More diverse more adaptive: Comprehensive Multi-task Learning for Improved LLM Domain Adaptation in E-commerce

Piao Tong*
piaot@std.uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, China

Pei Tang*
richpp523@gmail.com
Xiaohongshu Inc.
Shanghai, China

Zhipeng Zhang*
zhangzhipeng.work@bytedance.com
ByteDance Inc.
Shanghai, China

Jiaqi Li
slrrr8848@gmail.com
Sichuan University
Chengdu, China

Qiao Liu
qliu@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, China

Zufeng Wu✉
wuzufeng@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, China

ABSTRACT

In recent years, Large Language Models (LLMs) have been widely applied across various domains due to their powerful domain adaptation capabilities. Previous studies have suggested that diverse, multi-modal data can enhance LLMs' domain adaptation performance. However, this hypothesis remains insufficiently validated in the e-commerce sector. To address this gap, we propose a comprehensive e-commerce multi-task framework and design empirical experiments to examine the impact of diverse data and tasks on LLMs from two perspectives: "capability comprehensiveness" and "task comprehensiveness." Specifically, we observe significant improvements in LLM performance by progressively introducing tasks related to new major capability areas and by continuously adding subtasks within different major capability domains. Furthermore, we observe that increasing model capacity amplifies the benefits of diversity, suggesting a synergistic relationship between model capacity and data diversity. Finally, we validate the best-performing model from our empirical experiments in the KDD Cup 2024, achieving a rank 5 in Task 1. This outcome demonstrates the significance of our research for advancing LLMs in the e-commerce domain.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

session-based recommendation, e-commerce, language models

ACM Reference Format:
Zhipeng Zhang, Piao Tong, Qiao Liu, Yingwei Ma, Xujiang Liu, Xu Luo. 2024. More diverse more adaptive: Comprehensive Multi-task Learning for Improved LLM Domain Adaptation in E-commerce. In *KDD Cup 2024 Workshop: A Multi-task Online Shopping Challenge for Large Language Models, August 25, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Recently, Large Language Models (LLMs) have been deployed across various domains[2, 6], leveraging their robust domain adaptation capabilities and the flexibility of generative models, which are not constrained by the task-specific fine-tuning requirements of discriminative models. This adaptability allows LLMs to easily transfer to different fields and achieve impressive performance.

The e-commerce sector is rich in textual data, including product titles, descriptions, reviews, and Q&A content[4, 5]. Researchers have begun to utilize these data sources as inputs for LLMs in e-commerce applications, yielding promising results. These studies encompass a wide range of tasks, from classic NLP extraction and classification to product sequence recommendation and query ranking. However, most research focuses on Prompt Engineering or fine-tuning for specific tasks. This trend raises an intriguing and pressing research question: How do large models perform in multi-task settings, and what is the optimal approach to constructing a multi-task framework for the e-commerce domain?

Early methodologies, such as P5[1], were proposed to introduce a wider range of tasks to explore model performance across various recommendation system-related scenarios. However, their multi-task design remained confined to the realm of recommendations. Subsequently, researchers attempted to collate open-source data from diverse e-commerce domains[3], manually categorizing task types to construct an ontological structure. This led to the delineation of categories such as "Product Understanding," "User Understanding," "Query Product Matching," and "Product QA." Nevertheless, the framework's integrated data lacked comprehensiveness, primarily due to its insufficiently high-level and generalizable task classification (e.g., inability to categorize reasoning tasks and multilingual tasks). The ShopBench framework, introduced in KDD CUP 2024, addressed these limitations. It constructed a comprehensive multi-task learning framework for large language models based on the sequence of human cognitive patterns, progressing from "common sense cognition to logical reasoning, followed by fine-tuning on downstream tasks." However, this framework also

* The first four authors contributed equally to this work.
✉ Corresponding Author.

KDDCup '24, August 25, 2024, Barcelona, Spain

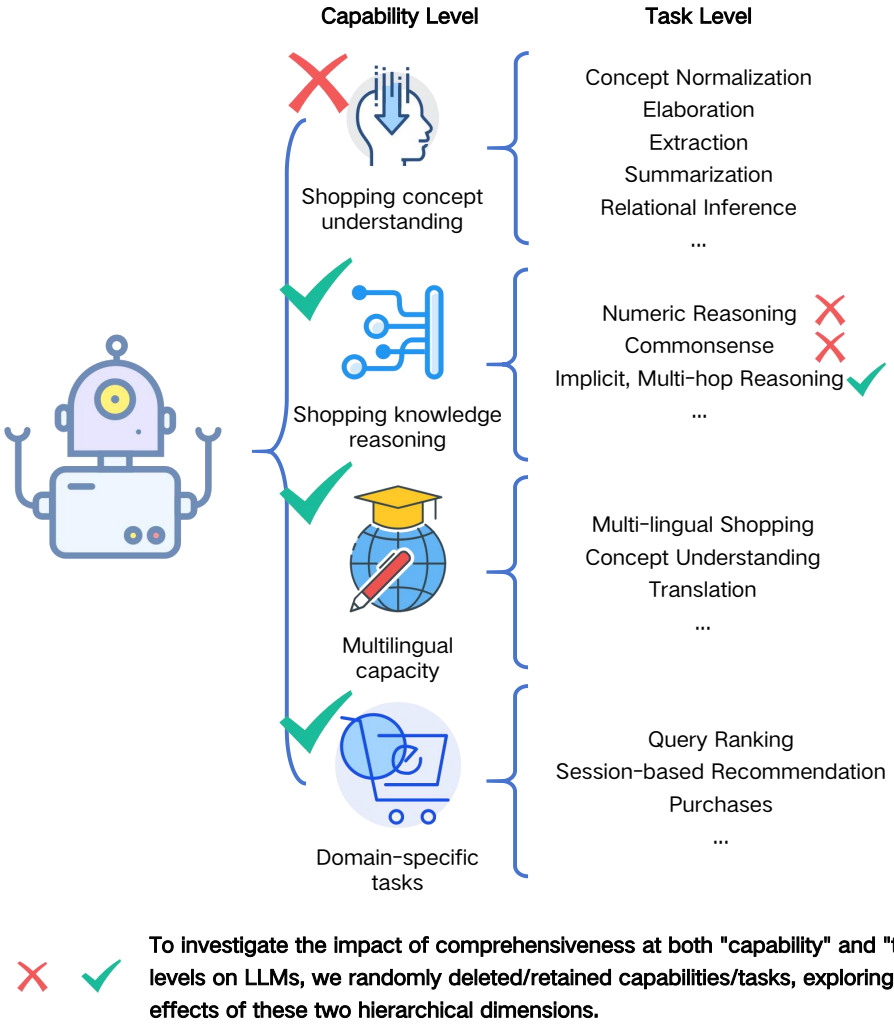

Figure 1: Our Experimental Framework: Ablation Studies at Capability and Task Levels to Evaluate the Impact of Data Diversity on Large Model Performance.

gave rise to new research questions, particularly regarding how the capabilities of large language models evolve with respect to both the comprehensiveness of the framework and the diversity of tasks.

In this paper, we address the capabilities integrated into large language models within the ShopBench framework by designing a diverse array of tasks that align with e-commerce business logic. These tasks are highly heterogeneous and numerous, designed to test various aspects of model performance. We employ multiple sensible approaches to acquire task-specific data, including prompt engineering and open-source data collection. Notably, we conducted extensive empirical experiments based on these data, aiming to explore the impact of diverse data and tasks on LLMs from two perspectives: "capability comprehensiveness" and "task comprehensiveness." Specifically, we investigated these perspectives by ablating major capability domains of the large language model and controlling the number of subtasks corresponding to each major capability domain. This research provides valuable insights into comprehensively constructing e-commerce multi-task learning frameworks for large language models and expanding the capability boundaries of e-commerce LLMs. Additionally, we posit that as data diversity increases, the fine-tuning capacity of large language models must be concurrently enhanced to accommodate more diverse features. We substantiate this hypothesis by manipulating capacity parameters.

2 METHODOLOGY

2.1 Experimental Framework

We first designed corresponding subtasks for the four key capabilities of large language models mentioned in ShopBench. We then conducted ablation experiments at both "Capability Level" and "Task Level" to evaluate the impact of "Capability comprehensiveness" and "Task diversity" on LLM domain adaptation performance. When designing tasks, two key points were considered:

- Task heterogeneity: Even subtasks within a single Capability should be as diverse as possible.

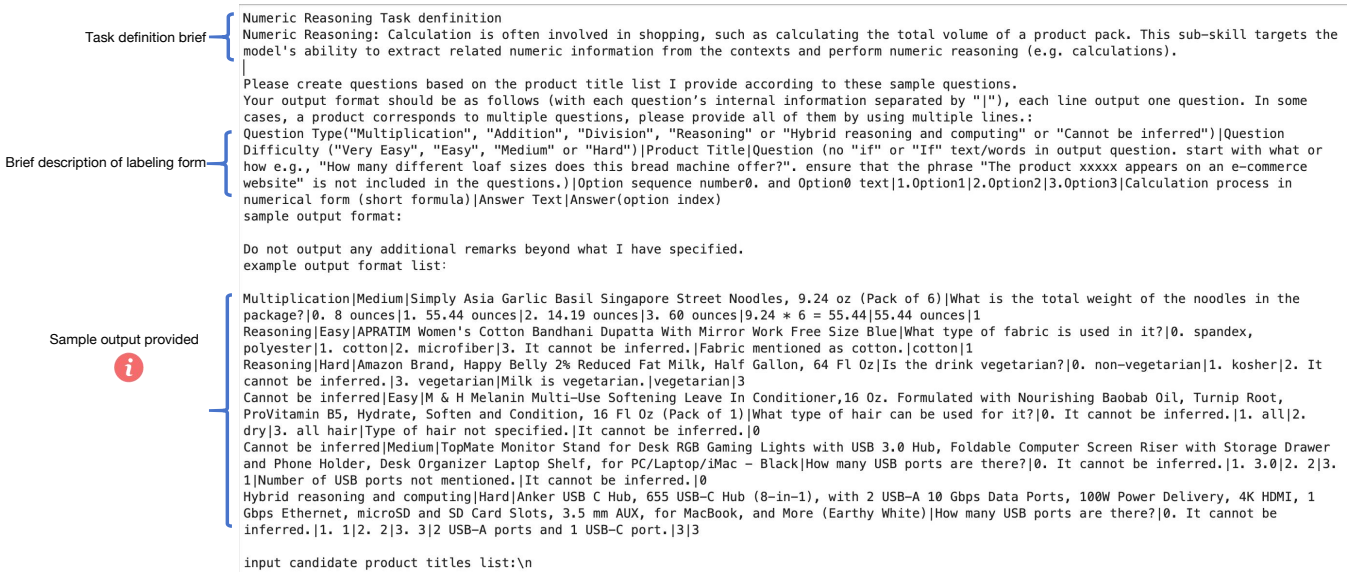

Figure 2: Exemplar Instructions for GPT-4 in Generating Training Data (Reasoning Tasks).

- Alignment with e-commerce business logic: Tasks unrelated to the e-commerce domain may potentially hinder LLM performance in e-commerce applications.

2.2 Data Acquisition for Training

For each task we designed, corresponding training data must be acquired. We obtain data through two approaches: (1) Collecting open-source data, and (2) Generating training data using high-performance LLMs like GPT-4.

Important tip: When aiming to make LLM-generated annotations possess certain characteristics, relying solely on prompt engineering may be insufficient. For instance, to include challenging samples in the output, an incorrect and correct approach are as follows:

- Incorrect: Including "Please ensure your annotations contain some difficult questions to thoroughly train my model" in the instruction.
- Correct: Explicitly incorporating desired output sample characteristics using delimiters in the "Sample output" provided to the LLM. As shown in Figure 3, we include "|Question Difficulty ("Very Easy", "Easy", "Medium" or "Hard")|" in the "Sample output" to prompt the model to consider question difficulty when generating the dataset.

2.3 LoRA Rank Control During Fine-tuning

A logical approach suggests that as data diversity increases, the parameter volume during model fine-tuning should correspondingly increase to accommodate the data's diversity. Currently, LoRA is the predominant fine-tuning method, utilizing two low-rank matrices to generate matrices of the same size as LLM components, which are then added to the original parameter matrices. By controlling the rank of these low-rank matrices, we can regulate the number of fine-tuning parameters. While previous studies have examined the relationship between parameter quantity and LLM performance, few have investigated how rank changes affect LLM performance under varying data diversity conditions. Therefore, we attempt to control both diversity and rank to explore the relationship between fine-tuning parameter capacity and data diversity.

3 EXPERIMENTS

3.1 Overall Performance

Our approach achieved considerably performance gain over the baseline solution, and ranked top-5 in task-1. The main results are shown in Table 1.

Dataset	Metric	Ranking
track1	Score=0.803	**5th**
track4	Score=0.707	7th
track5	Score=0.745	6th

Table 1: Performance of our approach.

3.2 Ablation Studies

We conduct ablation studies on the training data from two perspectives: "capability comprehensiveness" and "task comprehensiveness," while simultaneously controlling the rank of LoRA to investigate the synergistic effects of model parameter capacity and data diversity on LLM performance. Our experiments yield two key findings:

- When developing LLMs, incorporating a more comprehensive and diverse set of capabilities through varied data leads to qualitative improvements in model performance. Conversely, blindly increasing data volume with homogeneous content can actually degrade model effectiveness.
- As data diversity increases, it is crucial to concurrently expand the parameter capacity during LLM fine-tuning, enabling the model to effectively accommodate and leverage the enhanced data diversity.

Variants	Track1 Score
Only generating and training Task3 data	0.692
Preliminary addition of Track1 data	0.745
Reducing data from 30,000 to 2,000 samples	0.758
Adding tasks related to Track3 and Track2	0.774
Adjusting from 2,000 to 6,000 samples	0.784
Increasing LoRA rank from 16 to 32	**0.803**

Table 2: Ablation Study on Model Selection, Training Data Choice, and LoRA Rank.

4 CONCLUSION AND FUTURE WORK

This study, through empirically designed experiments, underscores the critical importance of comprehensive and diverse data in developing effective LLMs. Furthermore, the synergistic relationship between data diversity and LoRA rank variations in enhancing LLM performance demonstrates that increased parameter capacity is necessary to accommodate diverse information within the data. In future work, we plan to explore additional methods for expanding model capacity, such as MMoE and multi-LoRA, and further investigate their synergistic effects with data diversity in advancing LLMs within the e-commerce domain

REFERENCES

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

[2] Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian McAuley. 2023. Text is all you need: Learning language representations for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1258–1267.

[3] Bo Peng, Xinyi Ling, Ziru Chen, Huan Sun, and Xia Ning. 2024. eCeLLM: Generalizing Large Language Models for E-commerce from Large-scale, High-quality Instruction Data. *arXiv preprint arXiv:2402.08831* (2024).

[4] Piao Tong, Zhipeng Zhang, Qiao Liu, Xujiang Liu, Xu Luo, and Huhao Ran. 2024. Interpretable prediction model for decoupling hot rough rolling camber-process parameters. *Expert Systems with Applications* (2024), 124872.

[5] Piao Tong, Zhipeng Zhang, Qiao Liu, Yuke Wang, and Rui Wang. 2024. CARE: Context-aware attention interest redistribution for session-based recommendation. *Expert Systems with Applications* (2024), 124714.

[6] Zhipeng Zhang, Piao Tong, Yingwei Ma, Qiao Liu, Xujiang Liu, and Xu Luo. 2023. Language-Enhanced Session-Based Recommendation with Decoupled Contrastive Learning. *arXiv preprint arXiv:2307.10650* (2023).