - Product: OtterBox SYMMETRY CLEAR SERIES Case for iPhone 8 Plus & iPhone 7 Plus (ONLY) - Retail Packaging - EASY BREEZY (CLEAR/EASY BREEZY).
- Output:
  - Ground truth: No
  - eCeLLM: No
  - GPT-4 Turbo: Yes

J.8. Query Product Ranking (QPR)

- Instruction:
  - Rank the products A, B, C, ... based on their relevance to the provided query, and produce a ranked list with the most relevant product positioned at the top of the list.
- Input:
  - Query: everyone loves raymond dvd complete series
  - Product A: Everybody Loves Raymond: The Complete Series - Seasons 1-9
  - Product B: Perry Mason: The Ninth and Final Season, Vol. 2
- Output:
  - Ground truth: A, B
  - eCeLLM: A, B
  - GPT-4 Turbo: A

J.9. Answerability Prediction (AP)

- Instruction:
  - Given a question and its related document, determine if the question is answerable by analyzing the information in the document. Output yes if the document addresses the question, or no otherwise.
- Input:

  - Question: if i order 10 of these, are the keys the same for each lock? In other words, can I use any of the keys to open any of the locks? Thanks!
  - Documents:
    * Document 1: I purchased two of these locks and within two uses with each one I had trouble opening the lock. The third time and I cannot get either opened at all. They are stuck on the tree. I'll order a different brand and then go cut them off. They keys each turn but they won't come off. Wish I had read the reviews.
    * Document 2: Locks are great my only complaint is I ordered two and I wanted them keyed the same, there was no where in placing the order to specify this and they sent two different keys. Otherwise all is good.
    * Document 3: I like the Python. It's easy to use & I feel it is well made. The problem is I use multiple game cameras & you can not buy the locks keyed alike. My local locksmith who is a Master Lock dealer said can not order them alike. I called Master Locks corporate office-and was told the same. If you need one they're great. I'm forced to look elsewhere.
    * Document 4: ...

- Output:
  - Ground truth: Yes
  - eCeLLM: Yes
  - GPT-4 Turbo: No

J.10. Answer Generation (AG)

- Instruction:
  - Answer the given question by extracting information from the supporting document.

- Input:
  - Question: It says it can support 2 DIMMs and 8GB of RAM. Does this mean I can use a single 8gb stick, or do I have to use 2x4?
  - Documents:
    * Document 1: Not going to write a full review here, it's a cheap motherboard, it's $50, expect to get what you pay for. I bought this from another retailer, but wanted everyone to know that this board does in-fact support 16GB of RAM. I am using this board in a home lab VMware ESXi host, and I needed 16GB at least. Here is what you need to do: (1) Update bios to at least 1303 (here is linky) [...] (2) Use RAM on supported RAM list (here is linky)[...] I am using 16GB of Corsair VengeanceCorsair Vengeance 16GB (2x8GB) DDR3 1600 MHz (PC3 12800) Desktop Memory (CMZ16GX3M2A1600C10)and it works like a champ.
    * Document 2: I bought this mobo to replace my ASUS MVA2 HDMI series when upgrading my system. As said in other reviews it can definitely use a 125 watt CPU as I checked the manual before I started my upgrade. Originally this was designed for Phenom 2 X6 cpus which run over 95 watts typically on load.On my rig I'm using an AMD FX8350, 8gb of Kingston DDR31600 ram (2 sticks of 4gb in each slot dual channel), ASUS AMD Radeon 7750 1gb, Kingston 120gb ssd, 1tb Western Digital hdd, Windows XP 32bit and 7 64 bit, in an Antec desktop case with an upgraded Antec 660 watt power supply.
    * Document 3: This is the PLUS version and will take a 125W processor (I have a AMD 965 quad core Deneb). It will take 8GB RAM in each of the two slots. It's a good board with all the Asus extras on it. Some considerations - the legacy connectors on the back might be useful for you or not. Most people would prefer extra USB's (or a couple of fast USB). The front panel connector is on the side next to the USB connectors so if you have a short panel connector cord(s) that *might* be an issue for you.
    * Document 4: Always check the manufacturer's website for reliable specs. I built a computer using this board and it works perfectly.Reviewer who called this mobo "junk" and gave it a "1" was attempting to use a CPU this board isn't rated for- board rated for up to 95W cpu. No mobo will run with wrong RAM, CPU, inadequate power supply, etc. Double check all components, OS, peripherals, BIOS version and settings, etc. for compatibility.Make sure you know what you're doing before attempting to build your own computer or you can end up wasting a lot of money.

    - Document 5: This board is an excellent choice if you are working on a budget and are making a relatively low end system. I used it along with an FX-6100 processor and have had no problems.I would recommend this to anyone building their first system and/or hoping to keep it a cheap build.I would not recommend this as an upgrade but rather a starting point.pros:cheaphigh connectabilitysmallgood SATA port placement.excellent support and manualsimplecons:no USB 3.0 or SATA IIIno IDEonly 4 USB ports on boardlimited ram supportonly PCIe 2.0 x16 @ 8xonly 2 fan connectors (one for cpu and one for case)
    - Document 6: ...

- Output:
  - Ground truth: You can use 8Gb modules, there's only 2 DIMMS. I'm using Corsair DDR 3 16GB kit of 2 8Gb1600Mhz
  - eCeLLM: You can use 2 sticks of 4GB or 1 stick of 8GB.
  - GPT-4 Turbo: You can use a single 8GB stick.