Shopping MMLU: A Massive Multi-Task Online Shopping Benchmark for Large Language Models

Yilun Jin, Zheng Li, Chenwei Zhang, Tianyu Cao, Yifan Gao, Pratik Jayarao, Mao Li, Xin Liu, Ritesh Sarkhel, Xianfeng Tang, Haodong Wang, Zhengyang Wang, Wenju Xu, Jingfeng Yang, Qingyu Yin, Xian Li, Priyanka Nigam, Yi Xu, Kai Chen, Qiang Yang, Meng Jiang, Bing Yin

Correspondence: yilun.jin@connect.ust.hk, amzzhe@amazon.com

amazon

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

SIGILLUM UNIVERSITATIS DOMINAE NOSTRAE A LACU
VITA DULCEDO SPES

Shopping MMLU Highlights

- Massive multi-task online shopping benchmark for Large Language Models (LLMs).
  - 57 tasks, 4 major skills, ~20k questions.
  - Mostly real-world data from Amazon.
- Extensive experiments uncover insights on domain-specific LLMs.
  - General abilities, fine-tuning, in-context learning, etc.
- With Shopping MMLU, we host KDD Cup 2024 with over 500 participating teams.

Paper

Dataset

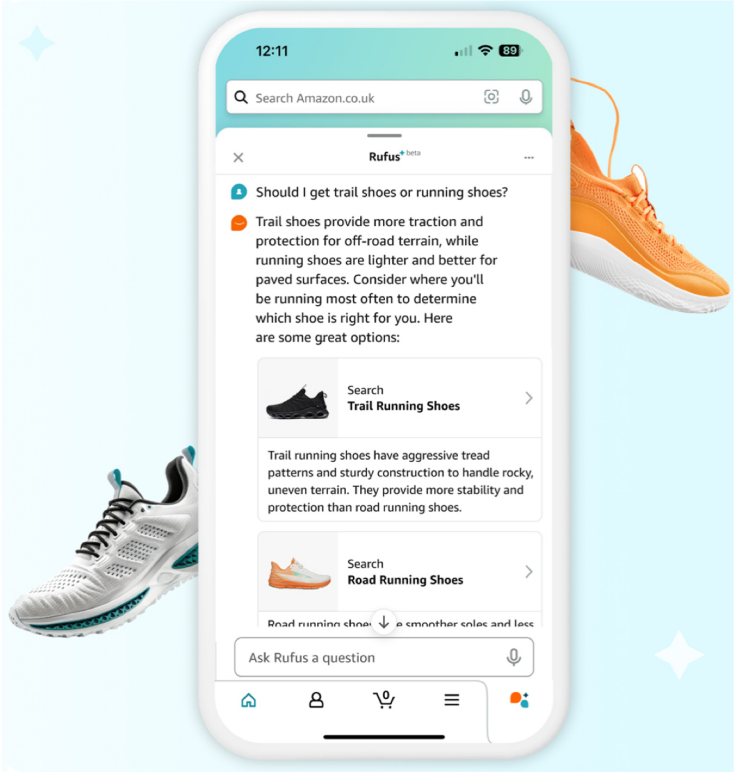
KDD Cup 2024 Challenge

Workshop And Winning Solutions

LLMs for Online Shopping

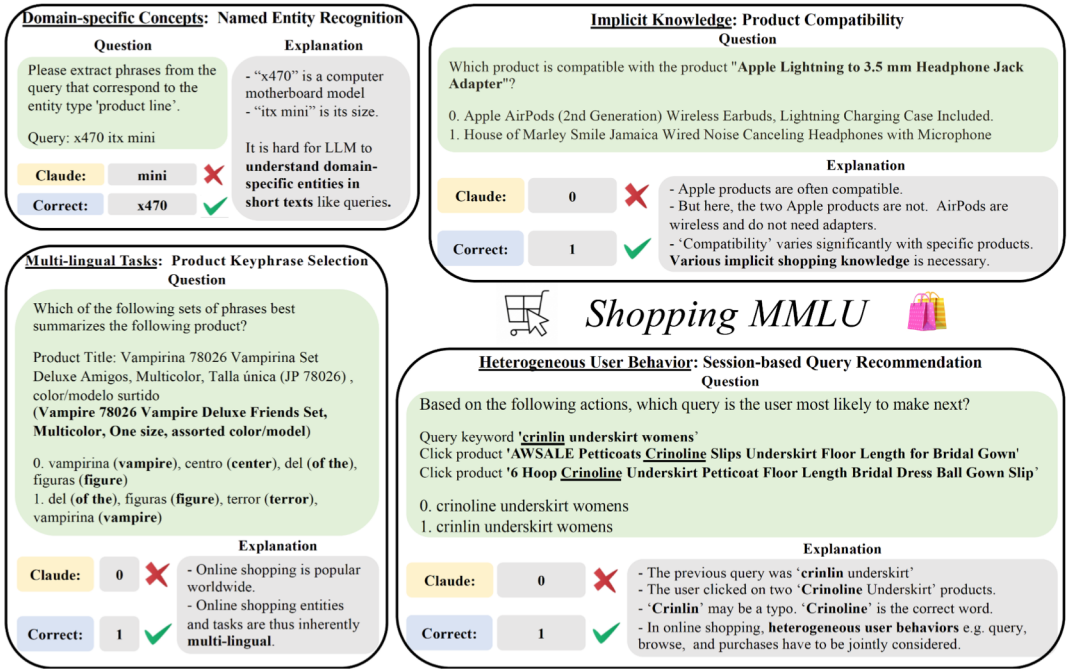
Advantages:

- **Multi-task:** One model for multiple shopping tasks. Less task-specific engineering.
- **Few-shot Adaptation:** Fast adaptation to new users, new products, etc.
- **Conversational Shopping:** Interactive, real-time feedback to customer questions.
    - Amazon Rufus Shopping Assistant



https://www.aboutamazon.com/news/retail/how-to-use-amazon-rufus

Unique Properties of Online Shopping



Shopping MMLU Organization

4 main shopping skills, 57 tasks:

- Shopping concept understanding
- Shopping knowledge reasoning
- User behavior alignment
- Multi-lingual abilities

More comprehensive skill and task coverage than existing datasets!

Figure 2: A brief taxonomy of Shopping MMLU including all skills and sub-skills.

Dataset	Unified Text-Gen Formulation	# Tasks	Concept Understanding	Knowledge Reasoning	User Behavior	# Languages
MAVE [48]	No	1	Partially	No	No	1
Amazon-M2 [16]	No	3	No	No	Partially	6
Amazon ESCI [33]	No	3	No	No	Partially	3
EComInstruct-Test (EcomGPT) [20]	Yes	12	Yes	No	No	2
ECInstruct (eCeLLM) [30]	Yes	10	Partially	No	Yes	1
Shopping MMLU	**Yes**	**57**	**Yes**	**Yes**	**Yes**	**6**

Experiments on Shopping MMLU

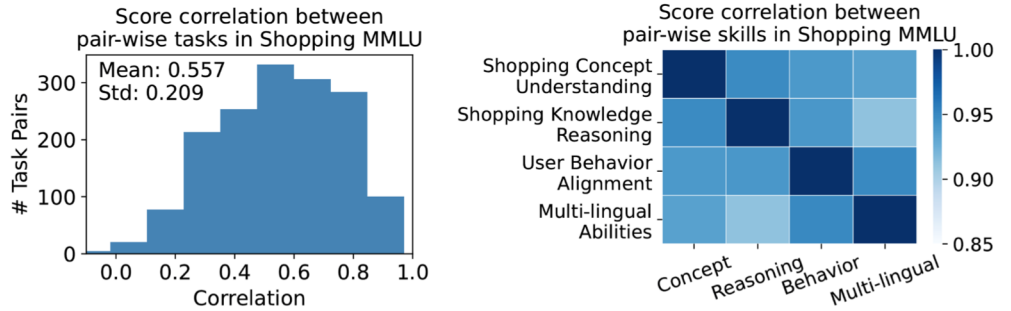
- **Observation 1**: Claude-3 Sonnet performs the best overall.
- **Observation 2:** Strong open-source LLMs catch up with proprietary ones.
- **Observation 3:** Domain-specific LLMs (eCeLLMs) are not always the strongest.

Model Type	# Params.	Model	Shopping Concept Understanding	Shopping Knowledge Reasoning	User Behavior Alignment	Multi-lingual Abilities
Proprietary	N/A	Claude-3 Sonnet	**80.75**	**71.63**	**70.17**	**67.76**
		Claude-2	75.46	65.50	63.53	65.24
		ChatGPT	75.63	64.97	59.79	60.81
Open-Source	70B	LLaMA3-70B-Instruct	**75.24**	**69.29**	**67.67**	62.00
		QWen1.5-72B	71.67	68.92	64.12	**64.84**
		LLaMA3-70B	69.59	63.56	55.77	58.95
		LLaMA2-70B-chat	61.84	40.73	44.20	47.04
		LLaMA2-70B	61.05	55.87	43.24	47.85
		Mixtral-8x7b	59.43	54.32	55.31	44.69
	14B	QWen1.5-14B	**67.22**	**60.92**	**54.92**	55.21
		eCeLLM-L	61.54	54.84	54.55	**59.64**
		Vicuna-13B	59.64	52.63	49.81	49.64
		LLaMA2-13B-chat	51.79	45.01	39.95	42.99
		LLaMA2-13B	45.86	39.47	39.43	44.23
	7B	LLaMA3-8B-Instruct	**65.26**	**56.84**	**54.88**	55.37
		LLaMA3-8B	58.02	49.74	44.16	51.03
		QWen1.5-7B	58.89	52.34	49.81	50.14
		eCeLLM-M	63.29	48.94	53.78	**56.08**
		Zephyr	61.65	52.57	44.73	45.35
		Mistral-7B-instruct	62.03	46.36	42.21	43.32
		Mistral-7B	55.82	46.69	46.27	41.47
		Vicuna-7B	53.46	45.06	41.11	43.82
		LLaMA2-7B-chat	51.67	43.48	41.42	40.43
		LLaMA2-7B	38.22	32.81	32.56	27.71
	<5B	QWen1.5-4B	**57.21**	**52.56**	**42.74**	**49.78**
		Phi-2	49.34	42.83	36.38	32.91
		eCeLLM-S	49.40	39.06	36.33	32.79

Experiments on Shopping MMLU

Tasks in online shopping **share common knowledge** and can be **jointly improved**!

- Skills and tasks in Shopping MMLU bear strongly positive correlations with each other.



(a) Distribution of Task-wise Correlations

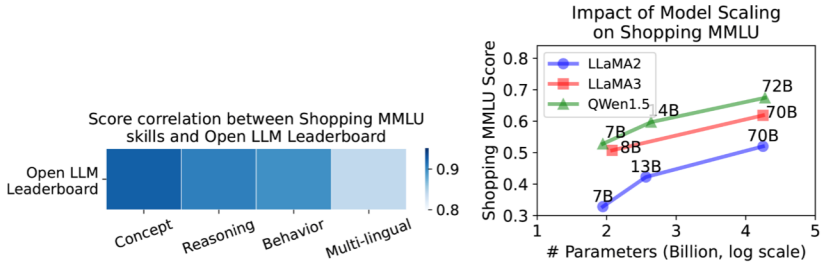
(b) Skill-wise correlations.

Figure 3: Task and skill-wise score correlations of Shopping MMLU.

Experiments on Shopping MMLU

General abilities transfer well to the specific domain of online shopping.

- Shopping MMLU scores highly correlate with Open LLM Leaderboard.
- Model scaling generally improves Shopping MMLU scores.



(a) Score correlations between Shopping MMLU skills and Open LLM Leaderboard [4]

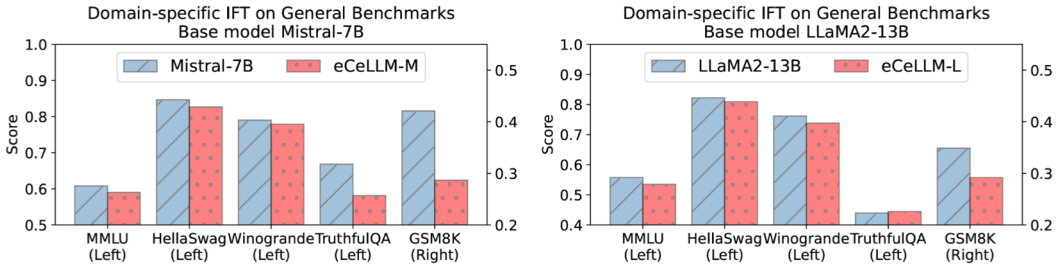
(b) Effects of model scaling on Shopping MMLU

Figure 4: General knowledge transfers well to online shopping.

Experiments on Shopping MMLU

Domain-specific fine-tuning may compromise general abilities.

- eCeLLMs perform generally worse than their base models on general LLM benchmarks.
- **Strong general abilities is the foundation of domain specialization!**



(a) Mistral-7B VS eCeLLM-M (b) LLaMA2-13B VS eCeLLM-L

Figure 7: Scores of eCeLLM and their base models on general LLM benchmarks.

Recap: Shopping MMLU Highlights

- Massive multi-task online shopping benchmark for Large Language Models (LLMs).
  - 57 tasks, 4 major skills, ~20k questions.
  - Mostly real-world data from Amazon.
- Extensive experiments uncover insights on domain-specific LLMs.
  - General abilities, fine-tuning, in-context learning, etc.
- With Shopping MMLU, we host KDD Cup 2024 with over 500 participating teams.

Paper

Dataset

KDD Cup 2024 Challenge

Workshop And Winning Solutions

Thanks!

Correspondence: yilun.jin@connect.ust.hk, amzzhe@amazon.com
Paper: https://arxiv.org/abs/2410.20745
Dataset: https://github.com/KL4805/ShoppingMMLU