LocalAdapt: Continuously Pre-training LLaMA for a Customized Local Life Model

Anonymous ACL submission

Abstract

Large language models (LLMs) have gained considerable attention due to their remarkable generalization capabilities, as exemplified by ChatGPT and GPT-4(OpenAI, 2023). However, these models exhibit limitations in specific domains, such as local life service scenarios, stemming from insufficient relevant knowledge and considerable disparities between local life industry data and general data. To address this issue, we first introduce a 170GB domain-specific corpus, LocalEvolve, for unsupervised continued pretraining. Second, we employ a low-rank adaptation approach to train a customized LLM, LocalAdapt, for local life service scenarios. Notably, we design a Multi-Task mapping system that transforms structured industry data into various Fundamental Reasoning Units (FRUs). Our LocalAdapt model demonstrates superior performance across different local life tasks compared to baseline models. Extensive empirical analysis further confirm the effectiveness of FRUs.

1 Introduction

Large Language Models (LLMs) signify a milestone in the field of natural language processing (NLP)(Shanahan, 2023)(Wei et al., 2023). Characterized by their billion-scale parameters and extensive pretraining on massive text data, LLMs exhibiting remarkable capabilities(Wei et al., 2022b), including in-context learning, instruction following, step-by-step reasoning and so on. Such abilities enable them to address complex challenges once considered insurmountable for NLP systems, catalyzing transformative changes across numerous industries(Rae et al., 2022).

Although large models exhibit remarkable generalization capabilities, their lack of specialization for the local life industry may lead to sub-optimal performance in related tasks(Xiong et al., 2023). This is due to the unique structure and word distribution of local life data compared to general data(Li et al., 2023). Local life data is derived from lifestyle services and live streaming platforms, predominantly consisting of structured attribute-value pairs and relational tuples, which challenges general LLMs' comprehension. Additionally, local life data contains prevalent industry-specific concepts with diverse representations across merchants, further complicating the task for general models. Therefore, there is a crucial need for a large model tailored to the local life industry, encompassing extensive industry knowledge and a thorough understanding of local life data.

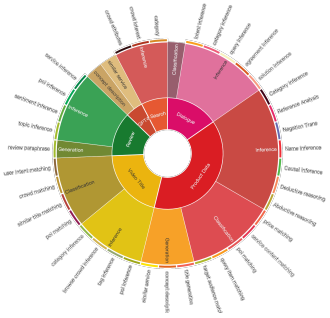

Figure 1: The complete schema of the FRUs.

In the LLM era, various efforts aim to improve model generalization in specific industries. One retrieval-based approach calls query-related instances for in-context learning(Asai et al., 2023)(Qian et al., 2023). However, this method fails to fundamentally address the lack of relevant industry knowledge in LLMs(Li et al., 2022), and accurately recalling related instances remains challenging(Cui et al., 2023a). Another approach attempt involves constructing multi-task instruction sets and employing instruct learning(Ouyang et al., 2022). Although effective, this method demands high complexity, diversity, and scale of instructions, thus increasing the burden of data annotation and collection for human(Sanh et al., 2022)(Wei et al., 2022a).

To imbue models with domain-specific knowl-

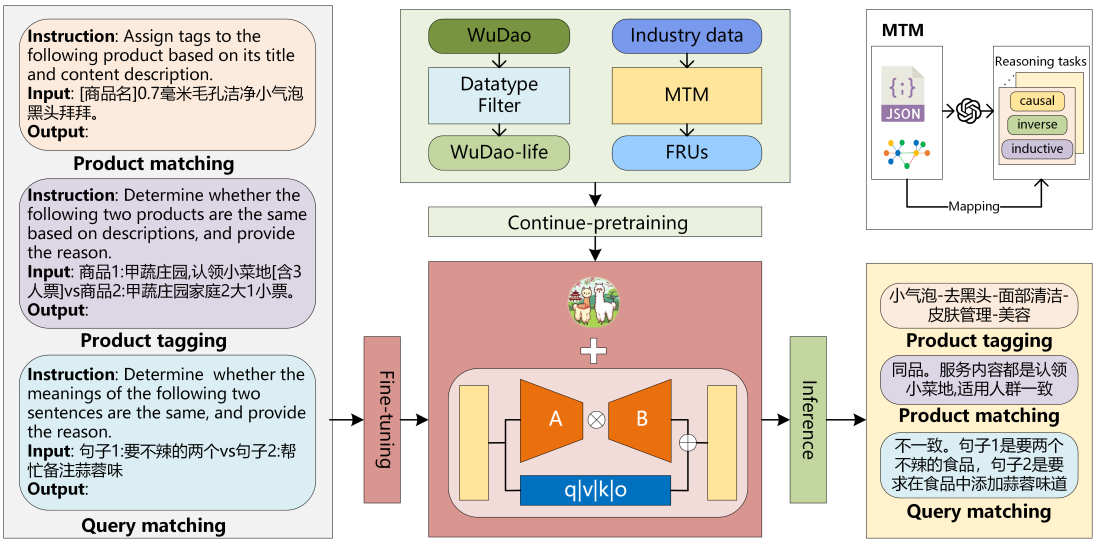

Figure 2: Overiew of LocalAdapt pipeline.

edge, we create LocalEvolve, a 170GB local life-related dataset comprising external and industry knowledge. External knowledge primarily comes from the WuDao corpora(Yuan et al., 2021), while industry knowledge mainly comes from platform data (e.g. product information, user reviews, video titles). Notably, we design a Multi-Task Mapping system(MTM) to assemble structured industry data into Fundamental Reasoning Units (FRUs), which can not only enhance the model's understanding of local life concepts, but also enable it to tackle complex reasoning tasks across various local life scenarios.

By continue-pretraining Chinese-LLaMA-13B on the LocalEvolve, we obtain a customized local life model, LocalAdapt. We adopt a parameter-efficient low-rank adaptation method(Hu et al., 2021), which not only improves training efficiency but also enhances the model's performance on downstream tasks, particularly in low-resource scenarios. LocalAdapt significantly outperforms the baseline in product matching, product tagging, and user query matching. Further ablation studies validate the effectiveness of incorporating FRUs, showing that multi-task reasoning learning during pretraining improves the model's reasoning abilities.

The main contributions of this paper are summarized below:(1)We use multi-task reasoning for organizing pretraining data, introducing a 170GB local life dataset, LocalEvolve. Our proposed MTM system effectively converts structured industry data into FRUs, significantly enhancing the diversity and inferential nature of pretraining corpora. (2)Using a low-rank adaptation approach on pretrained Chinese-LLaMA-13b, we create LocalAdapt, a customized local life model with exceptional performance across various local life tasks.(3)Extensive experiments demonstrate the effectiveness of LocalAdapt compared to other baseline models.

2 Approach

2.1 Overview of the LocalEvolve

In this section, we introduce LocalEvolve, a large-scale, high-quality dataset consisting of 170GB, primarily derived from two sources. Firstly, we collected life service-related data from public sources. These datasets are typically of high quality as they have been carefully curated by experts in the field.

Secondly, We collect data from life service sectors, which can be categorized into basic types: product information, user reviews, live video ASR, and video titles. For the 80% structured data, We design an MTM system, constructing a series of FRUs based on atomic reasoning tasks such as causal inference, primary and secondary reasoning, and inductive reasoning. A detailed description of the construction methods will be provided in the subsequent sections.

2.2 Acquiring External Data from WuDao Corpora

For public datasets, we primarily chose WuDao Corpora, a large-scale, high-quality dataset released by the Beijing Academy of Artificial Intelligence(BAAI). With 200GB of open-source data, it covers 50+ industries, including military and tech-

Models	Product Tagging			Product Matching			Query matching		
	P	R	F1	P	R	F1	P	R	F1
ChatGPT	76.84	67.16	71.67	87.03	47.11	61.02	79.02	42.11	54.88
Bert	89.01	77.69	81.62	75.62	69.41	72.38	43.10	26.86	39.77
Chinese-LLaMA-7b	96.85	74.61	80.70	73.92	79.06	76.41	75.74	73.54	74.63
Chinese-LLaMA-13b	83.22	86.01	83.69	75.17	83.02	78.90	76.01	84.04	79.82
Chinese-LLaMA-13b(LoRA)	96.36	80.91	85.83	73.89	86.42	79.66	76.73	87.22	81.64
LocalAdapt(ours)	92.36	**90.39**	**92.07**	77.80	**86.65**	**81.99**	**82.56**	**93.24**	**87.58**
w/o FRUs	89.01	77.69	81.62	74.98	86.18	80.19	76.74	87.07	81.58

Table 1: Precision (P), Recall (R), and F1 scores for ChatGPT, open-source general models, and LocalAdapt(ours) on product tagging, product matching, and query matching tasks.

nology, We primarily focus on lifestyle-related industries(entertainment, tourism, gaming, education and so on)to include in our LocalEvolve, amounting to 156GB in total.

2.3 Mapping Structured Data to FRUs

life service-related data primarily consists of four types of data: platform product data, user review data, live video ASR, and video titles, with structured data accounting for 80%. Due to the significant differences in form between structured data and common fluent text data, organizing pretraining corpora through simple linearization(Jiang et al., 2023) result in suboptimal performance for the model during the supervised fine-tuning phase. Specifically, the customized model experiences hard-to-suppress hallucinations during inference on downstream tasks(Manakul et al., 2023), generating numerous undesired product descriptions. This severely degrades the model's performance on downstream tasks and significantly increases the inference time because of the generation of excessive unrelated tokens.

To mitigate these hallucinations, we propose a Multi-Task Mapping (MTM) system that maps structured data to Foundation Reasoning Units (FRUs) to reduce data form differences and introduce more reasoning logic. Specifically, we randomly extract 100 data samples from each basic data type, and have GPT4 generate candidate reasoning tasks, detailed instruction design is presented in the appendixA.4. Then, human experts select atomic reasoning tasks from the candidates, and GPT4 designs mapping templates based on the structured data and atomic reasoning tasks. By fitting the structured data into these mapping templates, we obtain a series of natural language texts, namely FRUs, which encompass abundant domain knowledge and reasoning logic.

Tasks	**Size**	**In.len**	**Out.len**
product tag	2,000	75.12	8.67
product match	2,000	100.81	15.60
query match	1000	89.91	132.47

Table 2: Details of Data for SFT Tasks. "In.len" denotes the average length of inputs, "Out.len" denotes the average length of labels.

2.4 Training of LocalAdapt

We train LocalAdapt using Chinese-LLaMA-13b(Cui et al., 2023b), a customized model for Chinese, which supports efficient fine-tuning methods(Liu et al., 2022) like LoRA and acceleration techniques like DeepSpeed.

We adopt the LoRA approach to continue pretraining Chinese-LLaMA-13b on LocalEvolve and utilize DeepSpeed for acceleration. The continue-pretraining process was conducted using 24 A100 GPUs for a duration of 96 hours with a learning rate set to 2e-4 and weight decay set to 0.01. The batch size per device set to 4 and the gradient accumulation step set to 1. The low-rank adaption is applied to q,v,k,o and rank is set to 64 with alpha set to 32.

3 Experiments

3.1 Main Results

We selected three tasks, namely product tagging, product matching, and user query matching, to compare the performance between LocalAdapt and Chinese-LLaMA-13b. The data for the three tasks are derived from real-world business scenarios, with detailed descriptions provided in the appendixA.1. The dataset scale are detailed in Table 2. We allocate 20% of the dataset for validation and use precision, recall, and F1 scores as automatic metrics to measure the model's inference conclu-

Models	Read		Fact	
	flu.	gram.	corr.	compl.
ChatGPT	4.32	4.25	3.14	3.36
LLaMA-7b	4.09	3.60	3.57	3.22
LLaMA-13b	4.20	3.68	3.61	3.39
LLaMA-13b#	4.23	3.66	3.72	3.51
LocalAdapt	4.30	**4.26**	**3.98**	**3.82**
w/o FRUs	4.17	3.59	3.73	3.59

Table 3: Results of manual evaluation on product and query matching tasks, where 'LLaMA-13b#' refers to Chinese-LLaMA-13b trained using LoRA.

sions, and we assess the correctness and logicality of the model's inference process through manual evaluation. Tables 1 and 3 present the results of the automatic metric evaluation and manual evaluation, respectively. Table 1 shows that the F1 scores of LocalAdapt on all three tasks significantly outperform the baseline models, demonstrating that LocalAdapt has better reasoning performance in life service scenarios. Table 3 indicates that the readability and correctness of the LocalAdapt inference process surpass those of the baseline models, proving that LocalAdapt can generate more accurate and logical reasoning processes. The human evaluation guidelines and specific cases are presented in the appendixA.2 and A.5.

3.2 Ablation Experiments on FRUs

In order to validate the effectiveness of FRUs, we extract 15 million structured data and set up the following three experimental groups: Group 1: The structured data is simply concatenated without FRUs; Group 2: Halve atomic reasoning tasks and construct FRUs to form pre-training corpora; Group 3: All FRU reasoning tasks are retained. We continue pre-training the backone model under these three experimental settings, obtaining LocalAdapt–, LocalAdapt-, and LocalAdapt*. Subsequently, we fine-tune the three models on the product matching task. As shown in Table 4, LocalAdapt* achieves the highest F1 score, while the F1 metric of LocalAdapt- is higher than that of LocalAdapt–. This indicates that: (1) FRUs can enhance the model's reasoning capabilities; (2) A greater variety of atomic tasks leads to better inference performance in the model.

3.3 Effect of the LoRA Rank

We further investigate the impact of LoRA rank on the pre-training performance. Fig 3 presents the

Model	P	R	F1
LocalAdapt–	74.17	83.02	78.34
LocalAdapt-	76.67	84.12	80.23
LocalAdapt*	77.04	86.20	81.36

Table 4: Metrics for LocalAdapt–, LocalAdapt-, and LocalAdapt* after fine-tuning on the product matching task.

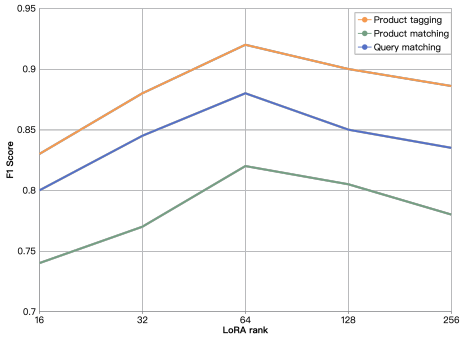

Figure 3: Metrics for LocalAdapt on product tagging, Product matching, and query matching tasks with different LoRA ranks in pre-training.

F1 score changes of LocalAdapt after fine-tuning on product tagging, product matching and query matching tasks when LoRA rank is set to 16, 32, 64, 128 and 256. The optimal LoRA rank is 64, as smaller ranks may not capture diverse tasks, while larger ranks may affect low-rank regularization and cause overfitting. Thus, for the LocalEvolve dataset, the best downstream performance is achieved with LoRA rank set to 64.

4 Conclusion

In this study, we improve the performance of LLM in local service scenarios by introducing a domain-specific corpus, LocalEvolve, and developing a customized LLM, LocalAdapt. Our Multi-Task Mapping (MTM) system effectively transforms structured data into Fundamental Reasoning Units (FRUs), enhancing the model's local life domain performance. Results show LocalAdapt outperforms baselines in various tasks, highlighting our approach's potential in enhancing LLMs' domain-specific applicability.

The empirical analysis confirms FRUs' efficacy in bridging local life industry and general data. Our work advances LLM refinement for specialized applications and domain-specific model adaptation research.

5 Limitations and Future Work

Our work is still in its early stages, and we outline the current limitations and future research directions as follows:

1. Exploring more suitable data ratios. At present, we have only evaluated the performance of LocalAdapt fine-tuning on three tasks. Finding an appropriate data ratio to ensure that all local living services can achieve performance improvements after fine-tuning with LocalAdapt remains a challenging task.

2. Automating the summarization of reasoning templates. The reasoning templates in our current work are manually selected from those constructed by GPT-4. In future work, we can consider developing a method for automatically evaluating the fit of templates, reducing the labor costs involved.

3. Mitigating the hallucination issue during the inference process. The current model's inference process still contains some errors. In subsequent work, we can consider analyzing the impact of pre-training tasks on the hallucination phenomenon during the inference process after fine-tuning.

References

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Acl 2023 tutorial: Retrieval-based language models and applications. *ACL 2023*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023a. Chatlaw: Open-source legal large language model with integrated external knowledge bases.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023b. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation.

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Hongjing Qian, Yutao Zhu, Zhicheng Dou, Haoqi Gu, Xinyu Zhang, Zheng Liu, Ruofei Lai, Zhao Cao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Webbrain: Learning to generate factually correct articles for queries by grounding on large web corpus.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis insights from training gopher.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin

Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

Murray Shanahan. 2023. Talking about large language models.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. 2021. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2:65–68.

A Appendix

A.1 Supervised Evaluation Dataset

We evaluate pre-trained models on product tagging, product matching, and query matching tasks.

The product tagging task aims to generate corresponding labels for products based on their names and descriptions, with labels derived from product category tags on the platform website.

The product matching task aims to compare whether two products under the same point of interest (POI) are identical based on their descriptions, with inference results in the labels coming from merchant notes, reasoning generated by GPT4, and verified by human experts.

The query matching task aims to compare whether two sentences from user inquiries in a customer service context express the same meaning, with inference results in the labels coming from customer service ticket notes, reasoning generated by GPT4, and verified by human experts.

Since product data and user queries are proprietary to the platform, we will not disclose the complete evaluation dataset to the public, thus preventing the leakage of private information and reducing security risks.

A.2 Human Annotation Guidelines

A.2.1 Readability Annotation Guidelines

For readability, we ask annotators to focus on the fluency and grammaticality of the reasoning and provide the following guidance:

Annotators determine whether the sentence is complete. If the sentence is incomplete, annotators rate both fluency and grammaticality as 1.

Annotators can understand the meaning of a complete sentence, but there are many grammatical issues in the sentence. Annotators rate fluency and grammaticality as 2 or 3.

Annotators can easily understand the meaning of the sentence, with only minor grammatical issues. Annotators rate fluency and grammaticality as 4 or 5.

A.2.2 Factuality Annotation Guidelines

For factuality, we ask annotators to focus on the correctness and completeness of the reasoning logic. The guidelines are as follows:

We ask annotators to check whether the given reasoning process correctly explains the concepts and whether the reasoning process leading to the conclusion contains complete reasoning logic.

If the match rate is less than 30%, annotators rate the score as 1 or 2;

if the match rate is greater than 30% and less than 60%, annotators rate the score as 3;

if the match rate is greater than 60% and less than 100%, annotators rate the score as 4 or 5.

A.3 Why Choose LoRA

We compare various methods for continuing pre-training the backbone model. Fig 1 demonstrates that LoRA (rank 64) significantly outperforms full parameter updates and P-Tuning V2 on query matching task. LoRA's superior performance is attributed to its regularization effect and preservation of pre-trained knowledge, preventing overfitting and improving generalization. In contrast, full-parameter fine-tuning may disrupt pre-trained knowledge, leading to suboptimal performance.

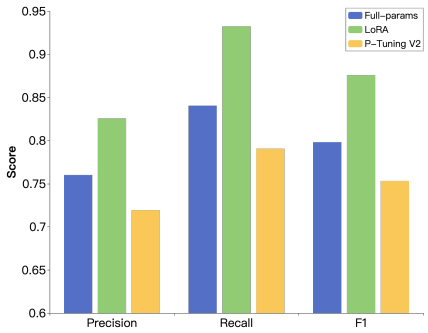

Figure 1: Metrics for LocalAdapt on query matching task with different pre-training strategies.

A.4 Example Prompts for Reasoning Task Design

We demonstrate the instructions driving the design of logical reasoning tasks for GPT-4 in Fig 2.

A.5 Case Study

In Figs 3, 4 and 5, we demonstrate the performance of various models on product tagging, product matching, and query matching tasks. Our LocalAdapt achieves state-of-the-art (SOTA) results in all three tasks.

Instruction:

你是一个商业逻辑教育专家，需要根据生活服务领域结构化的商品数据设计一段逻辑推理文本以提高青少年的商业认知和推理能力，你的目标是：
1. 增强青少年对生活服务相关商品的了解
2. 培养青少年商业思维和对商品的敏锐度
3. 提高青少年思维能力和思考分析

要求：
描述应当语言流畅，并使青少年容易理解。

你的回答包括：###逻辑任务和###具体的设计文本。
###逻辑任务：包括但不限于多步推理，主次判断，逆否命题转述，同义推理，因果推断，总结归纳，指代分析等等。
###具体的设计文本：鼓励多样，有深度的任务设计，形式包括但不限于推理文本，简述，选择，问答，逻辑符号推导等等。注意如果设计了问答，简述和选择的形式需要先描述商品信息，再设计问题，最后给出参考答案。

输出格式：
###逻辑任务：
###具体的设计文本：

Input:

商品名称：【私人订制】蛋白护理烫/染（男士专用）
商品描述：这款商品是一个私人订制的美发服务，针对男性顾客推出的蛋白护理烫/染服务，旨在为男性顾客打造更健康、更时尚的发型。

Output:

###逻辑任务：
商品领域推断
###具体的设计文本：
名为【私人订制】蛋白护理烫/染（男士专用）的商品
是一个私人订制的美发服务，针对男性顾客推出的蛋白护理烫/
染服务，旨在为男性顾客打造更健康、更时尚的发型，请问这个商品属于哪个领域？
A. 丽人-美发-美发
B. 美食-火锅-小火锅
C. 美容-护肤-面膜
D. 服装-女装-连衣裙
参考答案：A

Figure 2: Example Prompts for Reasoning Task Design.

Prompt	Model	Response
Instruction: Assign tags to the following product based on its title and content description	ChatGPT	Makeup removal and cleansing - Essence moisturizing - Hydration and firming
	Chinese-llama-7b	Cleansing - Beauty
Input: [Product Name] Facial Moisturizing SPA [Product Details] Makeup removal and cleansing || Deep hydration and peptide firming || Aier's refined series of lotion || Age-defying and rejuvenating || Hydro-oxygen warmth || Micro-conducting essence	Chinese-llama-13b	Facial cleansing - Skin management - Beauty
	Chinese-llama-13b(LoRA)	Facial cleansing - Facial hydration - Beauty
	LocalAdapt	Facial cleansing - Facial hydration - Skin management - Beauty
	LocalAdapt w/o FRUs	Facial hydration - Beauty
	Label	Facial cleansing - Facial hydration - Skin management - Beauty

Figure 3: Example output for product tagging task, where LocalAdapt achieves SOTA.

Prompt	Model	Response
Instruction: Determine whether the following two products are the same based on descriptions, and provide the reason.	ChatGPT	Different products. Product 1 is from a live-streaming platform, while Product 2 does not specify the source and is targeted at VIPs.
	Chinese-llama-7b	Different products. The services offered differ, but the target audience is consistent.
	Chinese-llama-13b	Different products. The services offered are inconsistent, but both target one adult and one child.
Input: Product 1: Bayannur Sesame Open Door Water World, live-streaming benefits, one adult and one child play pass vs Product 2: Sesame Open Door Water World - [VIP Parent-Child Ticket] one adult and one child pass	Chinese-llama-13b(LoRA)	Different products. The services offered are inconsistent, but the target audience is consistent (one adult and one child).
	LocalAdapt	Same product. The services offered are consistent, and the target audience is consistent (one adult and one child).
	LocalAdapt w/o FRUs	Different products. The services offered are inconsistent, but the target audience is consistent.
	Label	Same product. The services offered are consistent, and the target audience is consistent (one adult and one child).

Figure 4: Example output for product matching task, where LocalAdapt achieves SOTA.

Prompt	Model	Response
Instruction: Determine whether the meanings of the following two sentences are the same, and provide the reason.	ChatGPT	Not the same. In Sentence 1, the user is asking about cupping therapy, while in Sentence 2, the user is asking about dampness removal and mentions moxibustion. The focus of the inquiries is different.
	Chinese-llama-7b	Not the same. Sentence 1 asks about cupping therapy, while Sentence 2 asks about moxibustion, making them different.
Input: Sentence 1: "Does this store offer cupping therapy?" vs. Sentence 2: "Can they help with dampness removal here, like with moxibustion?"	Chinese-llama-13b	Not the same. Sentence 1 inquires if the store offers cupping therapy, while Sentence 2 asks about dampness removal, making them different.
	Chinese-llama-13b(LoRA)	Not the same. Sentence 1 asks about cupping therapy, while Sentence 2 asks about dampness removal. Cupping therapy is not a dampness removal method, making them different.
	LocalAdapt	The same. Sentence 1 asks about cupping therapy, while Sentence 2 asks about dampness removal. Cupping therapy is also a dampness removal method, making them the same.
	LocalAdapt w/o FRUs	Not the same. Sentence 1 asks about cupping therapy, while Sentence 2 asks about moxibustion. Cupping therapy and moxibustion are different.
	Label	The same. Sentence 1 asks about cupping therapy, while Sentence 2 asks about dampness removal. Cupping therapy is one of a method of dampness removal, making them the same.

Figure 5: Example output for query matching task, where LocalAdapt achieves SOTA.