Transfer-Prompting: Enhancing Cross-Task Adaptation in Large Language Models via Dual-Stage Prompts Optimization

Anonymous ACL submission

Abstract

Large Language Models (LLMs) face significant challenges in real-world applications that require simultaneously achieving high-quality responses and adhering to specific instructions. To address these issues, we introduce **Transfer-Prompting**, a novel two-stage framework designed to improve cross-task adaptation in prompt generation. The framework comprises two main components: (1) **source prompt construction**, which refines prompts on source task datasets to enhance their generalization capability, and (2) **target prompt generation**, which fine-tunes high-performing source prompts on task-specific datasets to optimize cross-task performance. In each optimization cycle, a reference LLM generates candidate prompts based on historical prompt-score pairs and task descriptions in the reference prompt. These candidate prompts are iteratively refined, with a scorer LLM evaluating their effectiveness using an objective prompt evaluator. This feedback loop facilitates continuous refinement, improving prompt quality and task-specific performance. We validate Transfer-Prompting through extensive experiments involving 25 LLMs, including 7 foundational and 18 specialized models, across 9 diverse datasets. The results demonstrate that Transfer-Prompting significantly enhances task-specific performance, highlighting its potential to improve cross-task adaptation in LLMs.

1 Introduction

Large Language Models (LLMs) have made significant advances in natural language processing, enabling high-quality text generation across various applications, including conversational agents, content creation, and machine translation (Wei et al., 2022). However, deploying LLMs in real-world scenarios presents unique challenges, particularly in balancing the generation of high-quality outputs with the ability to follow instructions effectively across diverse and complex tasks (Wang et al., 2023a; Chang et al., 2024).

These challenges are especially pronounced in tasks with multiple subtasks or stringent constraints, where LLMs often generate hallucinated outputs—responses that are syntactically coherent but factually incorrect or irrelevant (Ji et al., 2023; Bang et al., 2023). Moreover, LLMs may misinterpret user queries, leading to responses that fail to meet expectations or address the core of the question (Kulkarni and Tupsakhare, 2024). Such limitations undermine the utility of LLMs and expose them to significant risks in sensitive domains like healthcare, legal, and finance, where inaccurate or off-topic outputs can have serious consequences (Nori et al., 2023).

One potential solution to mitigate these challenges is the use of LLM-based automatic prompt optimization (Zhou et al., 2023; Pryzant et al., 2023). These methods typically involve iteratively optimizing prompts using an LLM to improve model performance on specific tasks. However, current optimization techniques predominantly focus on single-stage optimization aimed at enhancing a single evaluation metric (Yang et al., 2024; Sun et al., 2023). While effective in certain contexts, these methods often fail to account for the complexities of multi-objective tasks or tasks that require balancing multiple, sometimes conflicting, evaluation criteria. For instance, tasks that require balancing the tradeoff between maximizing output quality and maintaining high instruction-following accuracy remain particularly challenging for existing models. Furthermore, many current methods neglect the need for comprehensive evaluation across multiple performance dimensions, limiting insights into the model's overall effectiveness (Chen et al., 2024).

To address these limitations, we propose **Transfer-Prompting**, a novel two-stage framework designed to optimize prompts for LLMs in

complex tasks. This framework consists of two core components: (1) **source prompt construction**, which refines the original prompts on source task datasets to generate source prompts with enhanced generalization capability, and (2) **target prompt generation**, which improves the cross-task adaptation of target prompts by fine-tuning a set of high-performing source prompts on task-specific datasets.

In each optimization cycle, a reference LLM generates candidate prompts based on historical prompt-score pairs and task descriptions embedded in the reference prompt. The optimization terminates when the reference LLM fails to generate a higher-scoring prompt or when a predefined optimization step limit is reached. The scorer LLM evaluates the effectiveness of the candidate prompts using an objective prompt evaluator.

We validate the Transfer-Prompting framework through extensive experiments conducted on 25 LLMs, including 7 foundational models (e.g., GPT-3.5-Turbo (OpenAI, 2023a), GPT-4 (OpenAI, 2023b)) and 18 specialized models from the medical, legal, and financial sectors. The evaluation involves 3 heterogeneous reasoning datasets and 6 multi-task datasets tailored to these specialized models. The results demonstrate that Transfer-Prompting significantly enhances task-specific performance and cross-task adaptation, improving both instruction-following accuracy and overall output quality across diverse tasks.

Our main contributions are as follows:

- We propose **Transfer-Prompting**, a novel LLM-based automatic prompt optimization framework, which consists of two core stages: source prompt construction and target prompt generation.
- The optimization process relies on four key components. The reference LLM generates candidate prompts based on the requirements of the reference prompt, while the scorer LLM evaluates and provides feedback using an objective prompt evaluator.
- Extensive experiments on 25 LLMs (including both foundational and specialized models) show that Transfer-Prompting significantly improves task-specific performance, highlighting its potential to enhance cross-task adaptation in LLMs.

2 Related Work

Evaluation of Instruction Following and Output Quality in LLMs. LLMs exhibit impressive capabilities but often display uncertainty in predictions, necessitating effective calibration for reliable outputs. (Kuleshov et al., 2018) introduce a recalibration method that aligns confidence scores with empirical accuracy, without altering the model's architecture. (Zhang et al., 2017) enhance calibration through mixup training, which generates convex combinations of inputs and labels. (Guo et al., 2017) analyze calibration errors and propose metrics, such as Expected Calibration Error (ECE) and Maximum Calibration Error (MCE), for model comparison. For LLMs, (Desai and Durrett, 2020) apply temperature scaling, while (Zhao et al., 2021) use ensemble methods to achieve calibrated consensus. (Tian et al., 2023) assess LLM confidence through direct querying, and (He et al., 2023) evaluate calibration using ECE, AUROC, and AUPRC. (Lyu et al., 2024) introduce coherence sampling to refine LLM calibration further.

Prompt Engineering and Optimization. Prompt engineering has significantly advanced interactions with LLMs. Few-shot and zero-shot learning techniques minimize the need for large labeled datasets by leveraging minimal examples to guide models (Brown et al., 2020). Automated prompt generation methods, such as those proposed by (Liu et al., 2023), use reinforcement learning to discover optimal prompts. Recent studies emphasize the role of LLMs in prompt optimization. (Ma et al., 2024) show that LLMs can refine prompts to enhance task performance. To address distribution shifts, (Li et al., 2023c) propose Generalized Prompt Optimization (GPO), improving LLM generalization under subpopulation shifts. (Yang et al., 2024) demonstrate that LLM-generated prompts via the Optimization by PROmpting (OPRO) method outperform manually crafted prompts.

The key differences between Transfer-Prompting and OPRO can be summarized in two aspects: First, Transfer-Prompting is a two-stage optimization framework consisting of source prompt construction and target prompt generation, whereas OPRO operates in a single-stage process to directly generate optimized prompts. Second, Transfer-Prompting incorporates domain-specific reference prompts to better adapt to target tasks, providing greater flexibility for task customization.

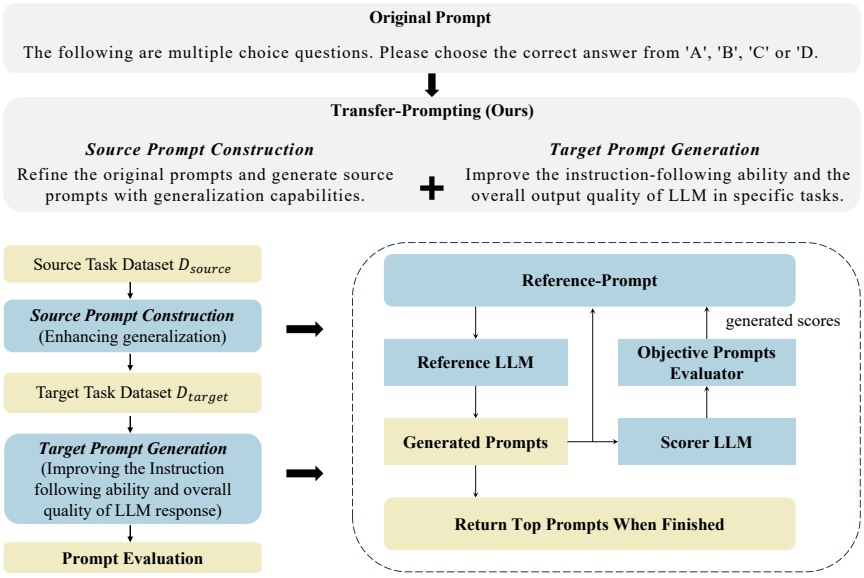

Figure 1: Illustration of the Two-Stage Prompt Optimization Framework in **Transfer-Prompting**: The framework consists of two main stages: source prompt construction and target prompt generation. It utilizes four key components: the reference LLM, reference prompt, scorer LLM, and objective prompt evaluator.

3 Methods

3.1 Preliminaries

We define two task sets: source tasks $\boldsymbol{\mathcal{S}} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_\kappa\}$ and target tasks $\boldsymbol{\mathcal{T}} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_\tau\}$, where κ and τ denote the number of source and target tasks respectively. Source tasks are designed to provide domain-specific knowledge, while target tasks focus on specific application scenarios. Each source task $\mathcal{S}_i$ is associated with a dataset $D_{\mathcal{S}_i} = \{(q_{i,n}, a_{i,n})\}_{n=1}^{M_i}$, where $q_{i,n}$ is the input and $a_{i,n}$ is the corresponding output, and M_i corresponds to the number of samples in source task $\mathcal{S}_i$. Similarly, each target task $\mathcal{T}_k$ is associated with $D_{\mathcal{T}_k} = \{(q_{k,m}, a_{k,m})\}_{m=1}^{N_k}$, where N_k denotes the sample count in target task $\mathcal{T}_k$.

The source task dataset is constructed by selecting related tasks from multiple datasets within the same domain to ensure domain consistency. This strategy enables the model to learn shared domain knowledge across similar tasks, enhancing its generalization capabilities. In contrast, the target task dataset is assembled by selecting specific tasks from datasets within a particular domain to maintain task focus. Details of the construction of the source task dataset are provided in Appendix A.2.

3.2 Transfer-Prompting Framework Design

LLMs often face challenges in balancing instruction-following, output quality, and other performance aspects, particularly on complex multi-task scenarios. To address these, we propose a novel LLM-based automatic optimization framework, **Transfer-Prompting**, designed to identify instructions that maximize target task performance.

As illustrated in Figure 1, the optimization process unfolds in two stages: (1) source prompt construction and (2) target prompt generation. The initial prompt, derived from domain expertise or random initialization, is first refined on the aggregated source task datasets $\bigcup_{i=1}^{\kappa} D_{\mathcal{S}_i}$ to produce generalized source prompts $\mathcal{P}_{\text{source}}$. Subsequently, high-performing $\mathcal{P}_{\text{source}}$ prompts are fine-tuned on the target task datasets $\bigcup_{k=1}^{\tau} D_{\mathcal{T}_k}$ to generate task-specific target prompts $\mathcal{P}_{\text{target}}$.

Prompt Optimization Strategy. At each iteration t, the reference LLM generates K candidate prompts $\{P_c^{(t)}\}_{c=1}^{K}$, which are scored using the objective prompt evaluator. The performance score $s_c^{(t)}$ aggregates the average accuracy of $P_c^{(t)}$ across datasets D as follows:

$$s_c^{(t)} = \sum_{d \in D} \phi(P_c^{(t)}, d), \tag{1}$$

where D represents the set of datasets under consideration (i.e., $D = \bigcup_{i=1}^{\kappa} D_{\mathcal{S}_i}$ for source tasks or $D = \bigcup_{k=1}^{\tau} D_{\mathcal{T}_k}$ for target tasks), and $\phi(P, d)$ denotes the average accuracy of prompt P over dataset d.

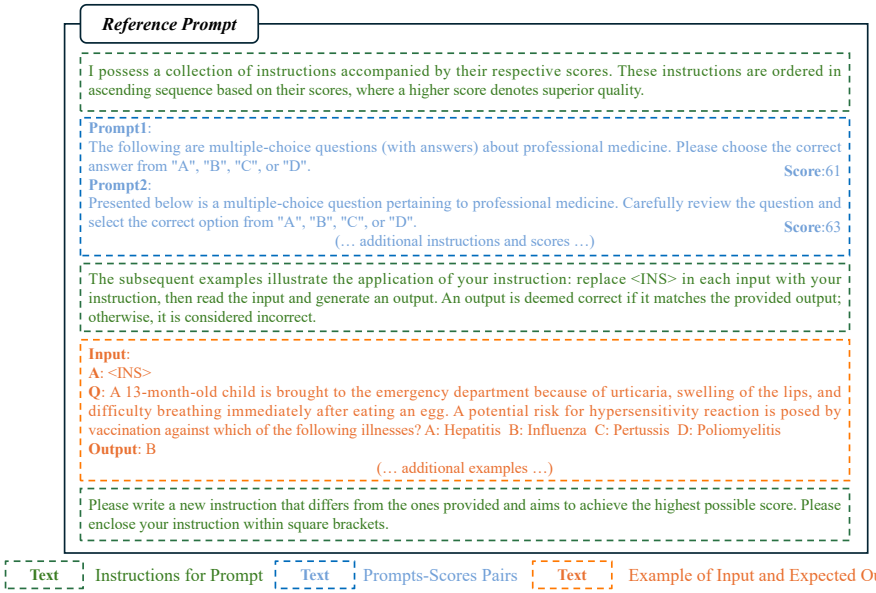

Figure 2: An example of the reference prompt for the reference LLMs (PaLM 2-L and PaLM 2-L-IT) on medically relevant datasets. The generated instruction is inserted at the position marked by <INS> in the input. The green text indicates instructions for prompts and scores, the orange text provides examples of how to apply the instruction, and the blue text displays the prompt-score pairs.

The optimization objective is to maximize the total accuracy across all prompts $\mathcal{P}$:

$$\mathcal{P}^* = \arg\max_{\mathcal{P}} \sum_{P \in \mathcal{P}} \sum_{d \in D} \phi(P, d). \quad (2)$$

This objective ensures a focused evaluation based on accuracy alone. The resulting scores $s_c^{(t)}$ guide the generation of new prompts until performance improvement is minimal or the maximum iteration limit is reached.

Reference LLM and Scorer LLM. In both stages of the optimization, we use advanced LLMs from different architectures as the reference LLM, which generates prompts based on the reference prompt. The most reliable LLM, selected for its robust and consistent performance, serves as the scorer LLM. As shown in Figure 2, the reference prompt consists of two components: (1) previously generated prompts with their corresponding scores and (2) a detailed description of the optimization problem, including task examples. The reference LLM generates new prompts at each iteration to improve instruction-following and overall task performance.

Objective Prompt Evaluation. The objective prompt evaluator uses accuracy as the sole optimization metric. To enhance efficiency, source prompt construction iteratively refines prompts on an aggregated dataset formed by selecting representative tasks from multiple domains. Target prompt generation is iteratively optimized on tasks selected from specific datasets to ensure domain consistency and task focus.

3.3 Source Prompt Construction

By refining the origin prompt on the aggregated source task datasets $\bigcup_{i=1}^{\kappa} D_{\mathcal{S}_i}$, we construct the source prompt set $\mathcal{P}_{\text{source}}$. The optimization goal is to identify a set of prompts $\mathcal{P}_{\text{source}}$ that maximizes domain-agnostic performance across source tasks:

$$\mathcal{P}_{\text{source}} = \arg\max_{P} \sum_{i=1}^{\kappa} \phi(P, D_{\mathcal{S}_i}), \quad (3)$$

where P denotes the current prompt being evaluated. At each training step, the reference LLM generates K candidate prompts based on the reference prompt (where K is a hyperparameter controlling the number of candidates per iteration). The scorer LLM then evaluates these prompts using the objective prompt evaluator. The highest-scoring prompts are selected for the next training step. The optimization process terminates when the reference LLM fails to generate new prompts with higher scores, or when the maximum number of optimization steps is reached. This results in the final source prompt set $\mathcal{P}_{\text{source}}$.

3.4 Target Prompt Generation

After constructing the source prompt set $\mathcal{P}_{\text{source}}$, we select a set of high-scoring prompts from $\mathcal{P}_{\text{source}}$ and fine-tune them on the corresponding target task datasets $\bigcup_{k=1}^{\tau} D_{\mathcal{T}_k}$, thereby generating a target prompt set $\mathcal{P}_{\text{target}}$ that is better suited for the target task.

The optimization objective for target prompt generation is defined as:

$$\mathcal{P}_{\text{target}} = \arg\max_{P} \sum_{k=1}^{\tau} \phi(P, D_{\mathcal{T}_k}), \qquad (4)$$

where P denotes the current prompt being evaluated. Starting from the highest-scoring source prompts selected from $\mathcal{P}_{\text{source}}$, the target prompt optimization process follows the same procedure as the source prompt optimization, resulting in the final target prompt set $\mathcal{P}_{\text{target}}$.

4 Experimental Setup

4.1 Models and Datasets

To evaluate the effectiveness of Transfer-Prompting, we tested 7 foundational models on 3 commonsense reasoning datasets: GPT-3.5-Turbo (OpenAI, 2023a), GPT-4 (OpenAI, 2023b), LLaMA-2 (7B & 13B) (Touvron et al., 2023), LLaMA-3-8B (AI@Meta, 2024), and Vicuna (7B & 13B) (Zheng et al., 2023). Additionally, we evaluated 18 specialized models on 6 multi-task datasets from medical, legal, and financial domains.

In the medical domain, we tested 6 specialized LLMs: ChatDoctor-13B (Li et al., 2023d), PMC-LLaMA-13B (Wu et al., 2023), MedAlpaca (7B & 13B) (Han et al., 2023), and Medicine-LLM (7B & 13B) (Cheng et al., 2023). In the legal domain, 6 law-specific LLMs were evaluated: DISC-LawLLM-13B (Yue et al., 2023), Lawyer-LLaMA-13B (Huang et al., 2023), ChatLaw-13B (Cui et al., 2024), LawGPT-7B (Zhou et al., 2024), and Law-LLM (7B & 13B) (Cheng et al., 2023). For the financial domain, we tested 6 LLMs: CFGPT-7B-Full (Li et al., 2023b), Tongyi-Finance-14B-Chat (JXY, 2024), FinGPT-13B-v2 (based on LLaMA-2-13B) (Yang et al., 2023a), FinMA-7B-Full (Xie et al., 2023), and Finance-LLM (7B & 13B) (Cheng et al., 2023). The foundational models were evaluated on 3 commonsense reasoning datasets: LogiQA (Liu et al., 2020), OpenBookQA (Mihaylov et al., 2018), and CosmosQA (Huang et al., 2019). For professional models, we used 6 multi-task datasets: **Medical Domain**: The corresponding medical models were evaluated using MMLU (Hendrycks et al., 2021), C-Eval (Huang et al., 2024), and MedMCQA (Pal et al., 2022) datasets. **Legal Domain**: Legal models were assessed with MMLU (Hendrycks et al., 2021), CMMLU (Li et al., 2023a), and AGIEval (Zhong et al., 2023) datasets. **Financial Domain**: Financial models were evaluated using CMMLU (Li et al., 2023a), C-Eval (Huang et al., 2024), and FinEval (Zhang et al., 2023) datasets.

4.2 Confidence Evaluation Methods

We employ the following methods to quantify model uncertainty:

Logits (Yang et al., 2023b): The model's predicted probabilities are interpreted as confidence scores, with the highest probability corresponding to the selected answer in multiple-choice questions.

Verbalized Confidence (Lin et al., 2022): By prompting LLMs, we obtain both answers and their associated confidence scores. These scores are used to evaluate the models' calibration by analyzing the relationship between accuracy and confidence across all valid responses.

4.3 Evaluation Metrics

We evaluate the instruction-following capabilities of LLMs using the instruction-following rate and accuracy. Additionally, we assess overall response quality using expected calibration error, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve.

Expected Calibration Error (ECE): ECE measures the alignment between predicted probabilities and actual outcomes, providing insight into model calibration quality. It is calculated as:

$$\text{ECE} = \sum_{i=1}^{n} \frac{|B_i|}{N} \cdot \left|\text{acc}(B_i) - \text{conf}(B_i)\right|, \qquad (5)$$

where n is the number of bins (defaulting to 10 in this study), B_i represents the samples in bin i, N is the total number of samples, $\text{acc}(B_i)$ is the accuracy within bin i, and $\text{conf}(B_i)$ is the mean predicted probability in bin i.

Area Under the Receiver Operating Characteristic Curve (AUROC): AUROC evaluates a binary classification model's ability to distinguish between positive and negative classes. It is derived

Table 1: Comparison of Zero-shot Learning Performance of Foundational Models Using Different Prompt Strategies on Commonsense Reasoning Datasets. The confidence is calculated by the verbalized confidence method. The best outcome is highlighted in **bold**.

Model	Method	LogiQA						OpenbookQA						CosmosQA					
		IFR ↑	ACC ↑	ECE ↓	ROC ↑	PR-P ↑	PR-N ↓	IFR ↑	ACC ↑	ECE ↓	ROC ↑	PR-P ↑	PR-N ↓	IFR ↑	ACC ↑	ECE ↓	ROC ↑	PR-P ↑	PR-N ↓
LLaMA-2-7B	Orign Prompt	0.40	**0.32**	0.54	**0.38**	**0.41**	**0.73**	0.48	**0.36**	0.47	**0.53**	**0.47**	**0.59**	0.45	0.33	0.58	0.42	0.43	0.72
	Transfer Prompt	**0.55**	0.29	**0.45**	0.36	0.38	0.78	**0.52**	0.35	**0.36**	0.49	0.44	0.73	**0.58**	**0.36**	**0.41**	**0.50**	**0.51**	**0.46**
LLaMA-2-13B	Orign Prompt	0.46	0.30	0.49	0.47	0.52	**0.67**	0.54	0.39	0.46	0.45	0.43	0.70	0.56	0.41	0.46	**0.59**	**0.57**	**0.54**
	Transfer Prompt	**0.57**	**0.37**	**0.34**	**0.57**	**0.54**	0.73	**0.65**	**0.45**	**0.27**	**0.56**	**0.54**	**0.55**	**0.64**	**0.43**	**0.30**	0.37	0.47	0.61
LLaMA-3-8B	Orign Prompt	0.66	0.44	0.42	0.63	0.55	0.59	0.72	0.43	0.35	0.61	0.55	0.49	0.69	0.46	0.26	0.67	0.66	0.45
	Transfer Prompt	**0.79**	**0.47**	**0.31**	**0.70**	**0.72**	**0.41**	**0.87**	**0.55**	**0.21**	**0.75**	**0.71**	**0.34**	**0.81**	**0.53**	**0.15**	**0.71**	**0.79**	**0.33**
Vicuna-7B	Orign Prompt	0.37	**0.29**	0.64	**0.44**	**0.36**	**0.75**	0.43	0.32	0.49	0.37	0.34	0.74	0.40	0.31	0.51	0.47	0.38	0.75
	Transfer Prompt	**0.46**	0.26	**0.44**	0.43	0.31	0.81	**0.51**	**0.36**	**0.45**	**0.50**	**0.42**	**0.64**	**0.52**	**0.34**	**0.43**	**0.63**	**0.58**	**0.70**
Vicuna-13B	Orign Prompt	0.43	0.32	0.49	0.45	0.46	0.72	0.53	0.36	0.48	0.49	0.40	0.67	0.51	0.36	0.49	0.55	0.44	0.64
	Transfer Prompt	**0.49**	**0.37**	**0.36**	**0.54**	**0.49**	**0.64**	**0.62**	**0.42**	**0.39**	**0.57**	**0.64**	**0.53**	**0.59**	**0.44**	**0.33**	**0.64**	**0.51**	**0.57**
GPT-3.5-Turbo	Orign Prompt	0.59	0.35	0.42	0.61	0.49	0.68	0.68	0.37	0.36	0.58	0.52	0.54	0.63	0.42	0.39	0.61	0.65	0.46
	Transfer Prompt	**0.71**	**0.39**	**0.27**	**0.73**	**0.68**	**0.40**	**0.77**	**0.49**	**0.23**	**0.70**	**0.69**	**0.37**	**0.75**	**0.51**	**0.20**	**0.68**	**0.71**	**0.35**
GPT-4	Orign Prompt	0.70	0.44	0.30	0.69	0.66	0.44	0.75	0.45	0.28	0.75	0.65	0.47	0.74	0.54	0.22	0.64	0.68	0.31
	Transfer Prompt	**0.82**	**0.50**	**0.18**	**0.81**	**0.74**	**0.32**	**0.89**	**0.58**	**0.16**	**0.83**	**0.76**	**0.29**	**0.87**	**0.59**	**0.14**	**0.74**	**0.85**	**0.19**

from the area under the ROC curve, which plots the true positive rate against the false positive rate across various thresholds.

Area Under the Precision-Recall Curve (PR-AUC) for Positive and Negative Classes (PR-P, PR-N): PR-P and PR-N measure a model's precision and recall for positive and negative classes, respectively. PR-P is particularly useful for evaluating performance on imbalanced datasets, while PR-N is essential for accurately identifying negative instances.

Instruction Following Rate (IFR): IFR quantifies the proportion of instances where the model's response adheres to the specified instructions. It is defined as:

$$\text{IFR} = \left(\frac{N_S}{N_T}\right) \times 100\%,$$

where N_S is the number of instances where the LLM's responses satisfy the specified requirements, and N_T is the total number of instructions attempted, including both successful and unsuccessful responses.

5 Results and Analysis

5.1 Performance Analysis on Commonsense Reasoning Datasets

This experiment evaluates the effectiveness of Transfer-Prompting in enhancing the performance of foundational LLMs on commonsense reasoning tasks. We selected three widely used benchmark datasets—LogiQA, OpenBookQA, and CosmosQA—to assess the reasoning capabilities of these models. The evaluated models include GPT-3.5-Turbo, GPT-4, LLaMA-2-7B, LLaMA-2-13B, LLaMA-3-8B, Vicuna-7B, and Vicuna-13B. The instruction-following ability and overall response quality of these LLMs were assessed under both zero-shot and five-shot settings. The Origin Prompt used examples as shown in Figure 1. In contrast, the Transfer Prompt was optimized using the second stage of the Transfer-Prompting framework, which generated high-scoring target prompts on the task-specific dataset.

The results in Table 1 demonstrate that Transfer-Prompting significantly improves the performance of most LLMs in the zero-shot setting, particularly GPT-4. Specifically, GPT-4's instruction-following rate (IFR) increases from 0.70 to 0.82, accuracy improves from 0.44 to 0.50, expected calibration error (ECE) decreases from 0.30 to 0.18, ROC increases from 0.69 to 0.81, PR-P increases from 0.66 to 0.74, and PR-N decreases from 0.44 to 0.32 on the LogiQA dataset. Additionally, the LLaMA series models, especially LLaMA-3-8B, show significant improvements, with IFR increasing from 0.66 to 0.79 on the LogiQA dataset. In contrast, the Vicuna series models show relatively smaller performance gains, potentially due to inherent architectural limitations. These results indicate that Transfer-Prompting significantly enhances both the instruction-following ability and overall response quality of LLMs, especially on complex commonsense reasoning tasks.

5.2 Performance Analysis on Sensitive Domains Datasets

This experiment evaluates 18 LLMs across three professional fields: medical, legal, and financial. The tasks from the MMLU and CMMLU datasets related to these fields are used for testing. The models evaluated include domain-specific models such as Med-Alpaca-13B, Law-LLM-13B, and Finance-LLM-13B. The performance indicators in-

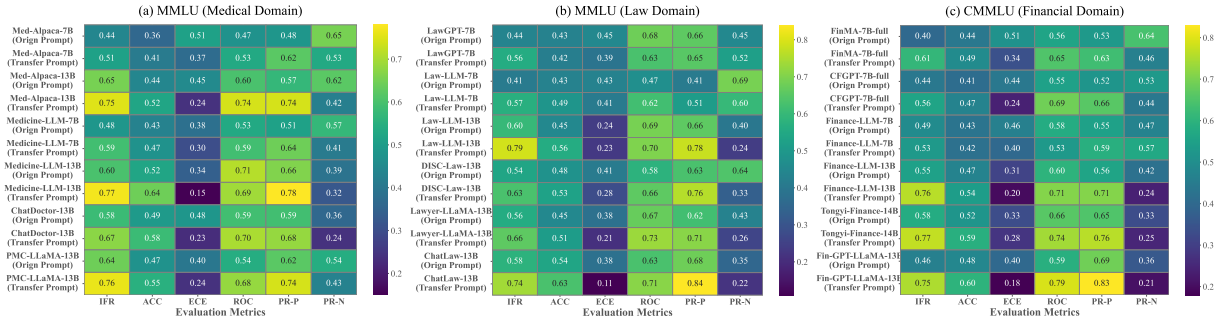

Figure 3: Comparative performance evaluation of various medical, legal, and financial models. The confidence is calculated by the verbalized confidence method.

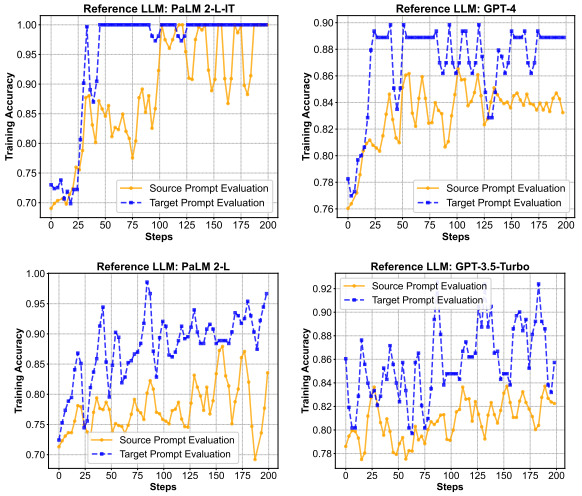

Figure 4: Score curves of the two-stage prompt optimization process of Transfer-Prompting on MMLU medical-related tasks.

clude IFR, accuracy (ACC), expected calibration error (ECE), receiver operating characteristic (ROC), and precision-recall metrics (PR-P and PR-N), providing a comprehensive evaluation of instruction compliance and overall response quality in these professional domains.

As shown in Figures 3 (a), (b), and (c), Transfer Prompt outperforms Origin Prompt across all fields. Specifically, in subfigure (a) of the medical MMLU dataset, Medicine-LLM-13B achieves the highest IFR (0.77), ACC (0.64), and PR-P (0.78), along with the lowest ECE (0.15) and PR-N (0.32) using Transfer Prompt. In subfigure (b) of the legal MMLU dataset, ChatLaw-13B achieves the highest ACC (0.63) and PR-P (0.84), with the lowest ECE (0.11) and PR-N (0.22) using the Transfer Prompt, significantly outperforming the Origin Prompt. Finally, in subfigure (c) of the financial CMMLU dataset, Fin-GPT-LLaMA-13B achieves the highest ACC (0.60), ROC (0.79), and PR-P (0.83), with the lowest ECE (0.18) and PR-N (0.21) using Transfer Prompt, again outperforming the Origin Prompt. These results demonstrate that Transfer Prompt consistently improves the model's instruction compliance and output quality across complex professional tasks.

5.3 Analysis of Source and Target Prompt Optimization Evaluation Process

In this experiment, we comprehensively evaluate the dual-stage prompt optimization process of Transfer-Prompting. A unified scorer LLM, PaLM 2-L, is used for the evaluation. Four reference LLMs—PaLM 2-L-IT, GPT-4, PaLM 2-L, and GPT-3.5-Turbo—serve as optimizers to generate candidate prompts for evaluation. The evaluation includes both source prompt evaluation (represented by the orange solid line) and target prompt evaluation (represented by the blue dashed line), with a total of 200 optimization steps performed on the MMLU medical-related task.

As shown in Figure 4, the dual-stage prompt optimization process of Transfer-Prompting significantly enhances the overall performance of the scorer LLM. The target prompt consistently outperforms the source prompt throughout the evaluation. In the case of the reference LLM PaLM 2-L-IT, near-perfect performance was achieved early in the optimization process, stabilizing quickly. Both GPT-4 and GPT-3.5-Turbo also showed steady improvements, with their scores eventually stabilizing between 0.88 and 0.9, further demonstrating the framework's adaptability and performance. Although PaLM 2-L exhibited some fluctuations, it displayed an overall upward trend, indicating that Transfer-Prompting can effectively optimize performance even for models with initially lower scores.

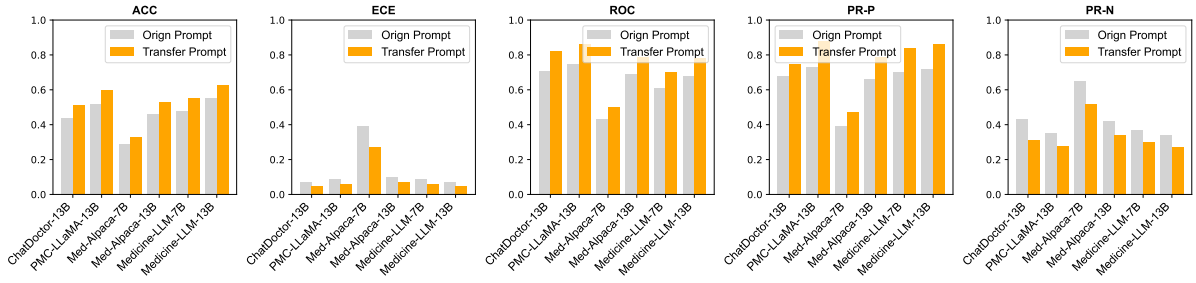

Figure 5: The zero-shot performance of different medical domain LLMs on MMLU medical-related tasks is evaluated using logits.

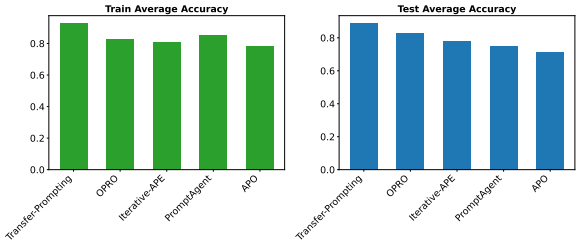

Figure 6: Performance Comparison with Baselines

5.4 Comparison with Baselines

To further demonstrate the effectiveness of the Transfer-Prompting method, we compared it with several state-of-the-art baseline approaches, including OPRO (Yang et al., 2024), Iterative-APE (Zhou et al., 2022), PromptAgent (Wang et al., 2023b), and APO (Pryzant et al., 2023). For optimization, PaLM 2-L-IT was used as the reference LLM, and PaLM 2-L was used as the scorer LLM. The dataset utilized for evaluation was the MMLU medical-related tasks. Data for Transfer-Prompting represents the average score from the second-stage optimization process.

As shown in Figure 6, Transfer-Prompting consistently outperforms all baseline methods. During the training phase (left), our method achieved the highest overall performance score. In the testing phase (right), Transfer-Prompting continued to demonstrate a significant advantage over the baseline methods. Additionally, the small difference between Transfer-Prompting's average scores in the training and testing phases suggests that the optimization process did not overfit the data. This sustained superiority highlights the robustness of our method, effectively generalizing to unseen data.

5.5 Analysis of Logits

To further validate the effectiveness of our method, this section evaluates Transfer-Prompting's impact on improving LLM performance in zero-shot and five-shot settings through Logits. The evaluation metrics include ACC, ECE, ROC, PR-P, and PR-N. We compared six medical-specialized LLMs—ChatDoctor-13B, PMC-LLaMA-13B, Med-Alpaca (7B & 13B), and Medicine-LLM (7B & 13B)—on MMLU medical-related tasks.

As shown in Figure 5, Transfer-Prompting significantly improves the performance of all models in the zero-shot setting. For example, the accuracy (ACC) of ChatDoctor-13B using Transfer Prompt increases from 0.44 to 0.51, indicating improved prediction accuracy. At the same time, its ECE decreases from 0.07 to 0.05, signifying better-calibrated predictions. Moreover, the ROC increases from 0.71 to 0.82, PR-P increases from 0.68 to 0.75, and PR-N decreases from 0.43 to 0.31, reflecting substantial improvements in the overall quality of the model output.

6 Conclusion

In this study, we introduce **Transfer-Prompting**, an innovative approach aimed at enhancing the generalization capabilities of LLMs by optimizing and adapting source prompts for specific target tasks. One of the key advantages of Transfer-Prompting is its ability to generate prompts that are finely tuned to the target dataset, leading to improved model performance. This adaptability makes it particularly well-suited for applications in diverse fields such as healthcare, legal, and financial services, where accurate and reliable model outputs are crucial. Furthermore, our approach is designed to address issues related to model calibration, ensuring that prediction confidence better aligns with actual accuracy. Extensive evaluations of both base models and domain-specific models show significant improvements in prediction accuracy, calibration, and instruction-following capabilities.

Limitations

Despite the promising results of the Transfer-Prompting framework, several limitations remain. One limitation is that the approach primarily focuses on optimizing prompts for LLMs, which may not fully address underlying issues with model architecture or fundamental capabilities. Additionally, while the framework demonstrates improvements across various tasks, it may not generalize equally well to all domains, especially those with highly specialized or dynamic requirements. Furthermore, the dependency on multiple LLMs for both prompt generation and evaluation may introduce computational resources and scalability challenges, particularly for large-scale implementations.

References

AI@Meta. 2024. Llama 3 model card.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, and 1 others. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Shijie Chen, Yu Zhang, and Qiang Yang. 2024. Multitask learning in natural language processing: An overview. *ACM Computing Surveys*, 56(12):1–32.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. *arXiv preprint arXiv:2309.09530*.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2024. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *Preprint*, arXiv:2306.16092.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Guande He, Peng Cui, Jianfei Chen, Wenbo Hu, and Jun Zhu. 2023. Investigating uncertainty calibration of aligned language models under the multiple-choice setting. *arXiv preprint arXiv:2310.11732*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *Preprint*, arXiv:2305.15062.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, and 1 others. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

JXY. 2024. Tongyi-finance-14b-chat. Accessed: 2024-06-15.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR.

Nilesh D Kulkarni and Preeti Tupsakhare. 2024. Crafting effective prompts: Enhancing ai performance through structured input design. *JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING (JRTCSE)*, 12(5):1–10.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023b. Cfgpt: Chinese financial assistant with large language model. *Preprint*, arXiv:2309.10654.

Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang, and Tat-Seng Chua. 2023c. Robust prompt optimization for large language models against distribution shifts. Association for Computational Linguistics.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023d. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2024. Calibrating large language models with sample consistency. *Preprint*, arXiv:2402.13904.

Ruotian Ma, Xiaolei Wang, Xin Zhou, Jian Li, Nan Du, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Are large language models good prompt optimizers? *arXiv preprint arXiv:2402.02101*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2023a. `https://chat.openai.com.chat`.

OpenAI. 2023b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Hong Sun, Xue Li, Yinchuan Xu, Youkow Homma, Qi Cao, Min Wu, Jian Jiao, and Denis Charles. 2023. Autohint: Automatic prompt optimization with hint generation. *arXiv preprint arXiv:2307.07415*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P Xing, and Zhiting Hu. 2023b. Promptagent: Strategic planning with language models enables expert-level prompt optimization. *arXiv preprint arXiv:2310.16427*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *Preprint*, arXiv:2304.14454.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *Preprint*, arXiv:2306.05443.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. *Preprint*, arXiv:2309.03409.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023a. Fingpt: Open-source financial large language models. *Preprint*, arXiv:2306.06031.

Yuchen Yang, Houqiang Li, Yanfeng Wang, and Yu Wang. 2023b. Improving the reliability of large language models by leveraging uncertainty-aware in-context learning. *arXiv preprint arXiv:2310.04782*.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *Preprint*, arXiv:2309.11325.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, and 1 others. 2023. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model. *Preprint*, arXiv:2406.04614.

Appendix

Contents

A Details of Models and Datasets

A.1 Details of Models

We selected a diverse set of models to evaluate the performance of both foundational and domain-specific LLMs. This selection allows us to assess the broad applicability of Transfer-Prompting and evaluate its effectiveness across specialized domains. By comparing these models, we aim to demonstrate the potential and advantages of Transfer-Prompting comprehensively.

For **foundational models**, we used *GPT-3.5-Turbo*, *GPT-4*, *LLaMA2-7B*, *LLaMA2-13B*, *LLaMA3-8B*, *Vicuna-7B*, and *Vicuna-13B* in our experiments. These models serve as baselines to evaluate the broader applicability of Transfer-Prompting across general-purpose LLMs.

For **domain-specific models**, we evaluated models tailored to three critical domains: medicine, law, and finance. This allows us to investigate how domain-specific adaptations enhance model performance when applied to specialized data.

Medicine: In the medical domain, we selected *ChatDoctor-13B*, *PMC-LLaMA-13B*, *MedAlpaca-7B & 13B*, and *AdaptLLM-Medicine-LLM-7B & 13B*. These models are designed to handle complex medical queries and generate accurate medical information, which is essential for real-world medical applications.

Law: For the legal domain, we evaluated *DISC-LawLLM*, *LawGPT-7B*, *Lawyer-LLaMA-13B*, *ChatLaw-13B*, and *AdaptLLM-Law-LLM-7B & 13B*. These models specialize in interpreting and generating legal text, making them crucial for legal research, document drafting, and case analysis.

Finance: In the financial domain, we selected *FinGPT-13B-v2 (LLaMA2-13B-based)*, *CFGPT-7B-full*, *Tongyi-Finance-14B-Chat*, *AdaptLLM-Finance-LLM-7B & 13B*, and *FinMA-7B-full*. These models are specialized in interpreting financial data and forecasting, which are vital for market analysis, risk assessment, and financial planning.

A.2 Details of Datasets

Our experiments comprehensively evaluate model performance on commonsense reasoning using three datasets and on multiple-question answering (MQA) tasks involving sensitive data across five distinct datasets. The commonsense reasoning datasets include LogiQA[1], OpenBookQA[2], and CosmosQA[3]. For evaluation, we selected the top 1,000 questions from the LogiQA and OpenBookQA test sets and the validation set of CosmosQA.

For MQA on sensitive data, we evaluated the following datasets:

MMLU[4]: MMLU (Massive Multitask Language Understanding) is a benchmark designed to evaluate language models across 57 subjects, consisting of approximately 16,000 multiple-choice questions. We selected specific MMLU subsets to evaluate the performance of medical-related LLMs, including *medical genetics*, *professional medicine*, and *college medicine*. Additionally, we chose *college law*, *legal and moral basis*, and *international law* to assess law-related LLMs.

C-Eval[5]: C-Eval is a comprehensive Chinese evaluation suite containing 13,948 multiple-choice questions across 52 disciplines and four difficulty levels. We selected data from C-Eval to evaluate medical-related LLMs, focusing on *physician*, *clinical medicine*, and *basic medicine*, and for law-focused LLMs, we chose datasets such as *law*, *legal and moral basis*, and *international law*.

CMMLU[6]: CMMLU is a benchmark with 11,582 multiple-choice questions across 67 subjects, designed to evaluate language models' knowledge and reasoning in a Chinese context. We selected data from CMMLU to assess the performance of finance-related LLMs, including *business ethics*, *economics*, *marketing*, and *professional accounting*.

MedMCQA[7]: MedMCQA is a large-scale medical multiple-choice question-answering dataset with over 194,000 questions, designed to advance research in intelligent question-answering systems within the medical domain. We selected the first 1,000 questions from the test split of MedMCQA for evaluation.

AGIEval[8]: AGIEval is a benchmark designed to evaluate foundation models in human cognition and problem-solving tasks, including law school admission tests and lawyer qualification exams. We used the law-related data to assess legal LLMs' understanding of judicial examination questions and

[1] https://paperswithcode.com/dataset/logiqa
[2] https://paperswithcode.com/dataset/openbookqa
[3] https://paperswithcode.com/dataset/cosmosqa
[4] https://paperswithcode.com/dataset/mmlu
[5] https://paperswithcode.com/paper/c-eval-a-multi-level-multi-discipline-chinese-1
[6] https://paperswithcode.com/paper/cmmlu-measuring-massive-multitask-language
[7] https://paperswithcode.com/dataset/medmcqa
[8] https://github.com/ruixiangcui/AGIEval

case analyses, specifically the first 1,000 questions from the *jec-qa-kd* and *jec-qa-ca* tasks.

FinEval[9]: FinEval is a compilation of high-quality multiple-choice and text-based quiz questions designed specifically for the Chinese financial sector. We selected *advanced financial accounting*, *financial markets*, and *corporate finance* datasets for the evaluation of finance-related LLMs.

B Prompt Templates for Source Prompt and Transfer Prompt

As shown in Table 3, the comparison involves two types of prompts: source prompts and transfer prompts. Source prompts provide general instructions for answering multiple-choice questions, enhancing their generalization ability. In contrast, transfer prompts incorporate specific medical context and guidance to improve the overall quality of LLM responses. For example, PaLM 2-L-IT achieves a score of 43% using source prompts, while the score increases to 56% when medical context is included in the transfer prompt. This comparison underscores the importance of tailoring prompts to the context of specific domains to enhance the performance of language models in specialized fields.

C More Results

C.1 Evaluation of Commonsense Reasoning Capabilities

Table 2 compares the five-shot learning performance of various models on commonsense reasoning datasets. The results demonstrate that Transfer-Prompting significantly enhances model performance, particularly for GPT-4, which exhibits substantial improvements across key metrics, including instruction-following rate (IFR), accuracy (ACC), and expected calibration error (ECE). Other models, such as LLaMA3-8B and Vicuna-13B, also show notable improvements, highlighting the effectiveness of Transfer-Prompting in enhancing score, confidence calibration, and generalization across various commonsense reasoning tasks. These findings emphasize Transfer-Prompting's robustness and its potential to elevate the capabilities of multiple LLMs in complex reasoning scenarios.

[9] https://huggingface.co/datasets/FinGPT/fingpt-fineval

C.2 Performance Analysis on Sensitive Domains

In the **legal field**, as shown in Figure 7, the application of Transfer-Prompting significantly improves model performance. For example, in the case of LawGPT-7B, after applying Transfer-Prompting, the instruction-following rate (IFR) increases from 0.65 to 0.78, and the expected calibration error (ECE) decreases from 0.32 to 0.21, demonstrating improvements in both inference quality and model calibration. Similarly, the IFR of Law-LLM-13B improves from 0.72 to 0.83, and its accuracy (ACC) improves from 0.48 to 0.57. These results highlight the potential of Transfer-Prompting in applications that require high accuracy and confidence, such as legal contexts.

In the **financial field**, as shown in Figure 7, Transfer-Prompting also leads to significant performance improvements. For instance, the IFR of Finance-LLM-13B improves from 0.69 to 0.81, and the ACC increases from 0.49 to 0.58. Additionally, the ECE decreases across all models. These results confirm that Transfer-Prompting is crucial for enhancing model performance in the financial domain.

In summary, the results from the legal and financial domains are consistent with those observed in the medical domain, further demonstrating the generalizability and effectiveness of Transfer-Prompting in improving LLM performance across various sensitive professional domains.

C.3 Analysis of Logits

In this study, we use the LLaMA-Factory[10] to evaluate logits and analyze the effectiveness of Transfer-Prompting in enhancing LLM performance. As shown in Figure 8, the five-shot results further confirm the effectiveness of Transfer-Prompting, demonstrating consistent improvements across various models and metrics, similar to the zero-shot findings. Transfer-Prompting significantly boosts accuracy (ACC), reduces expected calibration error (ECE) and precision-recall negative (PR-N), and improves receiver operating characteristic (ROC) and precision-recall positive (PR-P) values, particularly for complex models like Med-Alpaca-13B and Medicine-LLM-13B. These results highlight Transfer-Prompting's reliability and versatility, establishing it as a valuable technique for enhancing

[10] https://github.com/hiyouga/LLaMA-Factory

Table 2: Comparison of five-shot learning performance of foundational models using different prompt strategies on commonsense reasoning datasets. The confidence is calculated by the verbalized confidence. The best outcome is highlighted in **bold**.

Model	Method	LogiQA						OpenbookQA						CosmosQA					
		IFR ↑	ACC ↑	ECE ↓	ROC ↑	PR-P ↑	PR-N ↓	IFR ↑	ACC ↑	ECE ↓	ROC ↑	PR-P ↑	PR-N ↓	IFR ↑	ACC ↑	ECE ↓	ROC ↑	PR-P ↑	PR-N ↓
Llama2-7B	Orign Prompt	0.44	**0.37**	0.52	**0.45**	**0.48**	**0.70**	0.45	**0.40**	0.49	**0.55**	**0.54**	0.64	0.49	0.37	0.52	0.45	0.45	0.65
	Transfer Prompt	**0.57**	0.35	**0.40**	0.41	0.45	0.74	**0.60**	0.39	**0.43**	0.35	0.48	**0.57**	**0.63**	**0.42**	**0.45**	**0.56**	**0.57**	**0.51**
Llama2-13B	Orign Prompt	0.55	0.38	0.45	0.49	0.56	**0.59**	0.56	0.37	0.52	0.41	0.39	0.75	0.54	**0.46**	0.47	**0.63**	**0.67**	**0.35**
	Transfer Prompt	**0.63**	**0.41**	**0.38**	**0.59**	**0.66**	0.68	**0.69**	**0.48**	**0.30**	**0.59**	**0.65**	**0.51**	**0.66**	0.45	**0.32**	0.59	0.65	0.48
Llama3-8B	Orign Prompt	0.71	0.43	0.35	0.67	0.72	0.36	0.76	0.44	0.30	0.67	0.60	0.43	0.74	0.46	0.25	0.67	0.71	0.46
	Transfer Prompt	**0.80**	**0.47**	**0.21**	**0.79**	**0.77**	**0.25**	**0.89**	**0.57**	**0.17**	**0.81**	**0.76**	**0.28**	**0.87**	**0.53**	**0.11**	**0.78**	**0.83**	**0.25**
Vicuna-7B	Orign Prompt	0.42	**0.29**	**0.55**	0.41	**0.42**	**0.73**	0.46	0.27	0.55	0.42	0.34	0.77	0.43	0.29	0.55	0.41	0.33	0.83
	Transfer Prompt	**0.50**	0.27	0.47	**0.47**	0.30	0.76	**0.63**	**0.38**	**0.31**	**0.51**	**0.48**	**0.58**	**0.65**	**0.39**	**0.37**	**0.68**	**0.46**	**0.66**
Vicuna-13B	Orign Prompt	0.49	0.33	0.37	0.47	0.38	0.78	0.58	0.35	0.42	0.54	0.40	0.67	0.63	0.44	0.45	0.55	0.57	0.50
	Transfer Prompt	**0.63**	**0.37**	**0.34**	**0.53**	**0.49**	**0.65**	**0.66**	**0.44**	**0.36**	**0.58**	**0.59**	**0.51**	**0.67**	**0.49**	**0.28**	**0.71**	**0.64**	**0.44**
GPT-3.5-Turbo	Orign Prompt	0.68	0.37	0.37	0.61	0.56	0.61	0.74	0.42	0.32	0.64	0.57	0.49	0.67	0.48	0.35	0.68	0.70	0.32
	Transfer Prompt	**0.77**	**0.45**	**0.23**	**0.78**	**0.70**	**0.36**	**0.81**	**0.55**	**0.19**	**0.76**	**0.74**	**0.34**	**0.84**	**0.56**	**0.18**	**0.75**	**0.79**	**0.22**
GPT-4	Orign Prompt	0.78	0.47	0.26	0.75	0.67	0.39	0.83	0.50	0.21	0.78	0.76	0.44	0.75	0.55	0.18	0.70	0.75	0.30
	Transfer Prompt	**0.86**	**0.56**	**0.12**	**0.88**	**0.80**	**0.24**	**0.91**	**0.63**	**0.13**	**0.86**	**0.87**	**0.27**	**0.89**	**0.64**	**0.09**	**0.88**	**0.92**	**0.16**

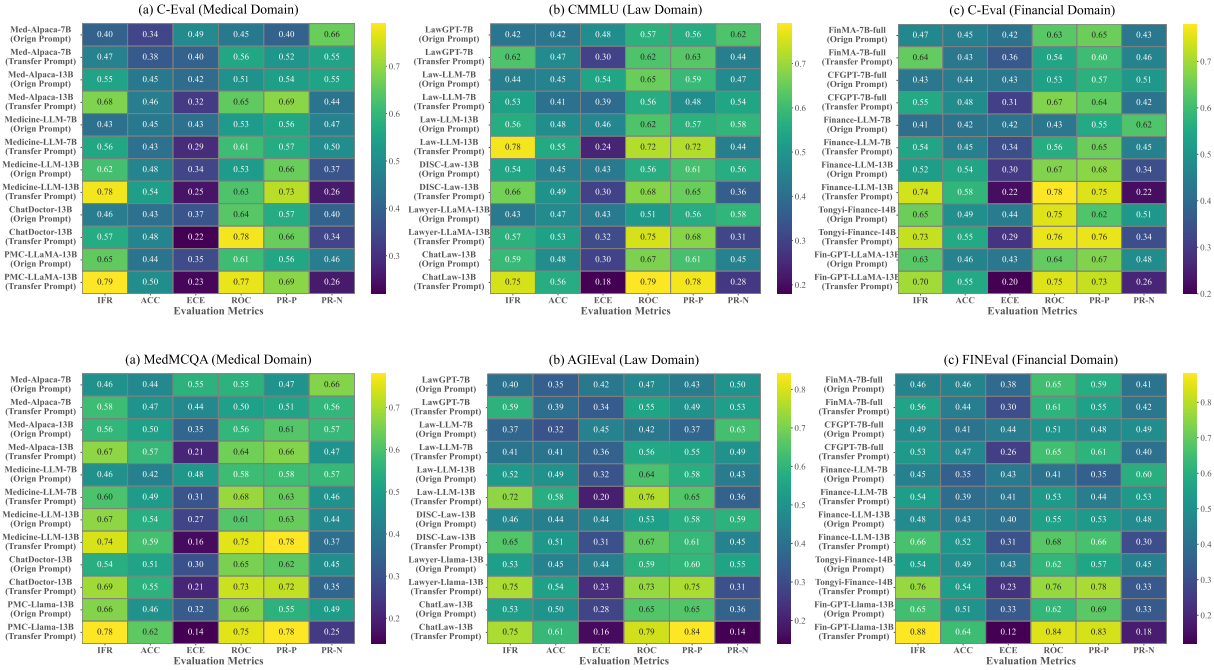

Figure 7: Comparative performance evaluation of various models in the medical, legal, and financial domains. The confidence is calculated by the verbalized confidence method.

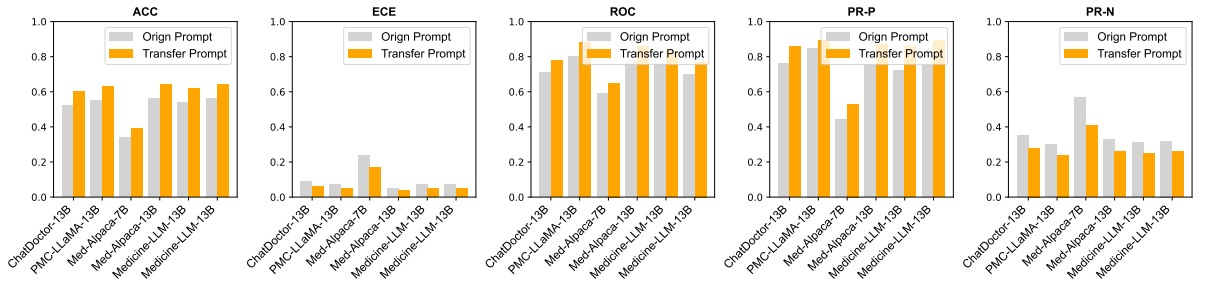

Figure 8: The five-shot performance of different medical domain LLMs on MMLU medical-related tasks is evaluated using logits.

LLM performance across diverse and critical domains.

D Reference-Prompt Template for GPT-3.5-Turbo and GPT-4

Figure 9 illustrates the optimization of a Reference-Prompt Template for GPT-3.5-Turbo and GPT-4 in

Table 3: An example of the two-stage prompt generation for medical-related tasks using the Transfer-Prompting method. These prompts are generated by the corresponding reference LLM, PaLM 2-L-IT, and the corresponding scorer LLM, PaLM 2-L, provides the respective scores.

Reference LLM	Prompt Type	Prompt	Score
PaLM 2-L-IT	Source	Answer the following multiple-choice questions by selecting the most accurate option from 'A', 'B', 'C', or 'D'. Use your general knowledge across various domains to provide the best answer.	46%
	Source	For each question below, choose the correct answer from 'A', 'B', 'C', or 'D'. Consider all relevant information to ensure accuracy.	43%
	Source	Carefully read each multiple-choice question and select the correct option ('A', 'B', 'C', or 'D') based on your comprehensive understanding of the subject matter.	37%
	Transfer	As a medical expert, answer the following questions by selecting 'A', 'B', 'C', or 'D'. Provide the most accurate answer based on medical knowledge and clinical evidence.	**61%**
	Transfer	Utilize your medical expertise to select the correct answer from 'A', 'B', 'C', or 'D' for each of the following medical questions. Ensure your choice reflects current best practices.	58%
	Transfer	Carefully read each medical question and choose the correct answer from 'A', 'B', 'C', or 'D'. Base your selection on established medical guidelines and evidence-based practice.	55%
	Transfer	Apply clinical reasoning to answer the following medical multiple-choice questions by selecting 'A', 'B', 'C', or 'D'. Choose the option that best fits the clinical scenario presented.	52%
GPT-4	Source	Answer the following multiple-choice questions by selecting the most appropriate option from 'A', 'B', 'C', or 'D'. Draw upon your broad knowledge base to ensure accuracy.	42%
	Source	For each question, select the correct answer from 'A', 'B', 'C', or 'D'. Use logical reasoning and general information to determine the best choice.	40%
	Source	Read the following questions carefully and choose the correct option ('A', 'B', 'C', or 'D') based on your overall understanding.	38%
	Transfer	As an experienced medical professional, answer the following questions by selecting 'A', 'B', 'C', or 'D'. Utilize critical thinking and advanced medical knowledge to provide the most accurate answer.	**57%**
	Transfer	For each of the following medical questions, select the correct answer from 'A', 'B', 'C', or 'D'. Use your knowledge of medical science and current clinical guidelines to inform your choice.	55%
	Transfer	Answer the following medical multiple-choice questions by selecting 'A', 'B', 'C', or 'D'. Ensure your answers are based on evidence-based medical practices and the latest research findings.	53%
	Transfer	Apply your medical expertise to select the most appropriate answer from 'A', 'B', 'C', or 'D' for each question. Base your choices on up-to-date medical knowledge and best practices.	50%

Table 4: An example of the two-stage prompt generation for legal-related tasks using the Transfer-Prompting method. These prompts are generated by the corresponding reference LLM, PaLM 2-L-IT, and the corresponding scorer LLM, PaLM 2-L, provides the respective scores.

Reference LLM	Prompt Type	Prompt	Score
PaLM 2-L-IT	Source	Answer the following multiple-choice questions by selecting the most accurate option from 'A', 'B', 'C', or 'D'. Use your general knowledge across various domains to provide the best answer.	44%
	Source	For each question below, choose the correct answer from 'A', 'B', 'C', or 'D'. Consider all relevant information to ensure Score.	39%
	Source	Carefully read each multiple-choice question and select the correct option ('A', 'B', 'C', or 'D') based on your comprehensive understanding of the subject matter.	35%
	Transfer	Analyze the following legal scenarios and choose the most legally sound answer from 'A', 'B', 'C', or 'D'. Apply principles of law and precedents to support your selection.	**55%**
	Transfer	Evaluate each case presented below and determine the correct legal outcome by selecting 'A', 'B', 'C', or 'D'. Use statutory interpretation and legal reasoning in your analysis.	52%
	Transfer	Review the following legal questions and select the appropriate answer from 'A', 'B', 'C', or 'D'. Consider current laws and judicial decisions in your decision-making process.	50%
	Transfer	Apply your understanding of legal concepts to answer the following multiple-choice questions by selecting 'A', 'B', 'C', or 'D'. Your answers should reflect accurate legal interpretations.	48%
GPT-4	Source	Answer the following multiple-choice questions by selecting the most appropriate option from 'A', 'B', 'C', or 'D'. Draw upon your broad knowledge base to ensure Score.	38%
	Source	For each question, select the correct answer from 'A', 'B', 'C', or 'D'. Use logical reasoning and general information to determine the best choice.	35%
	Source	Read the following questions carefully and choose the correct option ('A', 'B', 'C', or 'D') based on your overall understanding.	33%
	Transfer	Interpret the legal issues in the following questions and select 'A', 'B', 'C', or 'D' as the correct answer. Justify your choice based on legal doctrines and case law.	**53%**
	Transfer	For each legal problem below, determine the most appropriate resolution by choosing 'A', 'B', 'C', or 'D'. Incorporate relevant statutes and legal principles in your reasoning.	50%
	Transfer	Assess the following situations and choose the correct legal response from 'A', 'B', 'C', or 'D'. Your answers should be informed by an understanding of jurisprudence and legal ethics.	48%
	Transfer	Utilize your legal expertise to answer the following questions by selecting 'A', 'B', 'C', or 'D'. Consider the implications of your choice within the context of existing law.	46%

Table 5: An example of the two-stage prompt generation for financial-related tasks using the Transfer-Prompting method. These prompts are generated by the corresponding reference LLM, PaLM 2-L-IT, and the corresponding scorer LLM, PaLM 2-L, provides the respective scores.

Reference LLM	Prompt Type	Prompt	Score
PaLM 2-L-IT	Source	Answer the following multiple-choice questions by selecting the most accurate option from 'A', 'B', 'C', or 'D'. Use your general knowledge across various domains to provide the best answer.	42%
	Source	For each question below, choose the correct answer from 'A', 'B', 'C', or 'D'. Consider all relevant information to ensure the Score.	40%
	Source	Carefully read each multiple-choice question and select the correct option ('A', 'B', 'C', or 'D') based on your comprehensive understanding of the subject matter.	38%
	Transfer	Solve the following econometric problems by selecting 'A', 'B', 'C', or 'D' as the correct answer. Apply econometric theories and statistical techniques in your calculations.	**55%**
	Transfer	For each question related to econometric analysis, choose the most accurate answer from 'A', 'B', 'C', or 'D'. Use your knowledge of regression models and data interpretation to inform your choice.	52%
	Transfer	Examine the following econometrics questions and select the correct option from 'A', 'B', 'C', or 'D'. Consider assumptions of econometric models and statistical inference in your reasoning.	50%
	Transfer	Apply quantitative methods to answer the following multiple-choice questions by choosing 'A', 'B', 'C', or 'D'. Base your answers on sound econometric practices and empirical evidence.	48%
GPT-4	Source	Answer the following multiple-choice questions by selecting the most appropriate option from 'A', 'B', 'C', or 'D'. Draw upon your broad knowledge base to ensure Score.	40%
	Source	For each question, select the correct answer from 'A', 'B', 'C', or 'D'. Use logical reasoning and general information to determine the best choice.	38%
	Source	Read the following questions carefully and choose the correct option ('A', 'B', 'C', or 'D') based on your overall understanding.	35%
	Transfer	Analyze the econometric scenarios provided and select 'A', 'B', 'C', or 'D' as the correct answer. Utilize advanced econometric concepts and statistical analysis to support your decision.	**58%**
	Transfer	For each of the following econometrics problems, determine the correct answer by choosing 'A', 'B', 'C', or 'D'. Apply knowledge of time series analysis and econometric modelling.	55%
	Transfer	Evaluate the econometric questions below and select the appropriate option from 'A', 'B', 'C', or 'D'. Your choices should reflect an understanding of hypothesis testing and estimation techniques.	53%
	Transfer	Use your expertise in econometrics to answer the following questions by selecting 'A', 'B', 'C', or 'D'. Consider the statistical properties and limitations of the models involved.	50%

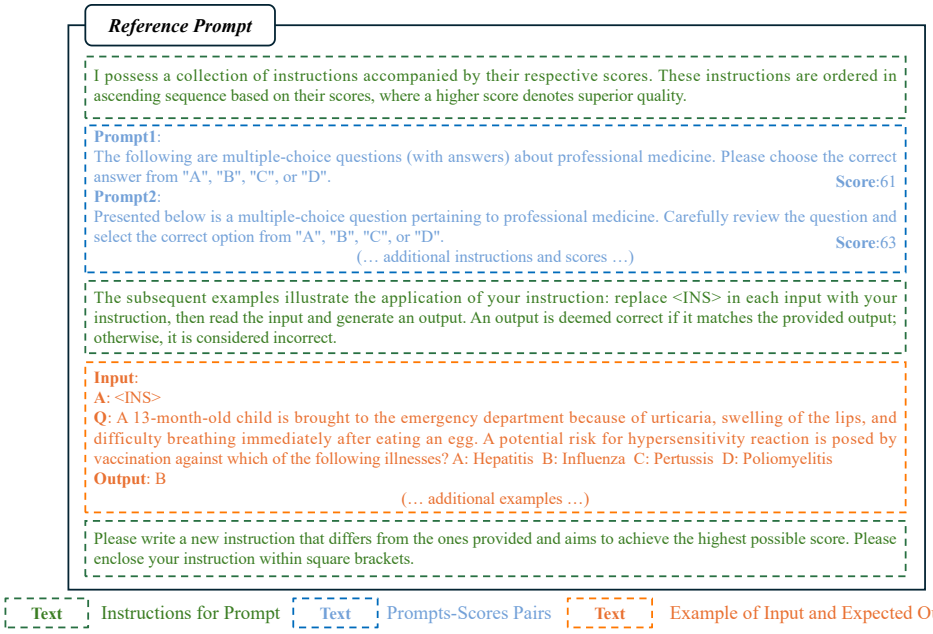

Figure 9: An example of the reference prompt for reference LLM (GPT-3.5-Turbo and GPT-4) on the medically relevant datasets. The generated instruction is inserted at the position marked by <INS> in the input. The green text displays instructions for prompts and scores; the orange text provides examples of how to apply the instruction; the blue text contains the prompts and scores pairs.

the context of medical multiple-choice questions. It provides examples of instructions along with their corresponding scores, with the task being to create a new instruction that improves performance. The figure also demonstrates how to integrate this new instruction into a prompt and evaluate its effectiveness.