Enhancing Large Language Models on Domain-specific Tasks: A Novel Training Strategy via Domain Adaptation and Preference Alignment

[1]Jingyang Deng, [1]Zeren Zhang, [1]Jo-Ku Cheng and [1]Jinwen Ma

[1]School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China

Introduction

As the complexity and specificity of real-world tasks increase, the limitations of general-purpose LLMs have become increasingly apparent, which are mainly caused by the following two factors:

- **Knowledge gap**. During the pre-training phase, general LLMs generally lack of exposure to domain-specific corpora, resulting in insufficient training for specific domain knowledge and terminology.
- **Value disagreement**. It is also worthy to note that general LLMs are trained to align with universal human values during the preference alignment phase, which may unfortunately not be beneficial to meet the requirements of specific domain tasks.

Contribution:

- **Pioneering case study in the unique domain of SOAEs** (Chinese State-Owned Assets and Enterprises). To our knowledge, this is the first work to utilize LLMs for this specific field, which has been underserved by general LLMs due to its specialized terminologies and intricate contexts.
- **Improved DAP with replay mechanism for mitigating catastrophic forgetting**. We incorporate a replay mechanism during the DAP phase, ensuring that the model not only acquires domain-specific knowledge but also retains its general language abilities thereby mitigating the issue of catastrophic forgetting.
- **Innovative data scheduling for instruction following and enhanced preference alignment**. Only a subset of domain-specific data is utilized for SFT, while others are combined with the low-quality labelled data to construct KTO-based preference alignment datasets. As a result, we effectively utilize even low-quality or imperfectly annotated data, which are often discarded in conventional training procedures.

Related Work

Existing Solutions and Limitations for Domain-Specific LLMs:

- Supervised Fine-tuning (SFT): Mainly improves instruction-following ability, but is poor at learning domain knowledge.
- Domain-Adaptive Pre-training (DAP) + SFT: Enhances domain-specific capabilities, but faces catastrophic forgetting, losing general-purpose abilities.
- It's challenging to collect domain-specific preference datasets, and high-quality data requires annotation by domain experts.

Existing Preference Alignment Methods for Domain-Specific LLMs:

- Proximal Policy Optimization (PPO): Highly reward-model-dependent, with complex and unstable training.
- Direct Preference Optimization (DPO): Simplifies training, eliminates the need for explicit rewards, and is mainstream. However, it still requires costly and time-consuming **pairwise** preference data collection.

Methodology

Step1. Adapting General-Purpose Pre-trained LLMs to the SOAEs Field

- Introduce a replay mechanism to prevent catastrophic forgetting.

$$X^{\mathrm{DAP}}_{\mathrm{domain}} \cup X^{\mathrm{DAP}}_{\mathrm{general}} \triangleq X^{\mathrm{DAP}} \qquad L_{\mathrm{DAP}}(\theta, X^{\mathrm{DAP}}) = \mathbb{E}_{x\sim X^{\mathrm{DAP}}}\left[-\sum_{t=1}^{T}\log p(w_i|w_{<i},\theta)\right]$$

Step 2: Supervised Fine-Tuning (SFT):

- Use subset of high-quality domain data $X^{\mathrm{SFT}}_{\mathrm{high}} \subset X_{\mathrm{high}}$

$$L_{\mathrm{SFT}}(\theta, X^{\mathrm{SFT}}_{\mathrm{high}}) = \mathbb{E}_{x\sim X^{\mathrm{SFT}}_{\mathrm{high}}}\left[-\sum_{t=a}^{T}\log p(w_i|w_{<i},\theta)\right]$$

Step 3: Preference Alignment with KTO:

- Combine low-quality data (discarded in traditional SFT) with remaining SFT data. $X^{\mathrm{KTO}}_{\mathrm{high}} = X_{\mathrm{high}} \backslash X^{\mathrm{SFT}}_{\mathrm{high}} \quad X^{\mathrm{KTO}} = X^{\mathrm{KTO}}_{\mathrm{high}} \cup X_{\mathrm{low}}$
- KTO Utility Function: Reward desired responses, penalize undesired ones **without paired preferences**.

$$L_{\text{KTO-SFT}}(\theta, X^{\mathrm{KTO}}) = L_{\mathrm{KTO}}(\theta, X^{\mathrm{KTO}}) + \mu L_{\mathrm{SFT}}(\theta, X^{\mathrm{KTO}}_{\mathrm{high}})$$

$$L_{\mathrm{KTO}}(\theta, X^{\mathrm{KTO}}) = \mathbb{E}_{x,y\sim X^{\mathrm{KTO}}}\left[\lambda(y)\left(1-\nu(x,y)\right)\right],$$

$$\nu(x,y) = \sigma(\beta\cdot y(r(x) - z_{\mathrm{ref}}))$$

$$r(x) = \log\frac{\pi_\theta(a|q)}{\pi_{\mathrm{ref}}(a|q)} \quad z_{\mathrm{ref}} = \mathbb{E}_{x\sim X^{\mathrm{KTO}}}\left[\mathrm{KL}(\pi_\theta(a|q)\|\pi_{\mathrm{ref}}(a|q))\right]$$

Experimental Results

1. Effectiveness of DAP & SFT 2. The necessity for domain-specific LLMs

TABLE I
COMPARISON OF TRAINING STRATEGIES FOR DOMAIN TASKS

metrics	open-ended QA B-4	open-ended QA R-1	info. extraction B-4	info. extraction R-1	classification Accuracy
GLM-4	7.77	24.56	24.57	47.55	53.50
baseline	8.29	24.11	19.24	42.37	8.50
SFT	12.73	30.05	77.14	85.54	57.00
DAP+SFT (ours)	**13.46**	**30.51**	**79.81**	**87.29**	**58.50**

3. Improved DAP combats catastrophic forgetting

TABLE II
COMPARISON OF TRAINING STRATEGIES FOR GENERAL TASKS

metrics	alpaca_zh B-4	alpaca_zh R-1	suolyer_webqa B-4	suolyer_webqa R-1
GLM-4	12.09	32.02	8.51	25.31
base	14.05	34.55	7.49	26.10
SFT	18.67	39.77	**10.35**	28.72
DAP+SFT (ours)	**18.68**	**40.02**	10.00	**29.51**

TABLE III
COMPARISON OF MODEL PERFORMANCE ON CMMLU BENCHMARK UNDER VARIOUS DAP STRATEGIES

	STEM	Social Sci.	Humanities	Other	Avg.
DAP-d	60.09	71.77	74.97	72.96	70.66
DAP (ours)	**62.90**	**72.95**	**75.21**	**74.02**	**71.51**

4. Low-quality data can also boost model performance via preference alignment using KTO

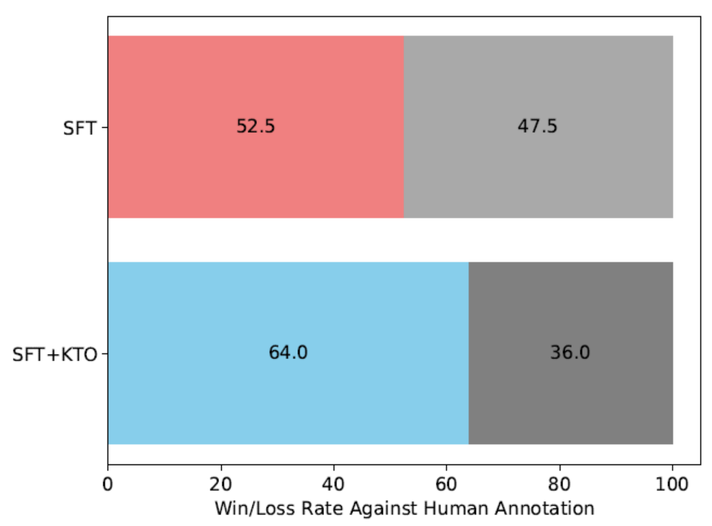

Fig. 1. Win (Left) and Loss (Right) rate for different models