Role Prompting Guided Domain Adaptation with General Capability Preserve for Large Language Models

Rui Wang[♡▲], **Fei Mi**[♠*], **Yi Chen**[♡▲], **Boyang Xue**[♣◇],
Hongru Wang[♣◇], **Qi Zhu**[♠], **Kam-Fai Wong**[♣◇], **Ruifeng Xu**[♡♦▲*]
[♡]Harbin Insitute of Technology, Shenzhen, China
[♠]Huawei Noah's Ark Lab [♣]MoE Key Laboratory of High Confidence Software Technologies
[◇]The Chinese University of Hong Kong [♦]Peng Cheng Laboratory, China
[▲]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
ruiwangnlp@outlook.com, mifei2@huawei.com, xuruifeng@hit.edu.cn

Abstract

The growing interest in Large Language Models (LLMs) for specialized applications has revealed a significant challenge: when tailored to specific domains, LLMs tend to experience catastrophic forgetting, compromising their general capabilities and leading to a suboptimal user experience. Additionally, crafting a versatile model for multiple domains simultaneously often results in a decline in overall performance due to confusion between domains. In response to these issues, we present the **R**ol**E** Prompting **G**uided Multi-Domain **A**daptation (REGA) strategy. This novel approach effectively manages multi-domain LLM adaptation through three key components: **1) Self-Distillation** constructs and replays general-domain exemplars to alleviate catastrophic forgetting. **2) Role Prompting** assigns a central prompt to the general domain and a unique role prompt to each specific domain to minimize inter-domain confusion during training. **3) Role Integration** reuses and integrates a small portion of domain-specific data to the general-domain data, which are trained under the guidance of the central prompt. The central prompt is used for a streamlined inference process, removing the necessity to switch prompts for different domains. Empirical results demonstrate that REGA effectively alleviates catastrophic forgetting and inter-domain confusion. This leads to improved domain-specific performance compared to standard fine-tuned models, while still preserving robust general capabilities.

1 Introduction

Large Language Models (LLMs) (Touvron et al., 2023a,b; Brown et al., 2020; Ouyang et al., 2022) have revolutionized the field of Natural Language Processing, demonstrating exceptional general capabilities, such as instruction-following (Ouyang et al., 2022; Longpre et al., 2023) and complex reasoning (Wei et al., 2022; Shi et al., 2023; Wang et al., 2023b). However, general-purpose LLMs might fall short in some specific areas requiring professional knowledge, due to the lack of exposure to data in relevant domains. Hence, there has emerged an increasing number of studies in developing domain-specific models by injecting domain knowledge into LLMs in some domains, e.g., medicine (Wu et al., 2023a; Wang et al., 2023a), law (Cui et al., 2023), and finance (Wu et al., 2023c; Zhang et al., 2023).

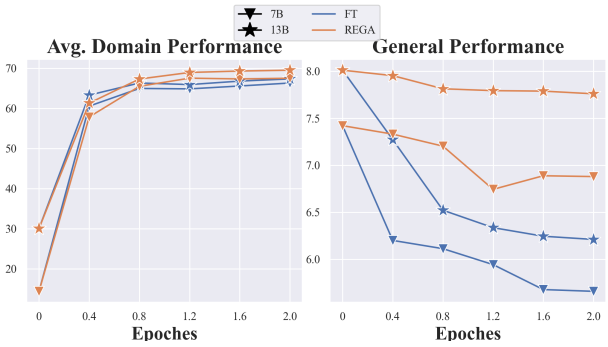

Figure 1: Performance Comparison of BELLE with varied sizes tuned by Standard Finetuning (FT) and REGA. The models tuned by FT suffer from a severe drop in general performance as the training epoch increases. Whereas, the counterparts tuned by REGA are better at preserving general capacities while achieving comparable domain-specific performance.

Nevertheless, adapting LLMs to specific areas risks triggering *catastrophic forgetting* (Fu et al., 2023; Arora et al., 2019; Lin et al., 2023; Zhang et al., 2023; Luo et al., 2023). As shown in Figure 1, the enhancement of specialized abilities comes at the cost of the generic ability to follow diverse instructions. This dilemma underscores the need for effective solutions that uphold the balance between domain-specific mastery and general applicability. Besides, directly adapting a single LLM to multiple domains simultaneously through standard finetuning may cause *inter-domain confusion* (Wang et al., 2023c), which negatively affects the model perfor-

*Corresponding Author.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2243–2255
June 16-21, 2024 ©2024 Association for Computational Linguistics

mance in each specific domain.

To this end, we propose the **R**ol**E** Prompt **G**uided Multi-Domain **A**daptation (**REGA**) strategy. As shown in Figure 2, given the instruction-following pairs from multiple target domains and our collected general-domain instructions, REGA reconstructs the training data for robust multi-domain adaptation through three key steps.

(1) Self-Distillation leverages the LLM itself to generate responses to the pre-collected diverse general-domain instructions before domain adaptation. The distilled instruction-following exemplars will be rehearsed during training to retain the generic abilities of the LLM, without the need to access the original, often private pre-training data.

(2) Role Prompting assigns the LLM with a unique expert role when adapting to distinct professional domains, and a generalist role by default when tackling general-domain data. This is done by concatenating a role prompt to the beginning of corresponding domain-specific or general-domain instructions. The role prompts act as system guidance to inform the LLM of clear domain boundaries during training, thus alleviating inter-domain confusion.

(3) Role Integration samples a small portion of data from each target domain and reuses them for training, all under the guidance of the central prompt. By guiding model training on the common domain-specific data, the different domain-sensitive expert roles are transferred and integrated into the generalist role of the central prompt.

During the inference stage, we directly use the central prompt to guide the model to handle instructions from various domains smoothly, alleviating the burden of role prompt engineering.

We conduct extensive experiments by adapting several LLMs in both Chinese and English datasets spanning three domains, including medicine, law, and finance. The experiment results exhibit that LLMs trained with REGA surpass other baselines in domain performance by a large margin while having a significant generic performance advantage. Furthermore, our detailed analysis underscores the effectiveness of each component of REGA. We reveal strong evidence that Self-Distillation is a reliable method for preventing the loss of general capabilities (§ 5.1). Additionally, Role Prompting is critical in reducing inter-domain confusion (§ 5.2). Lastly, Role Integration proves to be vital for the successful incorporation of knowledge from specific domain roles into a unified central role (§ 5.3), which is essential for the model's adaptability.

2 Related Work

Catastrophic Forgetting It has been observed that domain-specific tuning of LLMs can lead to catastrophic forgetting (Lin et al., 2023; Luo et al., 2023), where an LLM loses its ability to perform previously learned tasks effectively. This suggests a balance must be struck between domain specialization and general proficiency. To mitigate catastrophic forgetting, particularly in the context of continual learning, researchers have explored three kinds of strategies. *Exemplar replay* involves preserving and revisiting key training examples to maintain model performance (He et al., 2019; Lopez-Paz and Ranzato, 2017). *Regularization* methods introduce regulation functions in addition to the loss function to constrain the learning process (Lin et al., 2023; Li and Hoiem, 2018). *Architectural methods* adjust the model's structure by adding parameters specific to new tasks or domains (Zhu et al., 2022). Our task setting is to train an LLM that can competently handle multiple domains concurrently, with minimal impairment to its generalist capabilities, differentiating from continual learning where the model is exposed to tasks sequentially, striving to prevent significant forgetting of earlier tasks (Zhu et al., 2022).

Inter-domain Confusion Furthermore, training a single LLM for multiple domains risks triggering *inter-domain confusion* where the LLM may not perform as well in each domain due to the blending of domain-specific knowledge (Wang et al., 2023c; Sheng et al., 2021). Therefore, some studies have been directed toward identifying commonalities across domains to maintain model performance while preserving the unique characteristics of each domain (Wang et al., 2023c; Sheng et al., 2021). In this paper, we propose to utilize Role Prompting to alleviate inter-domain confusion.

Role Prompting Previous works found that role prompting can significantly improve the performance of LLMs. For example, Character.AI [1] proposes a dialogue agent mimicking diversified figures, which can bring enriched user experience. Similarly, Xu et al. (2022) proposes Cosplay to perform human-like conversations. Moreover, Wu et al. (2023b) found LLMs can effectively evaluate summarization results with diversified role prompts from varied perspectives. Kong et al. (2023) found

[1] https://beta.character.ai/

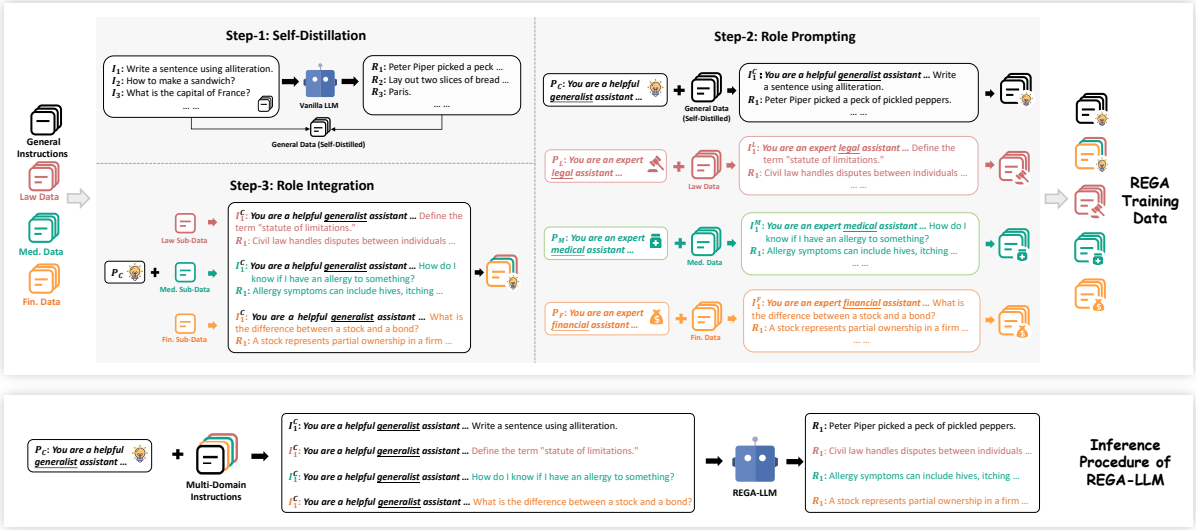

Figure 2: Overview of REGA. **For training**, REGA organizes the training data by: (1) *Self-Distillation*: The vanilla LLM generates exemplars according to a set of general-domain instructions to preserve generic abilities. (2) *Role Prompting*: The LLM is assigned a unique role through role prompts, which are concatenated with samples in corresponding domains. P_C is the central prompt indicating the generalist role for the general domain, while P_L, P_M, and P_F are the expert role prompts for law, medicine, and finance domains. (3) *Role Integration*: A fraction of data from each specialized domain is mixed with the general-domain data, all guided by the central prompt, which integrates various expert roles into the generalist role. **For inference**, the central prompt effectively guides the LLM tuned on REGA training data to respond to multi-domain instructions, without the need for role prompt selection.

role prompting can also boost the complex reasoning abilities of LLMs. Inspired by these findings, we propose to utilize role prompting to help LLMs distinguish samples among domains and assign domain-specific abilities to each role. Our experiments demonstrate that role prompting can effectively alleviate inter-domain confusion.

3 Method

3.1 Preliminaries

Consider that there is a large corpus whose domain distribution is known, which is $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\}$, where each $\mathcal{D}_i$ encompasses several sub-datasets about the i^{th} domain. D_i consists of instruction-response pairs, which means $(x_i, y_i) \in \mathcal{D}_i$. x_i and y_i represent the instruction and response respectively.

Our goal is to utilize $\mathcal{D}$ to train a language model θ to obtain θ' which has strong performance across n domains simultaneously without considerably compromising its general performance capability.

3.2 The REGA Tuning Strategy

As shown in Figure 2, REGA is a framework for organizing training datasets from multiple domains to obtain a final training corpus, which can improve the domain performance of LLMs without considerably compromising its general performance capability.

3.2.1 Self-Distillation

To alleviate catastrophic forgetting in the general domain, a straightforward and effective method is selecting exemplars in the training data and replaying (He et al., 2019; Lopez-Paz and Ranzato, 2017) them to LLMs besides the domain-specific data. However, the original training data of many LLMs are often proprietary and not open-sourced, so we try to partially replace it by devising the Self-Distillation. Specifically, we first collect a set of high-quality instructions $\mathcal{I} = \{(x_g,)\}$ from the general domain and let the LLM θ generate responses y_g for each x_g (as shown in the Step-1 part of Figure 2). This generated dataset $\mathcal{I} = \{(x_g, y_g)\}$, henceforth referred to as D_g, which is preserved as exemplars in the general domain and will be replayed in the following training process to restore the model's generic knowledge distribution. Now our training corpus can be denoted as $\mathcal{D}^+ = \{\mathcal{D}_g, \mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\}$

3.2.2 Role Prompting

Although the self-distilled D_g can alleviate the catastrophic forgetting, directly training θ on D^+ will degrade its performance on each one (Wang

et al., 2023c; Sheng et al., 2021) due to confusion among domains. To alleviate the inter-domain confusion, we introduce the Role Prompting to help LLMs distinguish among domains by assigning role prompts for data from each domain (as shown in the Step-2 part of Figure 2). In particular, the general domain is assigned a central prompt p_c, and each of n domains is assigned a unique role-prompt, forming a role-prompt set $P = \{p_c, p_1, p_2, ..., p_n\}$. Then each instruction-responses pair (x, y) is prefixed with its corresponding domain-specific role prompt, which means the current training dataset is $\mathcal{D}_r^+ = \{(p_c \oplus x_g, y_g)|(x_g, y_g) \in \mathcal{D}_g\} \bigcup \{(p_i \oplus x_i, y_i)|(x_i, y_i) \in \bigcup_{i=1}^{n} \mathcal{D}_i\}$.

3.2.3 Role Intergration

The Role Prompting can segregate domain-specific data during the training process but it also makes it crucial to determine which role prompt to use based on the domain of the input. To obviate the need for role prompt selection during inference, we design the Role Intergration that enables the central-prompt p_c to acquire the specialized abilities associated with each domain's role-prompt p_i. The key to this strategy is the reinforcement of the versatility of the central prompt, allowing the LLMs to process prompts from all domains using p_c. Concretely, a fraction of data is randomly selected from each domain's dataset D_i, denoted as D_i', is combined with the general domain data D_g and prefixed with the central prompt p_c. The composite data collection is thus structured as $\mathcal{T}_r^s = \{(p_c \oplus x_g, y_g)|(x_g, y_g) \in D_g \cup (\bigcup_{i=1}^{n} D_i'), D_i' \subset D_i\} \bigcup \{(p_i \oplus x_i, y_i)|(x_i, y_i) \in \bigcup_{i=1}^{n} D_i\}$.

3.2.4 Training Corpus of REGA

The final training corpus that REGA builds upon $\mathcal{D}^+$ is $\mathcal{T}_r^s$. Having trained the LLM θ on $\mathcal{T}_r^s$, we obtain the θ'. Besides, we introduce the mixing ratio r, quantifying the ratio of each selected subset D_i' to its full domain dataset D_i. The mixing ratio is defined as $r = |D_i'|/|D_i|$. This metric facilitates the calibration of domain exposure during the training process.

3.3 The REGA Inference Procedure

At the inference stage, we only need the central prompt to guide LLMs in the generation. For the given input x_u, the prediction process is represented as $y_u = \theta'(p_c \oplus x_u)$. This process bypasses the need for selecting different role prompts for each domain, thereby streamlining model deployment and ensuring consistency in responses across varied domains.

4 Experiment

4.1 Datasets

In this section, we introduce the domain datasets we utilized and the instruction set for **Self-Distillation**. We perform the experiments on three domains, medicine, law, and finance. We choose datasets carefully to contain both language understanding and generation tasks for more comprehensive evaluation of LLMs. The statistics and detailed metrics of datasets are shown in Appendix A.

English Datasets We encompass four English datasets across the medical, legal, and financial domains, including PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022), casehold_QA (Zheng et al., 2021), and FinBertQA[2].

Chinese Datasets For the Chinese portion of our study, we have sourced 11 datasets from three different sectors. For medical, we include cMedQQ (QQ) for paraphrase identification, cMedTC (TC) for sentence classification (Zhang et al., 2022) and cMedQA (MQA) for question answering (Zhang et al., 2018); in the legal domain, we have LawQA (LQA) for question answering[3] and LawSum (LS) for document summarization[4]; and for finance, datasets such as FNA, FQA, FNL, FRE, FFE, and FSP, which cover a range of tasks from sentiment analysis to entity relation classification, are adopted from Lu et al. (2023).

General Instruction Datasets In light of existing research underscoring the importance of data quality and diversity while training LLMs (Gunasekar et al., 2023; Wang et al., 2023d), we try to build high-quality and diversified instruction datasets to better preserve the models' generic capabilities in the constructed D_g after Self-Distillation. For the Chinese models, we have randomly extracted 50K instruction samples from both the Chinese-Alpaca[5] and MOSS (Sun et al., 2023) projects, resulting in a combined total of 100K samples. In the case of the English models, we have likewise randomly chosen 50K instruction samples from each of the WizardLM (Xu et al.,

[2]https://sites.google.com/view/fiqa
[3]https://github.com/pengxiao-song/LaWGPT/tree/main
[4]http://cail.cipsc.org.cn
[5]https://github.com/ymcui/Chinese-LLaMA-Alpaca

Model	General	Medicine			Law		Finance					
	CGev	QQ	TC	MQA	LQA	LS	FNA	FQA	FNL	FRE	FFE	FSP
					BELLE-7B							
0-shot	7.42	8.80	1.60	12.32	16.88	9.85	49.54	24.98	41.73	0.00	2.99	18.13
FT	5.41	83.10	75.94	36.84	**58.05**	44.82	61.07	72.83	93.47	50.75	67.39	85.41
FTSD	6.26	**85.21**	74.44	32.64	57.04	44.29	59.81	75.10	91.31	47.76	**69.57**	85.44
REGA^c	**6.87**	82.39	**76.69**	**37.95**	57.65	**46.90**	**61.68**	**78.18**	**95.65**	**55.22**	68.48	**85.68**
					BELLE-13B							
0-shot	8.01	33.10	3.01	41.72	56.47	35.15	41.53	23.85	41.30	7.46	33.70	13.20
FT	6.21	84.51	**81.96**	35.63	58.25	45.51	61.87	77.30	91.30	61.20	68.11	86.08
FTSD	6.92	84.37	78.95	36.47	**58.43**	44.58	61.67	74.69	**93.48**	59.29	72.83	85.76
REGA^c	**7.75**	**85.33**	79.22	**37.67**	58.11	**47.52**	**62.27**	**77.79**	92.33	**62.37**	**73.71**	**88.06**
Metrics	-	Acc.	u.F1	u.F1	u.F1	u.F1	u.F1	u.F1	Acc.	Acc.	Acc.	u.F1

Table 1: We present the performance of BELLE in different experimental conditions, with the top scores highlighted in **bold**. The superscript ^{c} indicates that the model's assessment was conducted using the central prompt p_c. Acc. or u.F1 means that the evaluation metric of this dataset is Accuracy or Uni-gram-F1 respectively. The mixing ratio of REGA is 0.1.

Model	General	Medicine		Law	Finance
	MTB	PMQA	MMQA	CQA	FQA
		Vicuna2-7B			
0-shot	6.23	42.68	31.28	18.60	24.02
FT	4.57	52.17	42.07	66.80	32.17
FTSD	5.68	60.87	**42.27**	67.20	39.12
REGA^c	**6.11**	**65.21**	41.41	**68.80**	**45.24**
Metrics	-	Acc.	Acc.	Acc.	u.F1

Table 2: The performance of Vicuna-7B is detailed below, with the highest scores emphasized in **bold**. Acc. or u.F1 means that the evaluation metric of this dataset is Accuracy or Uni-gram-F1 respectively. The mixing ratio of REGA is 0.1.

2023) and Alpaca[6] projects, amounting to a total of 100K samples.

Then these instructions are fed into the `BELLE` and `Vicuna` to obtain distilled exemplar set D_g. In the decoding process, we set the temperature to 0.7 and top-p to 0.95 for response generation.

4.2 Role Prompt Setting

We design role prompts for medicine, law, and finance domains respectively. However, we use the central prompt p_c in line with the original instruction-tuning process of the model rather than a fresh one. For instance, take the prompt used during Vicuna's instruction-tuning: "*A chat between a curious user and an artificial intelligence assistant. The assistant is designed to be helpful, detailed, and polite in responding to user queries.*" This same prompt is employed as the central prompt p_c in REGA to create our training dataset for Vicuna. Our goal of using the same p_c as the one in the instruction-following process is to preserve the foundational knowledge the model originally had.

4.3 Baselines

Zero-Shot We evaluate the `BELLE` and `Vicuna` on the domain and general test set with greedy decoding in a zero-shot setting.

Standard Finetuning (FT) We finetune the LLM θ on domain-specific datasets spanning three distinct domains, respectively represented by D_m, D_l, and D_f and the training corpus denoted as $T_{ft} = \{D_m \cup D_l \cup D_f\}$. This finetuning process results in a refined model θ_{ft}. For a given user input x_u, the inference stage of θ_{ft} is expressed as $y_u = \theta_{ft}(x_u)$.

Standard Finetuning with Self-Distillation (FTSD) We combine FT and Self-Distillation to diagnose catastrophic forgetting while training with FT and explore the effects of Self-Distillation. The FTSD approach integrates the self-distilled instruction-response dataset $\mathcal{D}_g$ into the fine-tuning corpus $\mathcal{D}$, resulting in the $\mathcal{T}_{ftsd} = \{\mathcal{D}_g \cup \mathcal{D}_m \cup \mathcal{D}_l \cup \mathcal{D}_f\}$. After training θ on T_{ftsd}, we obtain θ_{ftsd}. For a given user input x_u, the inference stage of θ_{ftsd} is denoted as $y_u = \theta_{ftsd}(x_u)$.

Standard Finetuning with Role Prompting (FTRP) We combine FT and Role Prompting to explore the existence of inter-domain confusion and the effects of Role Prompting. We use the same training corpus T_{ft} as FT in this setting but assign the role prompts to the instructions of each domain. Assume we have role prompts

[6]https://github.com/tatsu-lab/stanford_alpaca

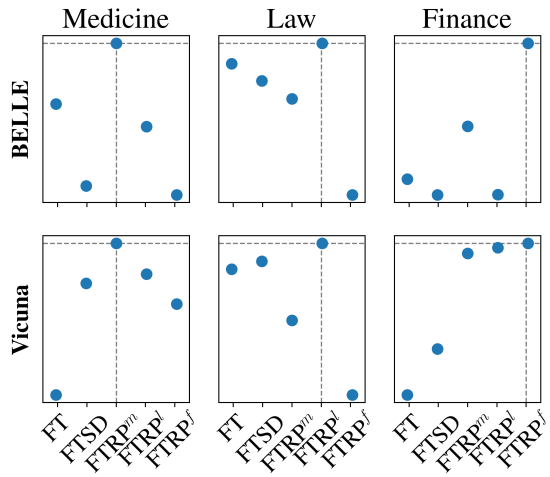

Figure 3: Performance of BELLE and Vicuna tuned by FTRP. FTRP^x indicate models are tested by the role prompts p_x of the x domain.

p_m, p_l and p_f for medicine, law, and finance domains, the training corpus can be denoted as $T_{ftrp} = \{(p_i \oplus x_i, y_i) | (x_i, y_i) \in \bigcup_{i \in \{m,f,l\}} D_i\}$. For a given user input x_u, we need to choose different role prompts according to the domain of x_u, while inferring with the obtained model θ_{ftap}. For example, if the x_u is from the medical role prompt, the inference procedure is $y_u = \theta_{ftrp}(p_m \oplus x_u)$.

REGA The REGA training and inference procedure are already described in Section 3.2.

For the existence of role prompts of REGA and FTRP, we also denote their inference process as REGA^x and FTRP^x, which means that the model trained with REGA or FTRP are using the role prompt p_x of the domain x.

4.4 Models

For our experiments with Chinese datasets, we have selected models from the BELLE series [7], specifically `BELLE-7B-2M` and `BELLE-13B-2M`. These models are iterations of `LLaMA-7B` and `LLaMA-13B` respectively (Touvron et al., 2023a). They have been further fine-tuned in a supervised manner on a Chinese dataset containing 2 million instruction-response pairs. Regarding English datasets, our choice of the base model is `Vicuna-1.5-7B` [8], which has been fine-tuned from `LLaMA2-7B` (Touvron et al., 2023b). We train these models in the LoRA (Hu et al., 2021) manner. The r and α of LoRA are 16 and 32 respectively. For all of the methods, batch size is set to 16, and the maximum number of epochs is set to 2. We test performance on the checkpoint obtained after the second epoch.

[7] https://github.com/LianjiaTech/BELLE
[8] https://github.com/lm-sys/FastChat

4.5 Evaluation

For domain performance, we evaluate the models on the corresponding test datasets using automatic metrics, including accuracy and uni-gram-F1 (also illustrated in Table 1 and Table 2).

As for the general performance, we evaluate the English models on MT-Bench (MTB) (Zheng et al., 2023) and the prompt format follows the exact setting of MT-Bench. Each response is evaluated by a numerical score ranging from 0 to 10. For the Chinese models, we collect an evaluation collection, **CGev**, consisting of 650 samples, to test model abilities across coding, reasoning, question answering, classification, and conversation tasks. The distribution of the tasks in **CGev** and the prompt format are shown in Appendix A. We evaluate `BELLE` series models on **CGev** by asking GPT-4 to give a numerical score (from 0 to 10) for the single response. All of the model's general performances are automatically evaluated by GPT-4-0613 with greedy decoding to reduce randomness.

4.6 Performance Analysis

To clearly illustrate the experiment results, we present the results of Zero-shot (0-shot), FT, FTSD, and REGA in Table 1 and Table 2. The performance of FTRP, FT and FTSD are in Figure 3. Several interesting observations can be noted.

Diagnosing Catastrophic Forgetting While FT can consistently improve domain performance, it tends to compromise the model's overall proficiency. As shown in Table 1 and Table 2, the general performance of these models decreases across languages and model sizes. In particular, the `BELLE-7B` model sees its score decrease from 7.42 to 5.41. Similarly, the `BELLE-13B` model's score declines from 8.01 to 6.21 after FT.

Diagnosing Inter-domain Confusion To investigate the existence of inter-domain confusion, we fine-tune `BELLE` using only medical datasets. The outcomes, depicted in Table 3, show that `BELLE` fine-tuned with FT solely on medical data, outperforms the variant trained across three domains in Table 1. This contrast confirms the presence of inter-domain confusion.

REGA Benefits Both General and Domain Performance. However, LLMs fine-tuned using the

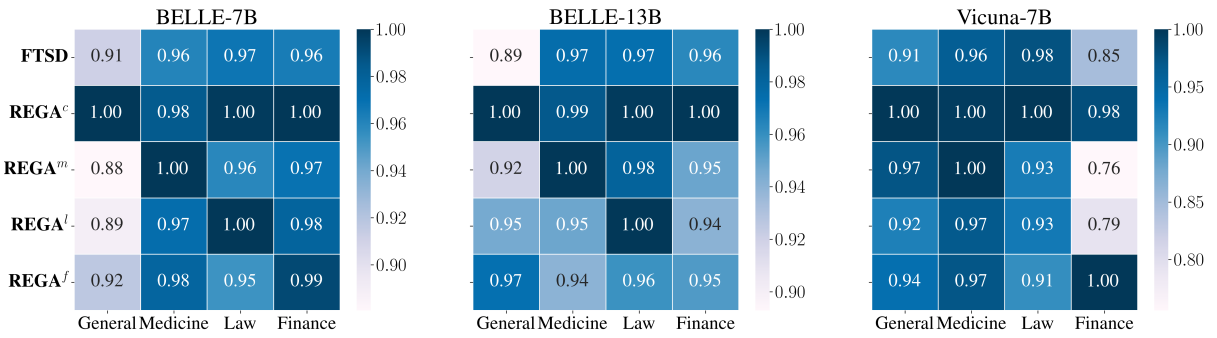

Figure 4: We present the normalized performance metrics for the BELLE and Vicuna-7B, which are fine-tuned using the **FTSD** and **REGA**. The notation **REGA**^{x} indicates that the model's inference is performed using the role prompt p_x. The normalization process involves dividing each score by the maximum score within the same column. The mixing ratio of REGA is 0.1.

REGA strategy exhibit superior domain-specific performance compared with baselines while maintaining a higher level of general abilities. To explore why the model trained REGA is better, we conduct further analysis in the following sections.

5 Further Analysis

In this section, we analyze the effects of the three components of REGA.

5.1 Effects of Self-Distillation

Self-Distillation effectively alleviates the catastrophic forgetting of generic abilities. As depicted in Table 1 and Table 2, the models trained with strategies with the Self-Distillation component (i.e., FTSD, REGA) achieve higher general scores than those trained with FT. For example, Vicuna achieves a score of 5.68 on the MT-Bench, which notably exceeds the 4.57 of the same model using the FT. This can prove that blending the training corpus with a self-distilled instruction dataset can alleviate the tendency of LLMs to forget their generic capabilities during the training process.

Furthermore, we also observed a disparity in domain-specific effectiveness when employing the FTSD to BELLE and Vicuna. As illustrated in Table 1, the domain-specific performance of both BELLE-7B and BELLE-13B, when trained using the FTSD strategy, is inferior to that of models trained under the FT approach. Conversely, the performance of Vicuna surpasses that of the FT configuration. We attribute this phenomenon to the English domain data excessively impairing general performance Vicuna, more than the Chinese domain data to BELLE. This is reflected in poor outcomes in FQA where the unigram-F1 score is low. Besides the limited diversity in English datasets (only 4 compared to 11 Chinese datasets) and the frequent requirement for shorter text responses might also be contributing factors to this issue.

5.2 Effects of Role Prompting

Figure 4 displays the performance of models trained using REGA with different role prompts and compares it to the FTSD method. Figure 3 shows the performance of models trained with FTRP in response to different role prompts, with comparisons to both FT and FTSD methods. These two figures allow us to conclude the following interesting observations:

(1) **Role Prompts alleviates inter-domain confusion.** The FTRP strategy outperforms those models trained with FT and FTSD across all tested domains. This superior performance is directly linked to the implementation of domain-specific role prompts throughout training and inference periods. This proves that Role prompts are crucial for assisting LLMs in recognizing and processing instructions tailored to specific domains by explicitly differentiating them.

(2) **Role Prompts elicit abilities within target domains.** LLMs, such as BELLE, trained with REGA or FTRP exhibit higher performance in the medical domain than the other two domains when utilizing the medical role prompt p_m, as shown in Figure 4 and Figure 3. This is also the same situation for the other two domain-specific role prompts.

5.3 Effects of Role Integration

The above section proves that the Role Prompting can effectively alleviate inter-domain confusion, leading to clear task distinction for the models. Concurrently, Role Integration simplifies the in-

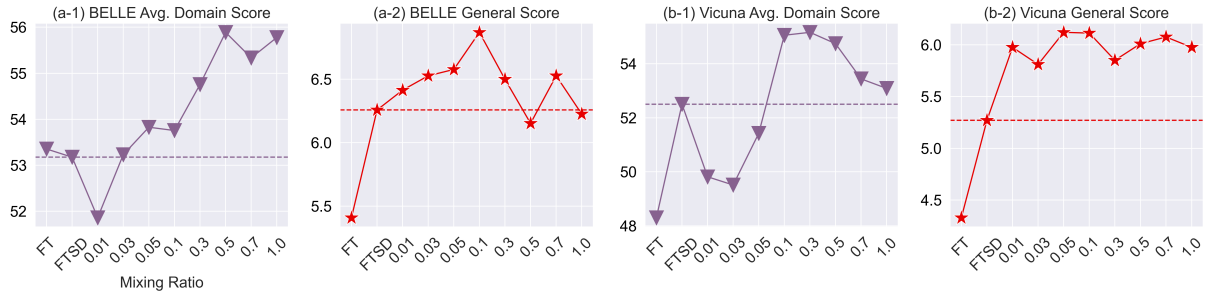

Figure 5: General and domain performance of `BELLE-7B` and `Vicuna-7B` trained with a varied mixing ratio of REGA.

Model	General	Medicine		
	CGev	QQ	TC	MQA
0-shot	7.42	8.80	1.60	12.32
FT	5.52	84.62	**83.60**	**39.60**
FTSD	6.55	82.60	79.20	36.65
REGA^c	**6.82**	**86.30**	83.14	37.38
REGA with Different Role Prompts				
REGA^m	6.33	84.50	81.20	39.17

Table 3: Performance of Belle on medicine and the best scores are in **bold**. The mixing ratio of REGA is 0.1.

ference process by removing the need for prompt selection and still ensures high performance across various areas. This is evident in Figure 4, where top performance typically corresponds with the matrix's diagonal and the REGA^c row.

Moreover, we have the following observations from Figure 4: (1) Adding only 10% of domain data to the central prompt is sufficient for REGA^c to exceed the domain performance of other role prompts, which use the full domain data set. (2) Even with access to the entire general and domain datasets, FTSD lags behind REGA^c across all domains. This indicates that the REGA model's domain proficiency isn't just a product of shared domain data; there's a clear contribution of knowledge transfer from domain-specific prompts.

Taken together, we argue that the knowledge transfer exists in the REGA-tuned model, which flows from the domain role prompts to the central prompt with the help of the shared domain data.

6 Discussion

6.1 REGA on Single Domain

We further extend our training to include the `BELLE-7B` model only within the medical domain, employing the strategies outlined in Section 4.3. The outcomes of these experiments are detailed in Table 3. Analysis of the data presented in Table 3 leads us to two key insights: firstly, the inter-domain confusion that can hamper performance is mitigated when focusing on a single domain, as evidenced by the **FT** approach yielding better results within the medical domain compared to training across multiple domains in Table 1. Secondly, the REGA strategy continues to demonstrate its efficacy by both reducing the loss of general language capabilities and enhancing the model's performance in the domain-specific context. This indicates that REGA still brings significant performance gains even when there is only a single-domain training requirement.

6.2 Choice of Mixing Ratio

Then we explore the impact of the mixing ratio r in Role Integration (shown in Figure 5). We have two observations: (1) Although a low mixing ratio (such as 0.01) is not enough for Role Integration to make REGA excel FT and FTSD in domain test sets, its generic abilities still stay at a superior position compared to other two methods. (2) The performance of the model train with REGA fluctuates with the change of mixing ratio, however, it still surpasses FT and FTSD by a large margin. As for the choice of mixing ratio, we recommend a safe interval $[0.05, 0.3]$ to simultaneously achieve higher domain and general performance.

7 Conclusion

In this paper, we attempt to strike a balance between domain specialization and generic abilities while adapting LLMs to multiple domains. Specifically, we propose the REGA, which consists of Self-Distillation to alleviate the catastrophic forgetting, Role Prompting to separate each domain while assigning each role prompt with domain-specific abilities to avoid inter-domain confusion, and Role Integration to transfer the domain-specific abilities

from the domain-specific role prompt to the central prompt. Extensive experiments on plenty of datasets and LLMs demonstrate the effectiveness and efficiency of our proposed method.

Limitations

In this paper, we introduce the **REGA** method for studying how to enhance LLMs with capabilities across multiple domains. However, **REGA** relies on pre-existing high-quality instruction sets to build general-domain exemplars. The quality of the instruction set determines the retention of the model's general capabilities. In this paper, we have made an effort to use open-source, high-quality data, as cited in the previous section.

Ethics Statement

In this paper, the datasets and models used are open-source and do not involve any issues related to privacy or contain harmful information. The approach proposed aims to enhance the domain capabilities of LLMs, focusing on improving their response accuracy and consistency. Additionally, all open-source resources employed in this research are cited or their sources explicitly stated. Accordingly, the models we have developed, which are built upon these open-source resources, do not present ethical concerns.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62176076, the Natural Science Foundation of Guangdong under Grant 2023A1515012922, Shenzhen Foundational Research Funding under Grant JCYJ20220818102415032, The Major Key Project of PCL under Project PCL2023A09, and Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies under Grant 2022B1212010005k. And we would like to express our gratitude to all the reviewers and editors for their valuable suggestions and feedbacks that have significantly improved our work.

References

Gaurav Arora, Afshin Rahimi, and Timothy Baldwin. 2019. Does an LSTM forget more than a CNN? an empirical study of catastrophic forgetting in NLP. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 77–86. Australasian Language Technology Association.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10421–10430. PMLR. ISSN: 2640-3498.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need.

Tianxing He, Jun Liu, Kyunghyun Cho, Myle Ott, Bing Liu, James Glass, and Fuchun Peng. 2019. Mix-review: Alleviate forgetting in the pretrain-finetune framework for neural language generation models.

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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577. Association for Computational Linguistics.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting.

Zhizhong Li and Derek Hoiem. 2018. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947.

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, and Tong Zhang. 2023. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning.

David Lopez-Paz and Marc' Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning.

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.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Xiang-Rong Sheng, Liqin Zhao, Guorui Zhou, Xinyao Ding, Binding Dai, Qiang Luo, Siran Yang, Jingshan Lv, Chi Zhang, Hongbo Deng, and Xiaoqiang Zhu. 2021. One model to serve all: Star topology adaptive recommender for multi-domain CTR prediction.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, and Xipeng Qiu. 2023. Moss: Training conversational language models from synthetic data.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023a. HuaTuo: Tuning LLaMA model with chinese medical knowledge.

Hongru Wang, Huimin Wang, Lingzhi Wang, Minda Hu, Rui Wang, Boyang Xue, Hongyuan Lu, Fei Mi, and Kam-Fai Wong. 2023b. Tpe: Towards better compositional reasoning over conceptual tools with multi-persona collaboration. *ArXiv*, abs/2309.16090.

Ximei Wang, Junwei Pan, Xingzhuo Guo, Dapeng Liu, and Jie Jiang. 2023c. Decoupled training: Return of frustratingly easy multi-domain learning.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023d. Self-instruct: Aligning language model with self generated instructions.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. Version: 1.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. PMC-LLaMA: Towards building open-source language models for medicine.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023b. Large language models are diverse role-players for summarization evaluation.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023c. BloombergGPT: A large language model for finance.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. WizardLM: Empowering large language models to follow complex instructions.

Chen Xu, Piji Li, Wei Wang, Haoran Yang, Siyun Wang, and Chuangbai Xiao. 2022. Cosplay: Concept set guided personalized dialogue generation across both party personas. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 201–211.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

S. Zhang, X. Zhang, H. Wang, L. Guo, and S. Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2023. XuanYuan 2.0: A large chinese financial chat model with hundreds of billions parameters.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset.

Qi Zhu, Bing Li, Fei Mi, Xiaoyan Zhu, and Minlie Huang. 2022. Continual prompt tuning for dialog state tracking.

Medicine: You are a knowledgeable assistant in the domain of healthcare and medicine, providing detailed answers to medical questions and successfully completing tasks in the medical domain.

Law: You are a knowledgeable assistant in the domain of law, and you provide detailed answers to users' legal inquiries and other legal requests. You excel at completing tasks in the legal domain.

Finance: You are a knowledgeable assistant in the domain of finance, capable of providing detailed answers to users' financial questions and completing tasks in the finance domain very well.

Table 4: Role prompts used in **REGA**. The Chinese version is translated from the above English prompts. The central prompt follows the original setting of LLMs.

A Dataset Statistics

In this section, we illustrate the datasets we utilized in Table 6 and Table 7, including 5 English datasets and 11 Chinese datasets in three domains, Medicine, Law, and Finance. Moreover, the evaluation metrics are also presented in the table.

B Prompt Settings

In this section, we introduce the setting of role prompts in Table 4 and the setting of the evaluation prompt in Table 5 of using GPT-4 to judge the Chinese LLMs' general performance.

C Genenal Performance Evaluation

The CGev dataset encompasses a range of tasks. The task and instance distribution are as follows: *Coding*: 19; *Information Extraction*: 30; *Classification*: 31; *Creative Writing*: 56; *Recommendation*: 50; *Dialogues*: 52; *Knowledge Testing*: 110; *Context-based Question Answering*: 23; *Open-domain Question Answering*: 23; *Rejection*: 25; *Summarization*: 51; *Math*: 55; *Planning*: 17; *Language Reasoning*: 60; *Writing*: 47.

```
请评价AI助手对用户问题的回复质量。
- - -
问题：{}

- - -
助手：{}

请分析助手的回复（综合考虑安全性、通顺性、相关性、正确性、信息性、专业性等）。
然后判断每个助手的回复是否存在以下错误：
无意义的重复
语句截断
不当的多语混用
语言不规范
回复与问题不相关
事实错误
违反逻辑规则
未遵循指令或约束
最后给每个助手的回复评分，最高10分，最低0分。

请按照以下JSON格式回答，对于错误判断，1代表存在相应错误，0代表不存在：
"分析": "...", "助手": "无意义的重复": ?, "语句截断": ?, ... "评分": ?
```

Table 5: Prompts we used to prompt GPT-4 to evaluate the general performance of LLMs. We request GPT-4 to give a numerical score ranging from 0 to 10.

	Medical		Law	Finance
	PMQA	MMQA	CHQA	FBQA
Training				
Nums.	1,000	10,000	10,000	10,000
P. Length	253.3	10.38	2058.5	62.6
R. Length	43.2	55.95	1.0	1034.6
Testing				
Nums.	50	500	500	500
P. Length	256.7	10.25	1925.7	63.0
R. Length	41.1	48.65	1.0	1034.5
Metrics	Acc.	Acc.	Acc.	uF1

Table 6: Statistics of 5 English datasets. **P. Length** and **R. Length** represents the average length of prompts and responses respectively. **Acc.** means accuracy and the **uF1** indicates the uni-gram-F1 score.

	Medical			Law		Finance					
	QQ	TC	MQA	LQA	LS	FNA	FQA	FNL	FRE	FFE	FSP
Training											
Nums	14,500	14,110	28,914	4,372	5,235	5,000	5,000	5,000	5,000	5,000	4,000
P. Length	83.9	709.3	31.4	67.3	1722.7	215.2	304.4	196.4	282.5	62.8	282.8
R. Length	1.0	12.8	119.4	136.0	247.1	25.4	6.3	5.2	3.5	2.0	7.5
Testing											
Nums	500	500	1,000	500	500	3,600	2,469	884	1,489	2,020	500
P. Length	83.6	708.6	31.5	67.0	1691.5	197.9	301.2	189.3	283.5	62.8	300.0
R. Length	1.0	12.8	122.1	137.6	250.7	26.0	6.3	5.1	3.5	2.0	6.7
Metrics	Acc.	uF1	uF1	uF1	uF1	uF1	uF1	Acc.	Acc.	Acc.	uF1

Table 7: Statistics of 11 Chinese datasets. **P. Length** and **R. Length** represents the average length of prompts and responses respectively. **Acc.** means accuracy and the **uF1** indicates the uni-gram-F1 score.