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

THE OHIO STATE UNIVERSITY
COLLEGE OF ENGINEERING

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

*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: ning.104@osu.edu

Motivation

E-commerce has been an integral part of daily life and drawn considerable attention from researchers. However, the conventional e-commerce models generally suffer from:

- Limited success in generalist e-commerce modeling;
- Unsatisfactory performance on new users and new products.

Contribution

To bridge the gap and develop e-commerce foundation models with real-world utilities for a large variety of e-commerce applications, we

- construct an open-sourced, large-scale, and high-quality benchmark instruction dataset ECInstruct for e-commerce realm;
- develop a series of e-commerce LLMs, denoted as eCeLLM, which substantially outperform baselines and exhibit excellent generalizability to out-of-domain settings.

ECInstruct Dataset

ECInstruct features 3 key design principles: broad coverage, realistic tasks, and high quality. ECInstruct dataset with diverse instructions covers with 116,528 samples from 10 real and widely performed e-commerce tasks. The 10 widely-performed tasks are

- attribute value extraction (AVE)
- product matching (PM)
- product relation prediction (PRP)
- sentiment analysis (SA)
- sequential recommendation (SR)
- multi-class product classification (MPC)
- product substitute identification (PSI)
- query-product ranking (QPR)
- answerability prediction (AP)
- answer generation (AG)

Quality Control:

To ensure the accuracy and high quality of ECInstruct dataset, we

- remove overlapping data between training and test sets to avoid data leakage;
- retain only data in English to ensure the unity of languages in texts;
- eliminate non-English notations such as HTML tags and Unicode;
- only select products with detailed information to allow sufficient product knowledge that LLMs can learn from;
- keep texts within a reasonable length following the convention in the literature;
- manually inspect all processed data.

We also conduct task-specific quality control on individual tasks.

Overview

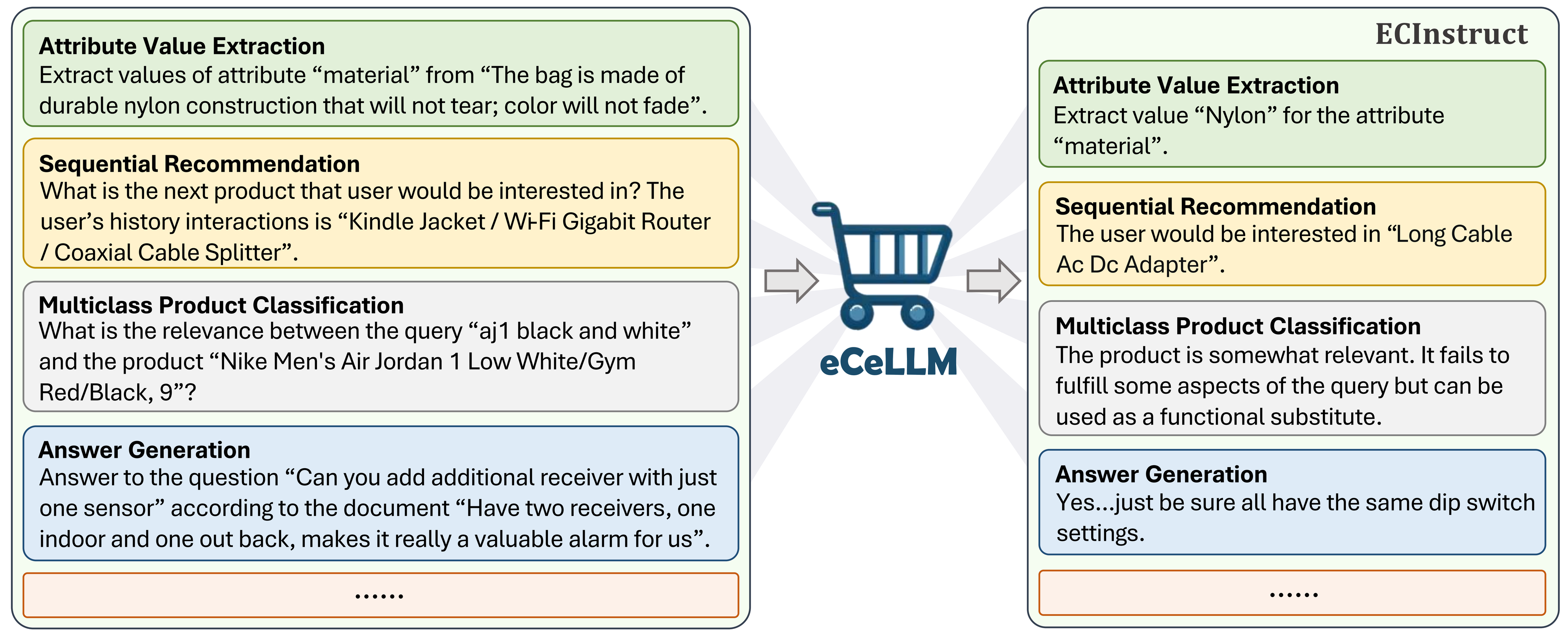

Figure 1. Overall scheme of eCeLLM instruction-tuned with ECInstruct

Comparison:

Table 1. Comparison among E-commerce LLMs

Comparison	LLaMA-E [2]	EcomGPT [3]	eCeLLM [1]
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.

Experiment Results

- 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 (Table 2).

Table 2. 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

- eCeLLM models show outstanding generalizability to OOD products and surpass the best baselines with a remarkable average improvement of 9.3% in OOD evaluation (Table 3).

Table 3. Overall Performance in OOD Evaluation

Model	AVE	PRP	SA	SR	AP	AG
	F1*	Macro F1	Macro 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

- By training over diverse instructions, eCeLLM is equipped with strong generalizability to unseen instructions (Table 4).

Table 4. 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

- Trained on all the tasks in ECInstruct together, eCeLLM exhibits similar or better performance than models trained on each individual task (Table 5).

Table 5. 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

Reference

[1] Peng, B., Ling, X., Chen, Z., Sun, H., and Ning, X. eCeLLM: Generalizing Large Language Models for E-commerce from Large-scale, High-quality Instruction Data. In Forty-first International Conference on Machine Learning.

[2] 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.

[3] 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.

https://ninglab.github.io/eCeLLM/