Deploying Multi-task Online Server with Large Language Model

Yincen Qu[1], **Hengyue Liu**[2], **Kun Wang**1, **Xiangying Dai**[1], **Hui Zhou**[1], **and Chao Ma**[2]
[1]Trip.com Group, Shanghai, China
[2]Independent Researcher
{yc.qu,kun_wang,xy.dai,hzhoug,ma_c}@trip.com
hengyueliu23@gmail.com

Abstract

In the industry, numerous tasks are deployed online. Traditional approaches often tackle each task separately by its own network, which leads to excessive costs for developing and scaling models, especially in the context of large language models. Although multi-task methods can save costs through parameter sharing, they often struggle to outperform single-task methods in real-world applications. To tackle these challenges, we present a three-stage multi-task learning framework for large language models. It involves task filtering, followed by fine-tuning on high-resource tasks, and finally fine-tuning on all tasks. We conducted comprehensive experiments in single-task and multi-task settings. Our approach, exemplified on different benchmarks, demonstrates that it is able to achieve performance comparable to the single-task method while reducing up to 90.9% of its overhead.

1 Introduction

In the industry, numerous natural language processing (NLP) tasks are deployed online, and all tasks are required to serve with punctuality and high accuracy. As the number of tasks increases, the demand for resources also grows. Preventing resource requirements from growing linearly with the number of tasks becomes one of the most critical challenge in cost-saving.

Traditional approaches tackle each task separately by its own network and pipeline. This leads to excessive workloads for development and maintenance, as well as increased latency and resource usage. Moreover, in the context of large language models (LLMs), it may also lead to excessive costs for scaling up models for each task. We propose utilizing multi-task serving to deploy LLMs instead of single-task serving. *single-task serving* and *multi-task serving* are two types of online serving strategies, and their paradigms are shown in

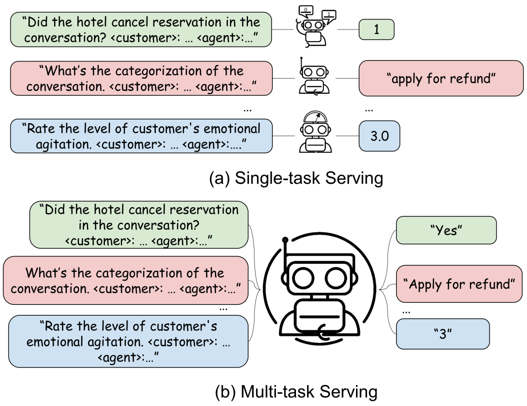

Figure 1: Two types of online serving strategies. (a) Independent single-task models are trained and deployed for each task. (b) One multi-task model is trained and deployed for all tasks.

Figure 1. Compared to single-task serving, multi-task serving reduces deployment efforts and saves more memory due to the sharing mechanism, thus alleviating resource wastage.

However, in real-world applications, multi-task methods often struggle to match the performance of single-task methods due to the data imbalance and task heterogeneity. Data imbalance consistently leads to overfitting in low-resource tasks. This occurs because early stopping is not a feasible solution for high-resource tasks; these tasks require many more epochs to converge. Additionally, heterogeneity may result in negative transfer between tasks. Different tasks require different gradient direction in model optimization, and tasks that are too divergent may conflict in terms of gradient direction.

In this paper, we propose a three-stage framework: filtering dissimilar tasks, fine-tuning on high-resource tasks, and fine-tuning on a mixture of all tasks. The task filtering strategy prevents the negative transfer between heterogeneous tasks. The strategy of fine-tuning on high-resource tasks fol-

Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, pages 483–495
January 19–24, 2025. ©2025 Association for Computational Linguistics

lowed by fine-tuning on the mixture effectively effectively enables early stop by allowing different tasks to have different training epochs, thus preventing overfitting of low-resource tasks or underfitting of high-resource tasks.

Through an extensive empirical study, we find that our algorithm achieves closer performance to the single-task setting compared to other multi-task baselines. We observed that the improvement in multi-task performance mainly comes from the sampling strategy, the task filtering and domain-specific continual pre-training.

Our main contributions can be summarized as follows:

(1) We propose a framework for multi-task serving that utilizes LLMs to facilitate the multi-task method that simultaneously handles multiple tasks and achieves comparable performance of that of the single-task method.

(2) We run a comprehensive set of experiments that suggest our scheme is practical across different benchmarks and capable of substituting for tasks trained in the single-task method. We also performed extensive experiments to gauge the importance of each of our components, such as task selection and sampling strategy.

(3) Our model was deployed to production to provide serving for a total of 11 downstream tasks. Compared to single-task serving, our model achieves comparable performance. We estimate that our system can reduce the total serving costs by up to 90.9% compared to single-task serving.

2 Related Works

Multi-task Learning. Multi-task learning (MTL) involves training a single model on multiple tasks simultaneously. Several studies have explored the effectiveness of MTL in various domains, such as natural language processing (Jean et al., 2019; Liu et al., 2019; Wei et al., 2022; Peng et al., 2023), computer vision (Kendall et al., 2018; Kang et al., 2020). Recently, T5 (Raffel et al., 2020), ExT5 (Aribandi et al., 2022) and Muppet (Aghajanyan et al., 2021) have been proposed to explore the application of Multi-Task Learning (MTL) techniques in Large Language Models (LLMs). However, they selected different checkpoints for each task without aiming to train the model to handle tasks simultaneously. Moreover, most of recent works such as FLAN (Chung et al., 2022), T0 (Sanh et al., 2022), and GPT-3 (Brown et al., 2020), etc., focused on zero-shot or few-shot performance and neglected to compare with the full fine-tuning method for single tasks. However, we found that it is not trivial to surpass single-task full-parameter fine-tuning method.

Data Imbalance. Due to the prevalence of imbalanced data distribution, data balancing has attracted increasing attention. Researchers have proposed static sampling to achieve a more balanced data distribution, which includes class-balanced sampling (Mahajan et al., 2018), temperature-scaled sampling (Pires et al., 2019). Previous works (Kurin et al., 2022; Xin et al., 2022) show evidence that static sampling approach yield optimal results in data rich regime (high-resources). Recently Chung et al. (2023); Choi et al. (2023) proposed to prevent model to overfit on the low-resource language in static sampling during multilingual pre-training. They focus on the performance of similar tasks under data imbalance, such as translation between different languages and multilingual pre-training. In our work, we integrated dissimilar tasks and explored whether data imbalance and heterogeneity could hinder multi-task performance.

3 Preliminaries

3.1 Sampling Strategies

In this section, we present three common sampling strategies that aim to re-balance the task distribution. We will utilize these three sampling methods as baselines for subsequent experiments.

Instance-balanced sampling. Instance-balanced sampling refers to sampling examples from each task based on the total size of each task's dataset. Specifically, the empirical distributions for different tasks are as follows.

$$p_l = \frac{n_l}{\sum_{l' \in L} n_{l'}} \quad (1)$$

where n_l is the data size of task l. Here data points from task l will be sampled with the probability p_l, which is proportional to the cardinality n_l of the task in the training set.

Class-balanced sampling. Class-balanced sampling refers to sampling examples from each task with equal probability. In each batch, each example is sampled uniformly from one of the tasks used for training.

Temperature-scaled sampling. Temperature-scaled sampling refers to re-scaling the sampling

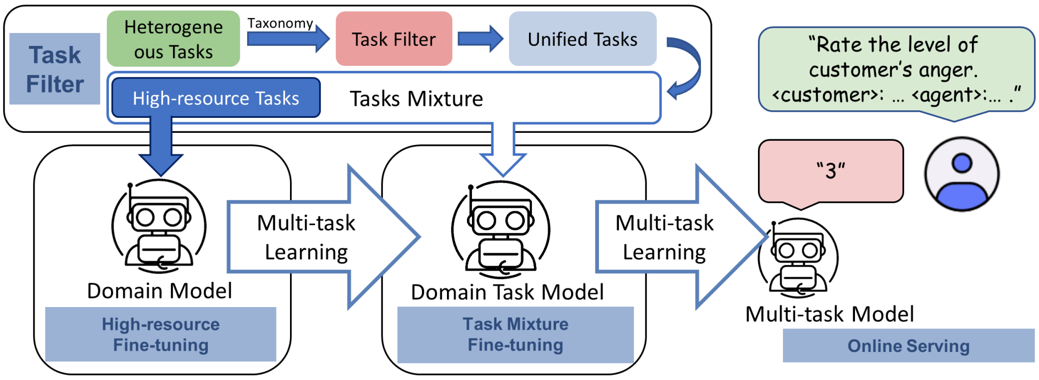

Figure 2: Pipeline of the proposed method. It starts with domain-specific continual pre-training, where the model undergoes self-supervised learning using domain-specific data. Next, we perform multi-task fine-tuning on high-resource tasks. Then, we perform multi-task fine-tuning on all tasks, enabling the model to learn from a mixture of tasks simultaneously. Finally, the multi-task model is deployed online to serve different tasks.

rates by a temperature τ. It uses a distribution q defined by exponentiating p.

$$q_l = \frac{p_l^{1/\tau}}{\sum_{l' \in L} p_{l'}^{1/\tau}} \tag{2}$$

When $\tau = 1$, this approach is equivalent to instance-balanced sampling. As τ increases, the mixing becomes more uniform across tasks. When $\tau \to \infty$, this approach is equivalent to class-balanced sampling. In practice, commonly used values for τ are $(1.43, 2, 3.33)$ (Pires et al., 2019; Blevins et al., 2022; Conneau et al., 2020; Xue et al., 2021).

3.2 Problem Setting

Given a set of target tasks L, our framework is dedicated to find the parameters θ of a model $\mathcal{F}$ that can achieve comparable performance to the single-task model in as many tasks as possible. This differs slightly from the common goal of multi-task learning, which aims to achieve high average performance across all training tasks. We refer to those tasks that attain 99% of the full fine-tuning baseline as qualified tasks and our goal is to deploy as many qualified tasks as possible with a single model. Besides, in real-world application, we have tasks of different types, each with varying amounts of training samples. Thus, we have to take heterogeneity and data imbalance into consideration.

4 Methodology

Our proposed framework in Figure 2 features a pipeline that consists of three steps: 1) Task filtering; 2) High-resource task fine-tuning; 3) Tasks mixture fine-tuning. We provide a detailed breakdown of these steps below.

4.1 Task Filter

4.1.1 Filtered Task

To prevent negative transfer between different tasks, it's important to filter out inappropriate tasks. We found that generation tasks and classification tasks would hinder each others' performance in multi-task training, as evidenced in the experiment sections. The output of classification tasks is fixed, whereas the output of generation tasks is flexible. For instance, the CLUE (Xu et al., 2020) tasks encompass single sentence classification, sentence pair classification, and machine reading comprehension. We categorize the single sentence classification and sentence pair classification as classification tasks, and machine reading comprehension as generation tasks.

Moreover, we also investigated that whether differences in input (such as single sentences or sentence pairs) or output (such as binary classification or multi-class classification) would further impede performance. We found that the more similar the tasks are, the higher the multi-task performance can be achieved, and the greater the number of qualified tasks becomes.

4.1.2 Unified Tasks

In order to train a unified model for various tasks, we cast all of the collected tasks into a format called "text-to-text." This format requires the model to be fed with some text for context and then generate output text for individual tasks. To indicate the specific task, we add a task-specific text prefix to the original input sequence prior to inputting it into the model.

4.2 Multi-task Fine-tuning

For tasks with imbalanced data, we utilize the multi-task learning approach to balance the performance of all tasks. However, the aforementioned sampling strategies are not ideal, as they sample all tasks with a constant probability throughout the entire training process, leading to over-fitting of low-resource tasks, while high-resource tasks still require learning.

We divide the tasks into two groups: high-resource and low-resource tasks. Since we have a variety of tasks with different training saturation steps, it is unfeasible to categorize them based on the amount of training data as in Choi et al. (2023). Instead, we categorize tasks based on the training saturation steps in the single-task setting. If a task achieves overfitting in fewer than 5 epochs, we refer to it as a "low-resource task." If a task achieves overfitting after more than 5 epochs, we refer to it as a "high-resource task."

For these task groups, we perform two-stage training, including high-resource task fine-tuning and tasks mixture fine-tuning.

(1) **High-resource task fine-tuning.** For high-resource tasks, we utilize the method of instance-balanced sampling to train them, given that they each have a similar amount of training data.

(2) **Tasks mixture fine-tuning.** After fine-tuning the model on high-resource tasks, we proceed to fine-tune it on the full mixture of tasks. We utilize temperature-scaled sampling and impose an artificial limit on dataset size to train all downstream tasks simultaneously. We set an artificial limit (K) on the dataset size to prevent over-fitting. The adjusted distribution of different tasks is as follows.

$$p_l = \frac{\min(n_l, K)}{\sum_{l' \in L} \min(n_{l'}, K)} \quad (3)$$

$$q_l = \frac{p_l^{1/\tau}}{\sum_{l' \in L} p_{l'}^{1/\tau}} \quad (4)$$

5 Experiments

In the following sections, we apply our proposed training method to CLUE (Xu et al., 2020) tasks and our domain application tasks. In the CLUE experiments, we show that inappropriate sampling strategy will lead to multi-task performance degradation and different tasks taxonomies also hinder multi-task performance. In the domain-related application tasks, we scale up the number of tasks, all of which are related to the customer service field, and show that our method remains equally effective in the real-world applications.

5.1 CLUE Tasks

5.1.1 Experiment Setup

The CLUE benchmark (Xu et al., 2020) is synthetic, consisting of six classification datasets: CWSC, TNEWS, CSL, AFQMC, IFLYTEK, and OCNLI. We provide details and references in Appendix B. For each task, we used accuracy rate as the primary evaluation metric. We reported the macro-average accuracy across all tasks within the benchmark. In the multi-task setting, we also provided the count of qualified tasks, which are defined as those achieving 99% of the performance of their single-task counterparts. To measure the parameter and computational efficiency, we introduced a ratio: the number of qualified tasks divided by the number of models deployed. This ratio is 1 for the single-task baseline, as it deploys one model per task. For multi-task models, the ratio is calculated as 1 divided by the number of qualified tasks. This metric is labeled as "overhead" in the header of Table 1.

In the experiment, we take the 7B Qwen2 (Yang et al., 2024) and 8B LLaMA3 (Touvron et al., 2023) as the base model. We present a comparative analysis of our two-stage sampling method against five benchmark approaches: few-shot prompting, single-task fine-tuning, instance-balanced sampling, class-balanced sampling, and UniMax (Chung et al., 2023). In the case of few-shot prompting, we prepend five random training instances $(q_i, a_i)_i$ as the example to guide the model's input.

5.1.2 Main Results

Table 1 shows the experimental results on the CLUE benchmark. We observed that an inappropriate sampling strategy would hinder the multi-task performance. The few-shot method performed the worst, suggesting that it is not yet capable of directly replacing current fine-tuning methods, particularly for multi-class classification tasks. Our 2-stage sampling strategy achieved the best performance among all sampling approaches, delivering the highest number of qualified tasks. Compared to our method without the two-stage training process,

Models	Methods	CWSC (Accuracy)	TNEWS (Accuracy)	CSL (Accuracy)	AFQMC (Accuracy)	IFLYTEK (Accuracy)	OCNLI (Accuracy)	Avg.	Num.	Overhead
LLaMA	Single-task	70.22	58.71	87.06	73.98	58.39	79.23	71.26	6	100%
	Few-shot	65.07	13.82	62.10	46.80	14.57	54.47	42.81	0	-
	Instance-balanced	68.75	56.20	85.02	73.52	59.13	80.05	70.44	3	33.3%
	Class-balanced	69.12	57.39	83.34	74.05	59.33	80.66	70.64	3	33.3%
	UniMax	68.01	56.55	84.65	74.75	57.56	82.42	70.65	2	50.0%
	ours	70.06	57.31	87.51	74.68	58.79	80.83	71.53	**5**	**20.0%**
	ours (w/o 2-stage)	70.22	56.32	87.03	73.03	60.11	81.91	**71.76**	4	25.0%
Qwen	Single-task	71.69	60.16	83.54	74.12	58.31	86.52	72.39	6	100%
	Few-shot	65.44	22.84	66.62	53.54	17.63	73.00	49.85	0	-
	Instance-balanced	68.75	59.51	82.81	74.44	59.56	82.76	71.30	3	33.3%
	Class-balanced	71.69	58.20	85.56	74.12	58.54	80.01	71.35	4	25.0%
	UniMax	70.59	59.63	82.74	74.14	59.68	83.37	71.69	4	25.0%
	ours	71.32	59.59	86.03	74.56	58.86	83.67	**72.33**	**5**	**20.0%**
	ours (w/o 2-stage)	71.32	58.32	86.60	74.18	59.36	82.99	72.13	4	25.0%

Table 1: Main results on 6 tasks and the average performance across them. The performance is evaluated on the development set. "Avg." refers to the macro average per-task performance of downstream tasks. "Num." refers to the amount of the qualified tasks. All metrics for tasks are multiplied by 100. Shaded numbers indicate that they attain 99% of the single-task fine-tuning baseline.

the two-stage training only marginally improves average performance. However, it significantly increases the number of qualified tasks. We hypothesize that this enhancement is due to the high-resource task training helps to balance the diverse training steps across various tasks.

Moreover, we noted that LLaMA's macro-average performance on Chinese tasks is inferior to that of Qwen, likely due to insufficient training on Chinese corpora. Given that Qwen has been pre-trained on Chinese corpora, it demonstrates superior multi-task performance in Chinese. Consequently, in Section 5.2, we carry out additional experiments to assess the performance of the generic model in comparison to the model that has undergone domain-specific pre-training.

5.1.3 Taxonomy Impact

In this section, we investigate the impact of taxonomy granularity on multi-task performance. We introduced the machine reading comprehension task CMRC into our task mixture, and trained a multi-task model with this expanded dataset. Unlike the original set of six classification tasks, CMRC, as a generation task, has a flexible output format. From the Table 2, we found that training generation and classification tasks concurrently significantly impacts the overall performance. It is particularly notable that the performance of the classification tasks not only lags behind their single-task counterparts but also fails to match the performance of the multi-task model that was trained only on classification tasks.

To delve deeper into whether task similarity can enhance performance, we categorized the tasks into groups based on differences in input and output types: single-sentence, sentence-pair, binary classification, and multi-class classification. A more detailed presentation of the tasks and their results is provided in Appendix D. From Table 9, we noticed that increased task similarity correlates with improved performance. However, the "overhead" metric does not decrease, as the number of models also rises. To meet our objective of cost saving, a lower overhead metric is desirable. Consequently, we decided against further subdividing these tasks into more similar categories.

Methods	Generation	Classification	Avg.	Num.
Single-task	51.27	72.39	69.37	7
instance-balanced	47.61	70.64 (71.30)	66.82	1 (3)
class-balanced	52.94	70.58 (71.35)	68.02	2 (4)
ours	48.79	71.87 (72.33)	68.57	3 (5)

Table 2: Taxonomy impact of on generation and classification CLUE tasks. The number in brackets refers to the multi-task performance trained solely with the classification tasks.

5.2 Application Tasks

In this section, we expand from a six-task setting to the setting with dozens of tasks, to verify whether task filtering and sampling methods would affect the multi-task performance.

5.2.1 Experiment Setup

We tested with 17 classification tasks, which are all related to the domain of customer service. The details of these tasks are demonstrated in Appendix C. We also reported macro average performance, the number of qualified tasks, and the overhead metrics for each method.

We took Qwen2 7B as the base model. We provided a comparison of our method with 5 baseline methods, as in the previous section. In addition, we performed domain-specific continual pre-training on Qwen2 to obtain Qwen$_d$. The details of the continual pre-training will be demonstrated in the Appendix E. We report the multi-task performance of the generic model Qwen and Qwen$_d$ to further investigate whether domain pre-training can enhance multi-task performance.

5.2.2 Application Results

Table 3 shows the experimental results on the industry benchmark. We found that when task number increases, inappropriate sampling strategy has more obvious effect on the multi-task performance. Our method outperforms other sampling baselines by consistently enhancing both the macro-average performance and the number of qualified tasks. With an overhead of only 9.1% compared to the single-task approach, our method can potentially reduce the serving cost by up to 90.9% relative to the single-task method.

We observed that Qwen$_d$ exhibits relatively high performance compared to Qwen. Specifically, Qwen$_d$ demonstrates a higher average performance than Qwen. Furthermore, any sampling method with Qwen$_d$ results in a greater number of qualified tasks than with Qwen. We attribute these improvements to domain adaptation. Given the substantial disparity between customer service conversations and the general domain text corpora utilized by original LLMs, incorporating domain-specific knowledge through continuous pre-training significantly aids in downstream task performance. Moreover, the amount of required updates for each task is reduced, leading to less conflict in gradient directions when training tasks concurrently.

5.2.3 Taxonomy Impact

Consistent with our previous experiment, we incorporated a generation task into our task mixture and trained them jointly with Qwen$_d$. From Table 4, we found that regardless of the sampling strategy employed, both classification and generation tasks experienced a significant decline in performance compared to their single-task counterparts. This suggests that the negative impact is indeed present, likely due to the substantial differences between the tasks.

Models	Methods	Avg.	Num.	Overhead
Qwen	Single-task	88.64	17	100%
	Few-shot	49.68	0	-
	Class-balanced	85.34	5	20.0%
	Instance-balanced	85.82	5	20.0%
	Unimax	86.33	8	12.5%
	ours	87.19	**9**	**11.1%**
Qwen$_d$	Single-task	89.65	17	100%
	Few-shot	54.27	0	-
	Class-balanced	85.29	5	20.0%
	Instance-balanced	86.05	6	16.7%
	Unimax	86.91	8	12.5%
	ours	87.74	**11**	**9.1%**

Table 3: Main results on 17 application tasks. "Avg." refers to the macro average performance. "Num." refers to the amount of the qualified tasks.

We then categorized the classification tasks into three types: binary classification, ordinal classification, and multi-class classification, and trained separate models for each category. From Table 5, we also observed that performance improved with the more granular categorization of tasks. However, since this approach required multiple models for these tasks, the overhead metric did not show improvement.

Methods	Generation	Classification	Avg.	Num.
Single-task	57.13	88.64	86.89	18
class-balanced	54.17	84.97 (85.29)	83.26	3 (5)
instance-balanced	52.58	85.09 (86.05)	83.28	3 (6)
ours	53.69	85.42 (87.74)	83.67	6 (11)

Table 4: Taxonomy impact on generation and classification application tasks.

Methods	Binary	Ordinal	Multi.	Avg.	Num.	Overhead
Single-task	87.62	95.49	94.05	89.65	17	100%
instance-balanced	87.43	94.19	92.34	89.09	9	16.67%
class-balanced	87.12	95.25	93.77	89.25	10	16.67%
ours	87.46	95.21	93.48	89.43	10	14.29%

Table 5: Taxonomy impact on binary, ordinal and multi-class classification application tasks.

6 Conclusion

In this work, we demonstrated the benefits of task filtering and two-stage multi-task training for multi-task optimization in the presence of task imbalance and heterogeneity. Through a variety of experimental setups, we show that inappropriate sampling and task selection strategies may hinder the overall multi-task performance. Our method, though straightforward, is a viable alternative to models trained with the single-task approach, potentially resulting in substantial cost savings.

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5799–5811. Association for Computational Linguistics.

Vamsi Aribandi, Yi Tay, Tal Schuster, and et al. 2022. Ext5: Towards extreme multi-task scaling for transfer learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3575–3590. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. 2020. 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*.

Dami Choi, Derrick Xin, Hamid Dadkhahi, Justin Gilmer, Ankush Garg, Orhan Firat, Chih-Kuan Yeh, Andrew M. Dai, and Behrooz Ghorbani. 2023. Order matters in the presence of dataset imbalance for multilingual learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hyung Won Chung, Xavier Garcia, Adam Roberts, and et al. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Hyung Won Chung, Le Hou, Shayne Longpre, and et al. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, and et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Sébastien Jean, Orhan Firat, and Melvin Johnson. 2019. Adaptive scheduling for multi-task learning. *CoRR*, abs/1909.06434.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020. Decoupling representation and classifier for long-tailed recognition. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7482–7491. Computer Vision Foundation / IEEE Computer Society.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and Pawan Kumar Mudigonda. 2022. In defense of the unitary scalarization for deep multi-task learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, and et al. 2018. Exploring the limits of weakly supervised pretraining. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11206 of *Lecture Notes in Computer Science*, pages 185–201. Springer.

Zhiyuan Peng, Vachik S. Dave, Nicole McNabb, and et al. 2023. Entity-aware multi-task learning for query understanding at walmart. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 4733–4742. ACM.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, and et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, page 20. IEEE/ACM.

Victor Sanh, Albert Webson, Colin Raffel, and et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *CoRR*, abs/1909.08053.

Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Tianwen Wei, Jianwei Qi, and Shenghuan He. 2022. A flexible multi-task model for BERT serving. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 785–796, Dublin, Ireland. Association for Computational Linguistics.

Derrick Xin, Behrooz Ghorbani, Justin Gilmer, Ankush Garg, and Orhan Firat. 2022. Do current multi-task optimization methods in deep learning even help? In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Liang Xu, Hai Hu, and et al. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, and et al. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Hyper-parameter	CLUE	Application
Learning rate	3e-5	3e-5
Batch Size	1	1
Gradient accumulation	8	8
Epoch (stage 1)	1	1
Epoch (stage 2)	10	10
K	20000	8000
τ	2	3.33

Table 6: Hyper-parameters used in our experiments.

An Yang, Baosong Yang, Binyuan Hui, and et al. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, and et al. 2021. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2:65–68.

A Experiment Setting

For a fair comparison, we have capped the training steps for different sampling methods at 15,000. The hyper-parameters (e.g. learning rate, mini-batch size, etc) used in our experiments are summarized in Table 6.

B CLUE Benchmark

Chinese Winograd Schema Challenge (CWSC). The CWSC dataset is designed for anaphora and coreference resolution. The model is asked to determine if a pronoun and a noun phrase whithin a sentence refer to the same entity. It's a binary classification task. It mirrors similar English datasets and consists of sentences carefully selected from 36 modern Chinese literary works. Their anaphora relations are meticulously annotated by linguists, resulting in a collection of 1,838 questions.

TouTiao Text Classification (TNEWS). TNEWS consists of Chinese news from TouTiao, comprising 73,360 titles in total. Each title is assigned a label among 15 different news categories, such as finance, technology and sports. The goal of this task is to predict which category the title belongs to.

IFLYTEK. The IFLYTEK is a Chinese multi-class classification dataset, comprising 17,332 descriptions of mobile applications. The objective is to categorize each description into one of the 119 available categories, including but not limited to food, car rental, and education. A data filtering method akin to that employed for the TNEWS dataset has been utilized in this process.

Chinese Scientific Literature (CSL). CSL dataset comprises abstracts from Chinese scientific papers and their associated keywords, sourced from various core journals across natural and social sciences. This dataset includes artificially generated keywords using the tf-idf method, which are combined with genuine keywords. The task involves identifying whether the provided keywords for a given abstract are authentic to the paper. This primarily assesses the models' capacity to determine if the keywords accurately encapsulate the content of the document.

Ant Financial Question Matching Corpus (AFQMC). AFQMC originates from Ant Technology Exploration Conference (ATEC) Developer competition. It presents a binary classification challenge designed to determine if two given sentences share a similar meaning.

Original Chinese Natural Language Inference (OCNLI). OCNLI is a natural language inference dataset using a similar methodology to the MNLI dataset. It consists of 56,000 inference pairs across five different categories: news, government documents, fiction, TV transcripts, and telephone transcripts. The source material for the premises is Chinese, and hypotheses were authored by university students specializing in linguistics. The level of agreement among the annotators is comparable to that of MNLI.

Chinese Machine Reading Comprehension (CMRC). CMRC is a machine reading comprehension dataset that is based on span extraction. It comprises approximately 19,071 questions, all of which are human-annotated and sourced from Wikipedia passages. Each entry in the CMRC dataset includes a context, a question, and the corresponding answer. The answers are segments of text extracted directly from the context.

Taxonomy	Task	Metrics	$\lvert\mathcal{D}\rvert$
Classification			
Single Sentence	CWSC	acc.	947
	TNEWS	acc.	49,726
	IFYTEK	acc.	11,425
Sentence Pair	CSL	acc.	19,836
	AFQMC	acc.	6,564
	OCNLI	acc.	50,437
Generation			
Reading Comprehension	CMRC	EM.	10,143

Table 7: Examples of different tasks. $|D|$ refers to the number of training instances.

C Application Tasks

Reservation Cancellation (RC). Reservation cancellation refers to the hotel canceling a confirmed booking and not allowing guests to check-in. This is a binary classification problem where the input is a conversation, and we need to determine whether there is a booking cancellation mentioned in the conversation. Depending on the source of the input, which can be either from a phone call or an online chat, the task of reservation cancellation is considered as two separate tasks. The source of phone call is referred to as RC-A (Automatic speech recognition), while the source of online chat is referred to as RC-I (Instant messaging).

Unforseen Circumstances (UC). Unforseen circumstances refers to unforeseeable and uncontrollable circumstances that prevent guests from checking in after the hotel has confirmed a reservation. This is a binary classification problem where the input is a conversation, and we need to determine whether there is a mention of unforeseeable circumstances in the conversation. Depending on the source of the input, which can be either from a phone call or an online chat, unforseen circumstances is considered as two separate tasks. The source of phone call is referred to as UC-A (Automatic speech recognition), while the source of online chat is referred to as UC-I (Instant messaging).

Poaching Guests (PG). Poaching guests refers to persuading or forcing guests to book hotels and pay bills through alternative channels. This is a binary classification problem where the input is a conversation, and we need to determine whether there is a mention of poaching guests in the conversation. Depending on the source of the input, which can be either from a phone call or an online chat, poaching guests is considered as two separate tasks. The source of phone call is referred to as PG-A (Automatic speech recognition), while the source of online chat is referred to as PG-I (Instant messaging).

Insult Detection (ID). Insult detection is a binary classification task that determines whether a customer service representative is insulting the customer. The input for this task is the historical conversation between the customer and the customer service representative.

Complaint Sentiment Analysis (CSA). Complaint sentiment analysis refers to analyzing whether a customer is likely to post negative feedback on public platforms. The input is the customer's historical conversations, and the output is a binary classification indicating whether the conversation is likely to result in negative publicity.

No Room upon check-in (NR). No room upon check-in refers to determining whether a customer has encountered a situation where there is no available room upon their arrival at the hotel. The input is the customer's historical conversations, and the output is a binary classification. Depending on the source of the input, which can be either from a phone call or an online chat, no room upon check-in is considered as two separate tasks. The source of phone call is referred to as NR-A (Automatic speech recognition), while the source of online chat is referred to as NR-I (Instant messaging).

Hotel Shuttle (HS). Hotel shuttle is a binary classification task that determines whether a hotel provides shuttle service, where the input is the conversation between the guest and the hotel.

Invoice and Deposit Matters (IDM). Invoice and deposit issues matters is a binary classification task. The input for this task is the conversation between the guest and the output is a binary classification indicating whether the guest requires an invoice or not.

Customer Service Quality Rating (CSQR). Customer service quality rating task involves evaluating the caliber of service provided during customer interactions. For this purpose, the input data comprises historical conversations between customer service agents and their clients. The task's output is categorized into four distinct levels, numbered from 1 to 4.

Scoring Extreme Emotion (SEE). Scoring extreme emotion involves rating the level of customer agitation based on the dialogues. The resulting score ranges from 1 to 5, reflecting the intensity of their emotional state.

Review Text Classification (RTC) is a multi-label multi-class classification problem for categorizing reviews, where the input is the multi-lingual review texts and the output includes categories related to the review, such as hotel facilities, service attitude, etc.

Car Services Classification (CSC). Car services classification is a multi-label multi-class classification task, where the input is the historical conversation of a customer when taking a taxi, and the output is the categories of taxi-related issues mentioned by the customer.

Email Categorization (EC). Email categorization refers to classifying incoming emails based on their content. By categorizing the emails, they can be assigned to different business lines for processing. This is a multi-classification task where the input is the email content, and the output is the category of the email.

Conversation Summarization (CS) . In the task of conversation summarization, the input consists of the historical dialogues between customer and service agents, and the goal is to produce a concise summary.

Taonomy	Task	Metrics	$\lvert\mathcal{D}\rvert$
Classification			
Binary	RC-A	acc.	17,059
	RC-I	acc.	6,056
	UC-A	acc.	1,950
	UC-I	acc.	8,624
	PG-A	acc.	2,341
	PG-I	acc.	2,108
	ID	acc.	6,884
	CSA	acc.	5,011
	NR-A	acc.	40,397
	NR-I	acc.	19,726
	HS	acc.	1,328
	IDM	acc.	1,200
Ordinal	CSQR	acc.	2,489
	SEE	acc.	9,314
Multiclass	RTC	acc.	8,447
	CSC	acc.	8,168
	EC	acc.	6,564
Generation			
Summarization	CS	EM.	1,822

Table 8: Examples of different tasks. $|D|$ refers to the number of training instances.

D CLUE Taxonomy Impact

For our CLUE dataset, we divided them into two combinations: single-sentence and sentence-pair classification, binary and multi-class classification. The single-sentence classification includes the CWSC, TNEWS, and IFLYTEK tasks, while

Taxonomy	Methods	CWSC (Accuracy)	TNEWS (Accuracy)	CSL (Accuracy)	AFQMC (Accuracy)	IFLYTEK (Accuracy)	OCNLI (Accuracy)	Avg.	Num.	Overhead
-	Single-task	71.69	60.16	83.54	74.12	58.31	86.52	72.39	6	100%
SS	Instance-balanced	70.96	60.42	87.16	74.03	59.13	82.38	72.34	4	50.0%
	Class-balanced	73.16	59.60	87.40	74.63	59.44	83.64	72.97	5	33.3%
	ours	73.14	60.18	87.49	74.32	59.92	83.41	73.08	5	33.3%
BM	Instance-balanced	68.01	60.06	87.10	74.31	59.29	84.79	72.25	4	50.0%
	Class-balanced	73.16	60.10	86.46	70.86	59.60	84.18	72.39	4	50.0%
	ours	72.97	59.91	86.44	73.79	59.90	84.01	72.83	5	33.3%

Table 9: Results on 6 tasks with different dividing strategy.

the sentence-pairs classification includes the OCNLI, CSL, and AFQMC tasks. The binary classification includes the CWSC, CSL, and AFQMC tasks, and the multi-class classification includes the TNEWS, OCNLI, and IFLYTEK tasks. We refer to the division strategy of Single-sentence and Sentence-pairs as "SS", and the division strategy of Binary classification and Multi-class classification as "BM".

We report the detailed performance of each task in Table 9. As before, we also report the macro average performance, the number of qualified tasks, and the overhead. Since we have multiple models for the same benchmark, the calculation method for the "overhead" metric is slightly different from the previous one; we calculate the "overhead" by dividing 1 by the maximum number of qualified tasks per model.

E Continual Pre-training

We continually pre-train the open-source foundation model on pre-processed domain-specific corpus. The following paragraphs illustrate the pre-training process, covering data sourcing, data processing, tokenization, and pre-training strategy.

Data sourcing. We have collected domain-specific and general data, and mixed them together to enhance the model's general and domain-specific knowledge. Specifically, in our domain, we collect proprietary data such as customer service training materials, introductions to tourist attractions and businesses, and domain-related dialogues. Additionally, we also sample partial data from WuDaoCorpora (Yuan et al., 2021) as general data to supplement general knowledge. This produces an approximately 150 GB collection of the pre-training corpus.

Data processing. We establish a comprehensive data processing pipeline to enhance pre-training data quality. This pipeline comprises four modules: document-wise filtering, line-wise corrections, exact deduplication, ML-based filtering, and fuzzy deduplication. Figure 3 outlines the full data processing pipeline. After cleaning the original data, we obtain approximately 20 billion tokens of the domain-specific corpus.

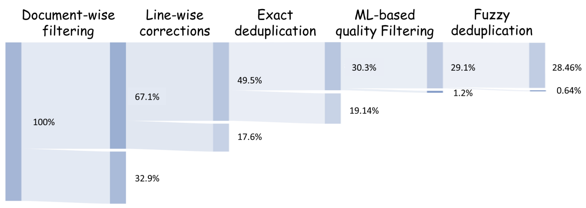

Figure 3: Pipeline of data processing.

Tokenization. We add more domain-specific phrases as new tokens for faster training and inference. We utilize the Byte-Pair Encoding (BPE) algorithm implemented in Sentencepiece (Kudo and Richardson, 2018) to train a domain-specific tokenizer with a vocabulary size of 13,000. We subsequently merge the domain-specific tokenizer into the original tokenizer by taking the union of their vocabularies. Specifically, the vocabulary size of the tokenizer has increased from 125,696 to 127,008. The compression rate in our domain-specific corpus has decreased from 0.6458 to 0.6104.

Pre-training strategy. We utilize the self-supervised learning approach, i.e. causal language modeling, to pre-train our model on the processed corpus. Causal language models refer to models that are trained to predict the next word in a sentence based on the preceding context, capable of capturing the causal relationships between words and generating coherent text. For efficiency, we utilize Megatron (Shoeybi et al., 2019) and DeepSpeed (Rajbhandari et al., 2020) as foundational frameworks, and have integrated flash attention (Dao et al., 2022).

F Few-shot Prompt

We conducted few-shot experiments in the 6 classification tasks, which are CWSC, TNEWS, IFLYTEK, CSL, AFQMC, and OCNLI. Specifically, we design prompts tailored for each task, as shown in Figures 4- 9.

请分辨以下句子中的名词和代词是否指的是同一个实体。
示例1：句子：{裂开的伤口涂满尘土，里面有碎石子和木头刺，我小心翼翼把它们剔除出去。}，词语1：{碎石子和木头刺}，词语2：{它们} -> 否
示例2：句子：{一些候烧者愤愤不平，另一些候烧者忧心忡忡，他们担心二十五年以后怎么办？}，词语1：{另一些候烧者}，词语2：{他们} -> 是
示例3：句子：{我思忖应该找到生前最后的情景，这个最后的情景应该在记忆之路的尽头，找到它也就找到了自己的死亡时刻。}，词语1：{记忆之路}，词语2：{它} -> 否
示例4：句子：{她有时从这些嚼舌根的姑娘跟前走过，知道她们正在说着她如何被那些领导儿子们蹬掉的传言，她仍然向她们送去若无其事的微笑，她们的闲言碎语对于她只是无需打伞的稀疏雨点。}，词语1：{姑娘}，词语2：{她们} -> 是
示例5：句子：{她坐在飞机上，身旁是一个从美国留学归来的博士，这个男人刚刚自己创业，比她大十岁，有妻子有孩子，两个多小时的飞行期间，他满怀激情地向她描述了自己事业的远大前程。}，词语1：{博士}，词语2：{他} -> 是
以下为输入：
{input text} ->

Please distinguish whether the nouns and pronouns in the following sentences refer to the same entity.
Example 1: Sentence: {The cracked wound is covered with dust, with gravel and splinters inside, I carefully remove them.}, Word 1: {Gravel and splinters}, Word 2: {Them} -> No
Example 2: Sentence: {Some candidates are indignant, while others are anxious, they worry about what to do in twenty-five years?}, Word 1: {Others}, Word 2: {They} -> Yes
Example 3: Sentence: {I ponder that I should find the last scene before death, this last scene should be at the end of the memory road, finding it means finding the moment of my death.}, Word 1: {Memory road}, Word 2: {It} -> No
Example 4: Sentence: {Sometimes she passes by these gossiping girls, knowing that they are talking about how she was dumped by the sons of the leaders, she still sends them a nonchalant smile, their gossip is to her just a sparse rain that doesn't need an umbrella.}, Word 1: {Girls}, Word 2: {Them} -> Yes
Example 5: Sentence: {She sat on the plane, next to her was a doctor who had returned from studying in the United States, this man had just started his own business, ten years older than her, with a wife and children, during the two-hour flight, he passionately described to her his ambitious career prospects.}, Word 1: {Doctor}, Word 2: {He} -> Yes
Now The input is:
{input text} ->

Figure 4: Prompt for CWSC.

请将新闻分类到给定类别中，类别包括（news_entertainment、news_military、news_finance、news_tech、news_travel、news_culture、news_house、news_edu、news_agriculture、news_sports、...{其余类别}）。
示例1：句子：{出栏一头猪亏损300元，究竟谁能笑到最后！}-> news_finance
示例2：句子：{以前很火的巴铁为何现在只字不提？}-> news_tech
示例3：句子：{图解：全要素 多领域 高效益 天津智能科技军民融合发展}-> news_tech
示例4：句子：{美术生去北京画室参加集训会不会影响联考成绩？}-> news_edu
示例5：句子：{如何解读蚂蚁金服首季亏损？}-> news_tech
以下为输入：
{input text} ->

Please categorize the news into the given categories, which include (news_entertainment, news_military, news_finance, news_tech, news_travel, news_culture, news_house, news_edu, news_agriculture, news_sports, ... {other categories}).
Example 1: Sentence: {A loss of 300 yuan per pig sold, who can laugh last!} -> news_finance
Example 2: Sentence: {Why is the once popular 'Ba Iron' not mentioned anymore?} -> news_tech
Example 3: Sentence: {Illustration: Full elements, multi-field, high efficiency, Tianjin intelligent technology military-civilian integration development} -> news_tech
Example 4: Sentence: {Will art students attending training in Beijing studios affect their joint exam results?} -> news_edu
Example 5: Sentence: {How to interpret Ant Financial's first-quarter loss?} -> news_tech
Now The input is:
{input text} ->

Figure 5: Prompt for TNEWS.

请将句子分类到给定类别中，类别包括（社区超市、工具、社区服务、动作类、休闲益智、新闻、亲子儿童、绘画、直播、棋牌中心、求职、辅助工具、借贷、...{其余类别}）。
示例1：句子：{一款以海盗为题材的动作类游戏，采用QTE格斗的方式来进行战斗，你需要在合适的时机点击屏幕你的角色就会出现不同的攻击特效，除了QTE战斗之外还加入了角色培养，建筑培养等元素，还有更多丰富的角色卡牌等你来解锁收集。游戏特色1.探索世界，探索许多危险的地点进行你的冒险；2.追求飞翔的荷兰人，你有没有听说过传说中的鬼船，可从来不出口，并注定要永远驶向海洋吗；3.挑战其他的海盗，加入在线PVP对决} -> 动作类
示例2：句子：{朴朴快送超市创立于2016年，专注于打造移动端30分钟即时配送一站式购物平台，商品品类包含水果、蔬菜、肉禽蛋奶、海鲜水产、粮油调味、酒水饮料、休闲食品、日用品、外卖等。朴朴公司希望能以全新的商业模式，更高效快捷的仓储配送模式，致力于成为更快、更好、更多、更省的在线零售平台，带给消费者更好的消费体验，同时推动中国食品安全进程，成为一家让社会尊敬的互联网公司。,朴朴一下，又好又快,1.配送时间提示更加清晰友好2.保障用户隐私的一些优化3.其他提高使用体验的调整4.修复了一些已知bug} -> 社区超市
示例3：句子：{《怪物x联盟2》是由奥飞游戏重金打造的一款原创宠物收集养成类手游。研发团队方寸科技为国内知名游戏研发商，其前作《怪物X联盟》自发行以来成为国人原创人气手游，并且在东南亚，台湾等海外市场表现优异。} -> 经营养成
示例4：句子：{VoiceTube看电影学英语]是台湾最大的看电影学英语的社区，每天视频，提供了最高质量的英语学习内容。包括泰德会谈，TED爱德教育，CNN学生新闻，看漫画学英语，学英语，音乐，电影剪辑，视频游戏，学习超过1万部的英语学习视频的英语技能。在VoiceTube应用程序可以搜索你有兴趣学习的视频，我们还设计了很多Android的特异功能，让你学习更有效。应用特点视频 -> 视频
示例5：句子：{随时在线预览免税商品，关注最热商品，掌握优惠动态。方便您提前制定购物计划，从容应对SHOPPINGFIGHT。ENJOYYOURTRIPENJOYINSUNRISE,,为您优化了体验细节。} ->电商
以下为输入：
{input text} ->

Please categorize the sentences into the given categories, which include (Community Supermarket, Tools, Community Services, Action, Casual Puzzle, News, Parent-Child, Drawing, Live Streaming, Chess and Card Center, Job Seeking, Assistive Tools, Lending, ... {The rest of the categories}).
Example 1: Sentence: {A pirate-themed action game that uses QTE combat to fight, you need to click the screen at the right time, and your character will have different attack effects. In addition to QTE combat, it also includes character cultivation, building cultivation and other elements, as well as a wealth of character cards waiting for you to unlock and collect. Game features 1. Explore the world, explore many dangerous places for your adventure; 2. Pursue the Flying Dutchman, have you ever heard of the legendary ghost ship that never sails out and is doomed to sail the ocean forever; 3. Challenge other pirates, join online PVP duels} -> Action
Example 2: Sentence: {Pupu Quick Delivery Supermarket was established in 2016, focusing on creating a one-stop shopping platform for mobile terminals with 30-minute instant delivery. The product category includes fruits, vegetables, meat, poultry, eggs, milk, seafood, grain, seasoning, alcohol, beverages, leisure food, daily necessities, takeout, etc. Pupu Company hopes to become a faster, better, more, and more cost-effective online retail platform with a new business model and a more efficient and faster storage and distribution model, bringing consumers a better consumption experience, while promoting China's food safety process and becoming a respected internet company in society. Pupu, good and fast, 1. Delivery time prompts are clearer and friendlier 2. Some optimizations to protect user privacy 3. Other adjustments to improve the user experience 4. Fixed some known bugs} -> Community Supermarket
Example 3: Sentence: {"Monster X Alliance 2" is an original pet collection and cultivation mobile game created by Alpha Flight Games with a heavy investment. The development team, Square Technology, is a well-known domestic game developer, and its predecessor, "Monster X Alliance," has become a popular original mobile game since its release and has performed well in overseas markets such as Southeast Asia and Taiwan.} -> Business Cultivation
Example 4: Sentence: {VoiceTube Watch Movies and Learn English] is the largest community in Taiwan for learning English by watching movies, providing the highest quality English learning content every day. Including Ted Talks, TED Ed Education, CNN Student News, Comic English Learning, English Learning, Music, Movie Clips, Video Games, and learning over 10,000 English learning videos to improve your English skills. In the VoiceTube app, you can search for videos you are interested in learning, and we have designed many unique features for Android to make your learning more effective. Application features video -> Video
Example 5: Sentence: {Preview duty-free products online at any time, pay attention to the hottest products, and keep up with promotional dynamics. It's convenient for you to plan your shopping in advance and deal with SHOPPINGFIGHT calmly. ENJOYYOURTRIPENJOYINSUNRISE,, has optimized experience details for you.} -> E-commerce
Now The input is:
{input text} ->

Figure 6: Prompt for IFLYTEK.

请判断以下论文关键词是否全部为该摘要的关键词。
示例1：摘要：{为解决传统均匀FFT波束形成算法引起的3维声呐成像分辨率降低的问题,该文提出分区域FFT波束形成算法.远场条件下,以保证成像分辨率为约束条件,以划分数量最少为目标,采用遗传算法作为优化手段将成像区域划分为多个区域.在每个区域内选取一个波束方向,获得每一个接收阵元收到该方向回波时的解调输出,以此为原始数据在该区域内进行传统均匀FFT波束形成.对FFT计算过程进行优化,降低新算法的计算量,使其满足3维成像声呐实时性的要求.仿真与实验结果表明,采用分区域FFT波束形成算法的成像分辨率较传统均匀FFT波束形成算法有显著提高,且满足实时性要求.}，关键词：{水声学，FFT，波束形成，3维成像声呐} -> 是
示例2：摘要：{单核巨噬细胞表面表达的人类白细胞抗原DR（humanleucocyteantigen-DR，HLA-DR）是外源性抗原肽递呈过程中最重要的分子，其表达水平可在疾病早期反映外周免疫状态。HLA-DR由人类白细胞抗原基因Ⅱ类DR区域编码，主要表达于单核巨噬细胞、树突状细胞等抗原递呈细胞，在机体免疫系统中发挥着许多重要功能。近年来发现，急性脑卒中后外周单核细胞HLA-DR表达与机体免疫抑制密切相关，作者就HLA-DR在脑卒中后免疫抑制中的作用做一综述。}，关键词：{单核，脑卒中，递呈} -> 否
示例3：摘要：{以1-氨基乙内酰脲与芳香醛为原料,合成了6种1-氨基乙内酰脲芳香醛类席夫碱,收率为72．6%-89．2%,并对其反应条件进行了优化,得出在回流温度下,1-氨基乙内酰脲与芳香醛的投料摩尔比为1：1(对苯二甲醛为2：1)时反应1．5~2h产率最高。通过IR、1HNMR和元素分析表征了目标化合物的结构。}，关键词：{进行，合成，产率，2%} -> 否
示例4：摘要：{以1-氨基乙内酰脲与芳香醛为原料,合成了6种1-氨基乙内酰脲芳香醛类席夫碱,收率为72．6%-89．2%,并对其反应条件进行了优化,得出在回流温度下,1-氨基乙内酰脲与芳香醛的投料摩尔比为1：1(对苯二甲醛为2：1)时反应1．5~2h产率最高。通过IR、1HNMR和元素分析表征了目标化合物的结构。}，关键词：{席夫碱，合成，芳香醛，1-氨基乙内酰脲} -> 是
示例5：摘要：{通过研究Windows环境下USB设备的工作原理，应用操作系统与USB设备驱动通信获取设备描述和设备ID等信息的机制，提出了一种实用有效的USB设备监控技术。实现了在开机前后两种情况下对USB设备的实时监控，有效地避免了其他监控技术的漏洞。实验结果证明，该方法是可靠有效的。}，关键词：{设备描述，设备ID，Windows环境，安全监控} -> 是
以下为输入：
{input text} ->

Please determine whether the following paper keywords are all keywords for the abstract.
Example 1: Abstract: {To address the issue of reduced 3D sonar imaging resolution caused by traditional uniform FFT beamforming algorithms, this paper proposes a regional FFT beamforming algorithm. Under far-field conditions, with the constraint of ensuring imaging resolution and the goal of minimizing the number of divisions, genetic algorithms are used as an optimization means to divide the imaging area into multiple regions. In each region, a beam direction is selected to obtain the demodulation output of each receiving array element when receiving echoes from that direction, using this as the raw data for traditional uniform FFT beamforming in that region.}, Keywords: {Acoustics, FFT, Beamforming, 3D Imaging Sonar} -> Yes
Example 2: Abstract: {Human leukocyte antigen-DR (HLA-DR) expressed on the surface of monocytes is the most important molecule in the presentation of exogenous antigen peptides and can reflect the peripheral immune state at an early stage of disease. HLA-DR is encoded by the human leukocyte antigen class II DR region and is mainly expressed on antigen-presenting cells such as monocytes and dendritic cells, playing many important functions in the immune system. In recent years, it has been found that the expression of HLA-DR on peripheral monocytes after acute stroke is closely related to immune suppression in the body. The author reviews the role of HLA-DR in immune suppression after stroke.}, Keywords: {Monocytes, Stroke, Presentation} -> No
Example 3: Abstract: {Starting with 1-amino hydantoin and aromatic aldehydes as raw materials, six types of 1-amino hydantoin aromatic aldehyde Schiff bases were synthesized, with a yield of 72.6%-89.2%, and the reaction conditions were optimized to achieve the highest yield when the molar ratio of 1-amino hydantoin to aromatic aldehyde is 1:1 (2:1 for terephthalaldehyde) under reflux temperature for 1.5-2 hours. The structure of the target compounds was characterized by IR, 1H NMR, and elemental analysis.}, Keywords: {Proceed, Synthesis, Yield, 2%} -> No
Example 4: Abstract: {Starting with 1-amino hydantoin and aromatic aldehydes as raw materials, six types of 1-amino hydantoin aromatic aldehyde Schiff bases were synthesized, with a yield of 72.6%-89.2%, and the reaction conditions were optimized to achieve the highest yield when the molar ratio of 1-amino hydantoin to aromatic aldehyde is 1:1 (2:1 for terephthalaldehyde) under reflux temperature for 1.5-2 hours. The structure of the target compounds was characterized by IR, 1H NMR, and elemental analysis.}, Keywords: {Schiff Bases, Synthesis, Aromatic Aldehydes, 1-Amino Hydantoin} -> Yes
Example 5: Abstract: {By studying the working principle of USB devices under the Windows environment and applying the mechanism of communication between the operating system and USB device drivers to obtain information such as device description and device ID, a practical and effective USB device monitoring technology is proposed. Real-time monitoring of USB devices is achieved before and after booting, effectively avoiding the vulnerabilities of other monitoring technologies. Experimental results prove that the method is reliable and effective.}, Keywords: {Device Description, Device ID, Windows Environment, Security Monitoring} -> Yes
Now The input is:
{input text} ->

Figure 7: Prompt for CSL.

请判断两个句子是否表达同一件事情。
示例1：句子1：{双十一花呗提额在哪}，句子2：{里可以提花呗额度} -> 否
示例2：句子1：{花呗支持高铁票支付吗}，句子2：{为什么友付宝不支持花呗付款} -> 否
示例3：句子1：{赠品不能设置用花呗付款}，句子2：{怎么不能花呗分期付款} -> 否
示例4：句子1：{为什么这个订单不可以花呗支付}，句子2：{为什么支付时没有出现用花呗支付} -> 是
示例5：句子1：{花呗收款额度限制}，句子2：{收钱码，对花呗支付的金额有限制吗} -> 是
以下为输入：
{input text} ->

Please determine whether the two sentences express the same thing.
Example 1: Sentence 1: {Where is the Alipay credit limit increase for Double 11}, Sentence 2: {Where can I increase my Alipay credit limit} -> No
Example 2: Sentence 1: {Does Alipay support high-speed train ticket payment}, Sentence 2: {Why doesn't Youfubao support Alipay payment} -> No
Example 3: Sentence 1: {Free gifts cannot be set to pay with Alipay credit}, Sentence 2: {Why can't I pay in installments with Alipay credit} -> No
Example 4: Sentence 1: {Why can't this order be paid with Alipay credit}, Sentence 2: {Why didn't Alipay credit payment appear when paying} -> Yes
Example 5: Sentence 1: {Alipay credit collection limit}, Sentence 2: {Is there a limit on the amount of Alipay credit payment with the payment code} -> Yes
Now The input is:
{input text} ->

Figure 8: Prompt for AFQMC.

请分类两个句子的关系是蕴含、中立还是矛盾。
示例1：句子1：{身上裹一件工厂发的棉大衣,手插在袖筒里}，句子2：{身上至少一件衣服}-> 蕴含
示例2：句子1：{一些地方财政收支矛盾较大}，句子2：{地方经历了经济危机} -> 中立
示例3：句子1：{否则,我们的高要求得不到落实,也影响了我们的低目标的实现}，句子2：{我们的要求落实与否无所谓。} -> 矛盾
示例4：句子1：{因此,日本舆论曾盛传海部内阁是过渡性政权}，句子2：{此舆论在中国传播范围同样广泛} -> 中立
示例5：句子1：{阿嗲说,谢谢你们,生日还送了东西.}，句子2：{阿嗲收到生日礼物} -> 蕴含
以下为输入：
{input text} ->

Let's categorize the relationship between the two sentences as entailment, neutral, or contradiction.
Example 1: Sentence 1: {Wrapped in a cotton coat issued by the factory, hands in the sleeves}, Sentence 2: {At least one piece of clothing on the body} -> Entailment
Example 2: Sentence 1: {There is a significant contradiction in the financial revenue and expenditure of some places}, Sentence 2: {The place has experienced an economic crisis} -> Neutral
Example 3: Sentence 1: {Otherwise, our high demands will not be implemented, and it will also affect the realization of our low targets}, Sentence 2: {Whether our demands are implemented or not is indifferent} -> Contradiction
Example 4: Sentence 1: {Therefore, there was a strong rumor in Japan that the Kaifu administration was a transitional government}, Sentence 2: {This rumor is also widely spread in China} -> Neutral
Example 5: Sentence 1: {Ah Da said, thank you all, you still gave gifts for the birthday}, Sentence 2: {Ah Da received birthday gifts} -> Entailment
Now The input is:
{input text} ->

Figure 9: Prompt for OCNLI.