KaFT: Knowledge-aware Fine-tuning for Boosting LLMs' Domain-specific Question-Answering Performance

Anonymous ACL submission

Abstract

Supervised fine-tuning (SFT) is a common approach to improve the domain-specific question-answering (QA) performance of large language models (LLMs). However, recent literature reveals that due to the conflicts between LLMs' internal knowledge and the context knowledge of training data, vanilla SFT using the full QA training set is usually sub-optimal. In this paper, we first design a query diversification strategy for robust conflict detection and then conduct a series of experiments to analyze the impact of knowledge conflict. We find that 1) training samples with varied conflicts contribute differently, where SFT on the data with large conflicts leads to catastrophic performance drops; 2) compared to directly filtering out the conflict data, appropriately applying the conflict data would be more beneficial. Motivated by this, we propose a simple-yet-effective Knowledge-aware Fine-tuning (namely KaFT) approach to effectively boost LLMs' performance. The core of KaFT is to adapt the training weight by assigning different rewards for different training samples according to conflict level. Extensive experiments show that KaFT brings consistent and significant improvements (up to **+5.73%** average scores) across four LLMs. More analyses prove that KaFT effectively improves the model generalization and alleviates the hallucination.

1 Introduction

While large language models (LLMs) (OpenAI, 2023; Dubey et al., 2024; Zhao et al., 2023) have showcased powerful general-purpose capabilities, they often struggle to handle domain-specific question-answering (QA) tasks, *e.g.*, medical QA (Labrak et al., 2024). Hence, supervised fine-tuning (SFT), aiming to activate LLMs' internal knowledge and align LLMs' output with the desired behavioral norms, is usually required (Zhou et al., 2024; Zhang et al., 2024).

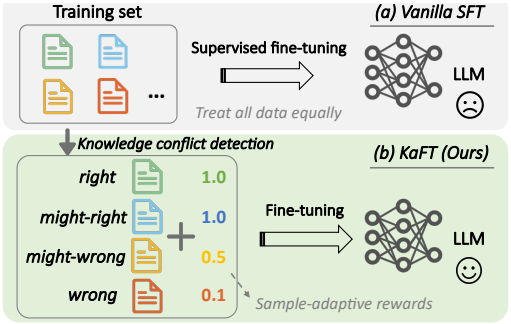

Figure 1: Comparison between **(a) vanilla SFT** and **(b) our KaFT**. Different from vanilla SFT treating all training data equally, KaFT uses sample-adaptive rewards to facilitate more effective learning of LLMs.

However, recent literature (Ren et al., 2024; Gekhman et al., 2024) reveals that domain-specific SFT usually suffers from a crucial problem: ***knowledge conflict***, which is the discrepancy between the LLMs' internal knowledge and the context knowledge of training data (Xu et al., 2024). Due to the long-tail distribution and timeliness of pretraining corpora, LLMs might struggle to learn sufficient domain-specific knowledge during pretraining. Conversely, SFT training datasets usually contain more up-to-date and professional knowledge. Unfortunately, SFT fails to learn additional knowledge (Ren et al., 2024), and enforcing LLMs to learn new knowledge through SFT would easily damage their prior abilities and lead to hallucination (Gekhman et al., 2024).

To tackle this problem, some empirical studies have been conducted (Ren et al., 2024; Gekhman et al., 2024; Ye et al., 2024). For instance, Ren et al. (2024) employ in-context learning (ICL) (Brown et al., 2020) to probe LLMs' internal knowledge and determine whether it conflicts with the training data. Based on this, they analyze the behavior of LLMs after SFT with conflict data. Despite providing some insightful findings, they still have some shortcomings: 1) the proposed conflict detection

methods are simply based on ICL, which is sensitive to few-shot examples and might introduce bias into the results (Min et al., 2022; Ye et al., 2024); 2) they alleviate the negative effect of knowledge conflict by directly filtering the conflict data, while neglecting how to make full use of these data.

To this end, we first improve the ICL-based conflict detection with a *query diversification* strategy to reduce the bias of few-shot examples. Based on it, we conduct a series of preliminary analyses to reveal the impact of knowledge conflict. Specifically, we calculate the conflict score for each training data and split the training set evenly into four subsets with varied conflicts. By fine-tuning LLMs with different subsets, we find that:

- Different subsets contribute differently, where SFT on the individual subset with more conflicts causes catastrophic performance drops.
- Compared to directly filtering the subset with more conflicts, appropriately applying these data might be more beneficial.

Based on these observations, we recognize that *not all training samples are equally important for SFT*, and LLMs should pay different attention to different samples. Motivated by this, we proposed a simple-yet-effective **K**nowledge-**a**ware **F**ine-**T**uning (namely **KaFT**) approach to effectively boost LLMs' QA performance. As illustrated in Figure 1, the core of KaFT is to assign different rewards to varied subsets and use these rewards to adapt the learning of LLMs. Specifically, for the data with more conflicts, KaFT assigns a smaller reward to alleviate its negative effect. Conversely, for the data with fewer conflicts, KaFT uses a larger reward to encourage its learning. By doing so, KaFT can not only avoid overfitting to conflict data, but also effectively activate its internal knowledge for more efficient domain adaptation.

We mainly evaluate our KaFT in the medical QA applications upon four popular LLMs, including LLaMA3-8B/3B (Dubey et al., 2024), Qwen1.5-7B (Bai et al., 2023), and Mistral-7B (Jiang et al., 2023). Extensive results show that KaFT surpasses the other baselines by a clear margin, and brings consistent and significant performance gains (up to **+5.73%** average scores) across all LLMs. More in-depth analysis prove that KaFT can be expanded to other domain-specific applications. More encouragingly, KaFT improves the model generalization and alleviates the hallucination effectively.

Contributions. To summarize, our contributions are three-fold: (1) We propose a query diversification strategy for robust conflict detection. Based on it, we conduct a series of preliminary analyses and reveal that training samples with varied conflicts contribute differently. (2) Motivated by this, we propose a simple-yet-effective knowledge-aware SFT (KaFT) approach, which employs sample-adaptive rewards to boost LLMs' QA performance. (3) Extensive experiments show that KaFT outperforms the vanilla SFT by a clear margin and improves the model generalization effectively.

2 Preliminary

2.1 Task Formulation

Given a domain-specific QA training dataset $D = \{(q_i, o_i, a_i)\}_{i=1}^{N}$ and a pretrained base LLM $\mathcal{M}_\theta$ parameterized by θ, where q_i, o_i and a_i denote the question, options and answer, and N denotes the number of all training samples. The goal of SFT is to use the $\mathcal{D}$ to fine-tune $\mathcal{M}_\theta$ with supervised learning, *i.e.*, maximum likelihood estimates, and obtain the final adapted LLM $\mathcal{M}_{\theta^*}$.

2.2 Knowledge Conflict Detection with Query Diversification Strategy

As mentioned in §1, SFT usually suffers from the knowledge conflict problem. To detect the knowledge conflicts in $\mathcal{D}$, Ren et al. (2024) propose an ICL-based probing method. Specifically, they randomly select some training samples as the few-shot examples and utilize them to probe $\mathcal{M}_\theta$'s response r_i with greedy decoding, *i.e.*, temperature=0, to each query (q_i, o_i). The response r_i is referred to as the model's parameter knowledge for this question q_i. Then, they determine whether the r_i is aligned with reference answer a_i, *i.e.*, $\mathbb{I}(r_i = a_i)$, where $\mathbb{I}(\cdot)$ is the indicator function, and regard the misaligned samples as the conflict data. To have a closer look, we provide a case in Appendix A.3.

Obviously, such a simple ICL-based approach is not robust, as it is sensitive to the few-shot examples and introduce bias. To this end, we improve this method with a *query diversification* strategy. The primary intuition of our strategy is that, if we replace the internal order of options o_i and $\mathcal{M}_\theta$ always fails to output the correct answer, $\mathcal{M}_\theta$ indeed does not learn the knowledge for the question. In practice, for each data point, we first replace the internal order of o_i and obtain N_o different queries $\{(q_i, o_i^j)\}_{j=1}^{N_o}$. Then, we feed the queries

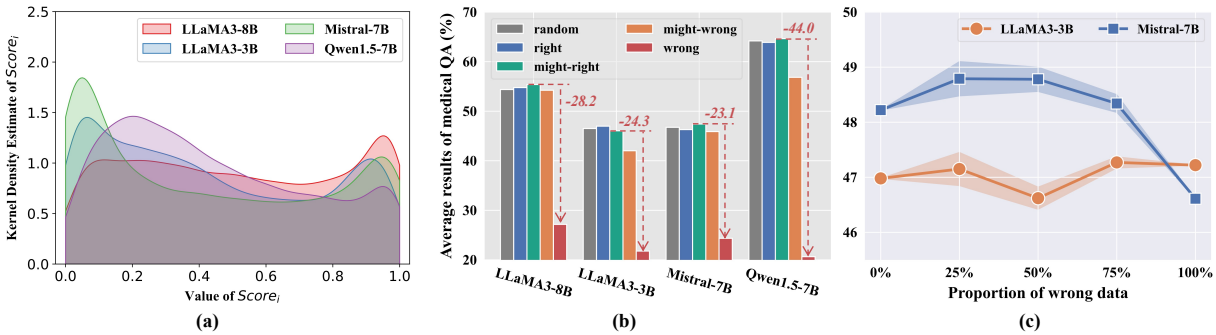

Figure 2: **(a) Illustration of distributions of $Score_i$ on MedQA across different LLMs**. We use the kernel density estimate for visualizing, where the larger density refers to more training samples. **(b) Performance comparison (%) of different subsets**. Note that all subsets hold the same number of training samples. **(c) Analysis of different proportions of `wrong` data**. Specifically, we randomly select varied samples from `wrong` and merge them with the other three subsets. We use three different random seeds for data sampling and report the average results.

into $\mathcal{M}_\theta$ to obtain its responses. Moreover, inspired by self-consistency (Wang et al., 2023), we set the temperature to 0.7 and sample N_r candidate responses $\{r_{ik}^j\}_{k=1}^{N_r}$ for each query (q_i, o_i^j). Lastly, the knowledge conflict can be measured as:

$$Score_i = \frac{\sum_{j=1}^{N_o}\sum_{k=1}^{N_r}\mathbb{I}(r_{ik}^j) = a_i)}{N_o \times N_r}, \quad (1)$$

where $Score_i$ denotes the conflict score (larger value refers to fewer conflicts) of i-th training data.

2.3 Empirical Analyses

Setting. We use a popular medical QA benchmark, *i.e.*, MedQA (Jin et al., 2021), as the testbed, containing 10,178 training data. We perform SFT on four cutting-edge LLMs, including LLaMA3-8B/3B (Dubey et al., 2024), Qwen1.5-7B (Bai et al., 2023), and Mistral-7B (Jiang et al., 2023). The tuned models are evaluated on six medical QA benchmarks, covering the test sets of MedQA, MedMCQA (Pal et al., 2022), MMLU* (Hendrycks et al., 2020)[1]), CMB (Wang et al., 2024b), CMExam (Liu et al., 2024b), and CMMLU* (Li et al., 2024). For conflict detection, we set the N_o and N_r to 10. The distributions of $Score$ are illustrated in Figure 2 **(a)**.

Findings. To investigate the impact of knowledge conflict, we conduct systematic analyses and empirically observe that:

❶ Different subsets contribute differently, where SFT on the individual subset with more conflicts causes catastrophic performance drops. First, we calculate the $Score$ for each training data and sort the $\mathcal{D}$ based on the score. Then, we split $\mathcal{D}$ evenly into four subsets with varied conflicts, denoted as `right`, `might-right`, `might-wrong` and `wrong`, where `right` has less conflicts and `wrong` has most conflicts. Notably, these subsets have the same number of training samples. We fine-tune the LLMs using different individual subsets and illustrate the comparative results in Figure 2 **(b)**. For reference, we also present the results of SFT on the randomly selected samples. As seen, LLMs tuned with different subsets perform differently. Similar to prior findings (Ren et al., 2024), SFT on the `wrong` leads to catastrophic performance drops, proving the negative effect of knowledge conflict. More interestingly, `right` is usually not the optimal subset, while `might-right` performs better among all LLMs. We conjecture that many `right` samples have been learned by LLMs and struggle to provide useful information. Conversely, `might-right` can help activate LLMs' internal knowledge and better boost their performance.

❷ Compared to directly filtering the subset with more conflicts, appropriately applying these data might be more beneficial. Intuitively, a straightforward way for alleviating the negative effect of knowledge conflict is to filter out the `wrong` subset from $\mathcal{D}$ and use the other subsets for SFT. However, as aforementioned, appropriate conflict data might help activate LLMs' internal knowledge and lead to better performance. To verify our conjecture, we introduce different proportions λ of conflict data from `wrong` into the collection of other less-conflict subsets, where λ ranges from

[1]Following Singhal et al. (2025), we select six medical subtasks from MMLU, and denote this subset as MMLU*. Similarly, we also collect the medical sub-tasks from CMMLU (Li et al., 2024) and denote it as CMMLU*.

0% to 100%. The results are illustrated in Figure 2 **(c)**, from which we find that compared to directly filtering the `wrong` (*i.e.*, $\lambda = 0\%$), introducing some conflict data (*e.g.*, $\lambda = 50\%$) could be more beneficial. This highlights the necessity of exploring more effective SFT methods to make full use of the conflict data.

3 Knowledge-aware Fine-tuning

Based on the observations in §2.3, we recognize that *not all training samples are equally important for SFT*, and LLMs should pay different attention to different samples. To this end, we propose a knowledge-aware fine-tuning (KaFT) approach to alleviate the negative effect of knowledge conflict and boost LLMs' performance. In this section, we introduce our KaFT in detail.

Motivation and Intuition. In addition to the empirical findings in §2.3, our KaFT is also inspired by a famous cognitive structure migration theory (Ausubel et al., 1978), *i.e.*, "The most important single factor influencing learning is what the student already knows", which highlights that *paying more attention to the new content relevant to prior learned knowledge can lead to more effective knowledge transfer*. Intuitively, for the data with more conflicts, *e.g.*, `wrong`, LLMs might easily over-fit the unfamiliar knowledge and lead to poor generalization. In contrast, for data with fewer conflicts, more in-depth learning is beneficial for transferring LLMs' internal knowledge and facilitating effective domain adaptation.

Implementation of KaFT. In practice, based on our proposed strategy in §2.2, we first calculate the conflict score $Score_i$ for each training data (q_i, o_i, a_i), and split $\mathcal{D}$ evenly into four subsets with varied conflicts, as done in §2.3. Then, we assign different rewards for different subsets, where `might-right` and `right` hold the larger rewards, and the `wrong` and `might-wrong` hold the smaller rewards. Lastly, the rewards are used to control the learning weights of different subsets. The learning objective can be formulated as:

$$R_i = \begin{cases} \alpha, & \text{if}(q_i, o_i, a_i) \in \texttt{wrong}, \\ \beta, & \text{if}(q_i, o_i, a_i) \in \texttt{might-wrong}, \\ 1, & \text{if}(q_i, o_i, a_i) \in \texttt{might-right}, \\ 1, & \text{if}(q_i, o_i, a_i) \in \texttt{right}, \end{cases}$$

$$\theta^* := \arg\min \mathbb{E}_{(q,o,a,R)\sim\mathcal{D}}[R \ \log \mathcal{M}(a|q,o)], \quad (2)$$

where R_i denotes the reward for i-th training data and θ^* denotes the parameters of final LLM $\mathcal{M}_{\theta^*}$. α and β are rewards between 0 and 1, where α is generally smaller than β. Empirically, we set α and β as 0.1 and 0.5, respectively.

4 Experiments

4.1 Setup

Tasks and Datasets. Similar to the settings of §2.3, we mainly apply our KaFT in the medical QA and fine-tune LLMs with the training set of MedQA. The tuned models are evaluated on six in-domain test sets, covering English medical QA (MedQA, MedMCQA, MMLU*) and Chinese medical QA (CMB, CMExam and CMMLU*). Moreover, we follow Ren et al. (2024) and use the constructed QA test sets from three domains: history, engineering and law, as the out-of-domain (OOD) benchmarks. For evaluation, we use the public `lm-evaluation-harness` toolkit and report the zero-shot accuracy for each benchmark. The details of all tasks are shown in Appendix A.1.

Models. We conduct extensive experiments on four cutting-edge LLMs across different model architectures and sizes, *i.e.*, LLaMA3-8B/3B (Dubey et al., 2024), Qwen1.5-7B (Bai et al., 2023), and Mistral-7B (Jiang et al., 2023). In the implementation of KaFT, the N_o and N_r are set to 10. We train each model with a batch size of 16 and a peak learning rate of 1e-4, except 2e-4 for LLaMA3-3B. All models are trained with the LoRA (Hu et al., 2021) for 1 epoch. The details of model training and inference can be found in Appendix A.2.

Baselines. We compare KaFT with a series of baselines: 1) ***Base*** denotes the original LLMs without SFT, 2) ***Vanilla SFT*** denotes directly fine-tuning LLMs on the full training set equally, 3) ***No-conflict*** denotes first removing the conflict data (`wrong` identified in §2.3) and then fine-tuning LLMs on the remaining training data, and 4) ***Self-aligning***, inspired by Ren et al. (2024), denotes first modifying the answers of `wrong` to match LLM's internal knowledge (*i.e.*, replacing a_i with r_i) and then fine-tuning LLMs on the combination of aligned `wrong` and the other original subsets.

4.2 Compared Results

The main results on medical QA benchmarks and OOD benchmarks are reported in Tables 1 and 2, respectively. From these results, we can find that:

Backbone	Method	English Medical Benchmark			Chinese Medical Benchmark			Score	
		MedQA	MedMCQA	MMLU*	CMB	CMExam	CMMLU*	Avg.	Δ ↑
Mistral-7B	Base	50.98	48.31	65.07	31.73	30.68	32.90	43.28	-
	Vanilla SFT	**59.86**	43.75	68.06	36.25	36.25	35.48	46.61	+3.33
	No-conflict	58.37	**51.11**	68.69	38.09	36.92	36.17	48.22	+4.94
	Self-aligning	55.62	50.99	**68.81**	36.70	36.81	35.13	47.34	+4.06
	KaFT (Ours)	59.54	49.87	68.47	**38.11**	**37.35**	**40.73**	**49.01**	**+5.73**
Qwen1.5-7B	Base	48.94	50.08	62.79	74.77	76.59	70.42	63.93	-
	Vanilla SFT	52.71	50.20	61.60	74.30	76.15	70.16	64.19	+0.26
	No-conflict	52.24	**50.54**	61.55	75.04	76.87	70.77	64.50	+0.57
	Self-aligning	51.77	50.11	63.18	75.23	77.05	70.71	64.67	+0.74
	KaFT (Ours)	**53.57**	49.82	**63.23**	**75.57**	**77.27**	**72.07**	**65.25**	**+1.32**
LLaMA3-8B	Base	59.62	56.51	72.66	45.50	46.04	44.62	54.16	-
	Vanilla SFT	61.82	55.75	73.34	45.99	45.85	44.95	54.62	+0.46
	No-conflict	61.98	56.11	73.40	47.17	47.72	45.59	55.33	+1.17
	Self-aligning	61.35	56.56	73.00	**47.53**	**48.34**	**46.55**	55.55	+1.39
	KaFT (Ours)	**64.10**	**56.94**	**74.01**	47.39	47.89	45.44	**55.96**	**+1.80**
LLaMA3-3B	Base	51.14	49.41	62.34	35.98	36.48	36.50	45.31	-
	Vanilla SFT	54.99	50.08	62.94	38.00	39.17	38.13	47.22	+1.91
	No-conflict	52.95	49.20	63.76	38.91	39.78	37.27	46.98	+1.67
	Self-aligning	52.79	49.82	62.93	38.53	38.66	38.55	46.88	+1.57
	KaFT (Ours)	**54.52**	**50.51**	**64.93**	**40.19**	**39.93**	**39.70**	**48.30**	**+2.99**

Table 1: **Performance comparison (%) on the medical QA benchmarks.** "Avg." denotes the average results, and "Δ ↑" refers to the gains against the base models. Best results are in **bold**, and second-best results are underlined.

KaFT surpasses the other baselines by a clear margin. As seen, compared to the vanilla SFT, "No-conflict" usually achieves better performance, highlighting the harmfulness of conflict data. "Self-aligning" can sometimes bring further performance gains against "No-conflict", *e.g.*, +0.22% average score in LLaMA3-8B. However, it might lead to worse performance in some cases. One possible reason is that r_i obtained by the method in (Ren et al., 2024) can not probe LLMs' internal knowledge well, thus leading to some noise. Conversely, our KaFT surpasses the other counterparts by a clear margin, proving its superiority.

KaFT brings consistent and significant performance gains among all model sizes and types. We see that KaFT not only achieves remarkable performance for LLaMA3-family models, but is also beneficial to the Qwen and Mistral models. Specifically, compared to the base models, KaFT brings **+5.73%**, **+1.32%**, **+1.80%** and **+2.99%** average gains for Mistral-7B, Qwen1.5-7B and LLaMA3-8B/3B, respectively. These results prove the effectiveness and universality of KaFT.

KaFT effectively improves the OOD performance. Additionally, we evaluate the tuned LLMs on the OOD benchmarks to verify LLMs' robustness. Due to space limitations, we only present the contrastive results of Mistral-7B and LLaMA3-3B models in Table 2. From the table, we observe that KaFT significantly outperforms the baselines among all domains, indicating that alleviating the negative effect of conflict data can avoid the overfitting of LLMs, continuing to prove the effectiveness of our proposed KaFT approach.

4.3 Ablation Study

Here, we investigate 1) the effect of different conflict detection methods, 2) the influence of different reward strategies in KaFT, and 3) the analyses of hyper-parameters in KaFT.

Effect of conflict detection methods. One of our contributions is to design a query diversification strategy for robust conflict detection. Here, to verify its effectiveness, we compare it with some variants: 1) "-w/o diverse query" means removing the query diversification process and obtaining multiple responses for the original query. 2) "-w/o response sampling" means using greedy decoding to obtain the model responses with the highest probability for diverse queries, respectively. 3) "-w/o both" means removing both processes and directly using greedy decoding to obtain the model response for each original query, as done in Ren et al. (2024). After obtaining the responses, we compared them

Method	Mistral-7B					LLaMA3-3B				
	History	Engineering	Law	Avg.	Δ ↑	History	Engineering	Law	Avg.	Δ ↑
Base	41.20	53.20	46.80	47.07	-	33.60	54.80	38.40	42.27	-
Vanilla SFT	46.00	59.20	50.00	51.73	+4.66	40.00	56.40	46.40	47.60	+5.33
No-conflict	45.20	**61.20**	49.60	52.00	+4.93	**40.80**	58.00	48.00	48.93	+6.66
Self-aligning	44.40	56.80	49.20	50.13	+3.06	39.60	56.40	48.00	48.00	+5.73
KaFT (Ours)	**50.40**	60.00	**51.60**	**54.00**	**+6.93**	40.40	**58.40**	**49.60**	**49.47**	**+7.20**

Table 2: **Performance comparison (%) of tuned medical LLMs on the out-of-domain QA test sets**. "Avg." denotes the average performance. Best results are in **bold**, and second-best results are underlined.

Method	Score	Δ ↓
Random	54.38	-
Ours	**27.16**	**↓ 27.22**
-w/o diverse query	38.96	↓ 15.42
-w/o response sampling	30.15	↓ 24.23
-w/o both	49.00	↓ 5.38

Table 3: **Performance comparison (%) of `wrong` sets selected by different conflict detection methods**. The LLaMA3-8B is used as the base model. "Δ ↓" denotes the performance drops against the random selection, where larger values refer to better performance.

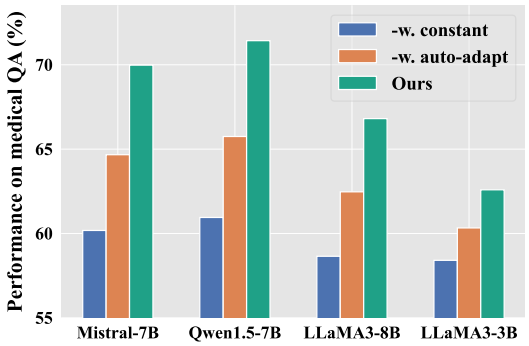

Figure 3: **Effect of reward strategies in KaFT**. The y-axis denotes the average performance of medical QA.

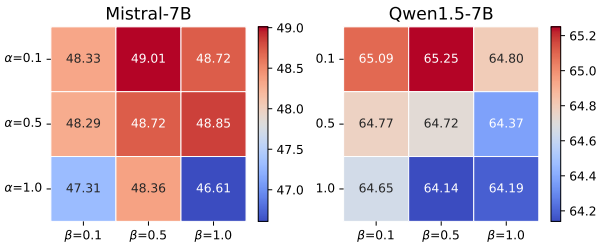

Figure 4: **Parameter analyses of KaFT**. The y-axis and x-axis denote the varied α and β, respectively. We report the average results on medical QA benchmarks.

with the references to calculate the conflict score. Based on it, we sort the training data and select the `wrong` set. Taking the LLaMA3-8B as an example, we present the medical QA results of models tuned with different `wrong` sets in Table 3. As seen, the `wrong` selected by our method leads to maximum performance degradation, *i.e.*, our method can effectively detect the conflict data and select the most conflict subset, proving its effectiveness.

Effect of reward strategies in KaFT. As mentioned in §3, KaFT empirically assigns the rewards for subsets with varied conflicts. In this part, we investigate this strategy by comparing it with two variants: 1) "-w. constant" refers to the constant reward for all subsets, *i.e.*, $R_i = 1.0$, and 2) "-w. auto-adapt" refers to using the conflict scores as the rewards, *i.e.*, $R_i = Score_i$. Comparative results of medical QA are illustrated in Figure 3. Both of ours and "-w. auto-adapt" outperform the "-w. constant" by a clear margin, proving the effectiveness of knowledge-aware SFT. Moreover, "-w. auto-adapt" usually performs worse than ours. One possible reason is that it assigns a relatively small reward for the more important `might-right` subset, thus hindering the activation of LLMs' internal knowledge. Conversely, our strategy can make full use of the training data and achieve the best performance.

Parameter Analysis. In Eq. 2, we use two hyperparameters, *i.e.*, α and β, to control the rewards for `wrong` and `might-wrong` subsets. In this study, we analyze their influence by evaluating the performance of KaFT with different α and β, spanning $\{0.1, 0.5, 1.0\}$. Figure 4 illustrates the average results of Mistral-7B and Qwen1.5-7B on medical QA benchmarks, from which we find that: 1) Increasing the α leads to a continuous performance decline, confirming the motivation to suppress the learning of conflict data. 2) Increasing the β appropriately brings better performance, but too large β (*i.e.*, 1.0) is harmful. We conjecture that the `might-wrong` could contain some conflict data, and overemphasizing its learning would cause over-

Method	Mistral-7B					LLaMA3-8B				
	QA	Dialogue	Sumarization	Avg.	Δ ↑	QA	Dialogue	Sumarization	Avg.	Δ ↑
Base	51.65	62.22	44.50	52.79	-	50.79	67.27	**49.49**	55.85	-
`Wrong`-only	48.92	56.71	44.40	50.01	-2.78	45.55	51.73	42.56	46.61	-9.24
Vanilla SFT	50.66	69.60	**49.24**	56.50	+3.71	45.43	72.04	48.77	55.41	-0.44
No-conflict	**54.54**	70.48	45.89	56.97	+4.18	49.16	71.98	48.55	56.56	+0.71
Self-aligning	54.22	72.41	46.11	57.58	+4.79	49.67	72.24	48.64	56.85	+1.00
KaFT (Ours)	54.20	**73.57**	47.68	**58.48**	**+5.69**	**50.85**	**72.65**	48.79	**57.43**	**+1.58**

Table 4: **Performance comparison (%) on the hallucination evaluation, *i.e.*, HaluEval** (Li et al., 2023a). **Green** and **red** results refer to the average performance gains and drops against the "Base" baseline, respectively. For references, we also report the results of "`Wrong`-only", which fine-tunes LLMs on the individual `wrong` subset.

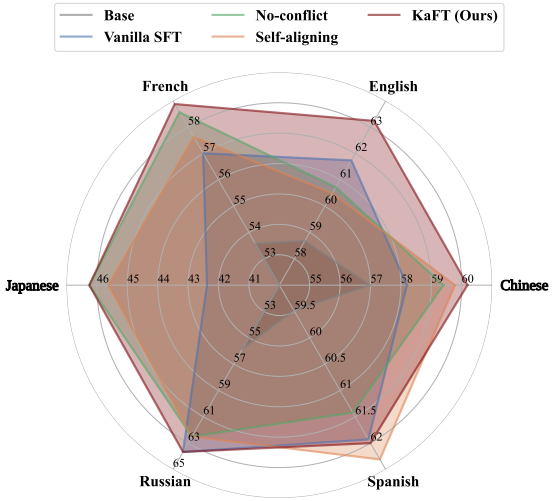

Figure 5: **Performance comparison (%) on multilingual medical QA**. LLaMA3-8B is used as base model.

fitting. More specifically, the case of $\alpha = 0.1$ and $\beta = 0.5$ performs best, thus leaving as our default experimental settings.

5 Discussion

Here, we conduct further analyses to discuss: 1) whether it gains better model generalization, and 2) whether KaFT still works in other scenarios.

5.1 Does KaFT improve the generalization?

Intuitively, by alleviating the negative effect of conflict data, KaFT can achieve better model generalization. To verify it, we further analyze its effect from the following aspects:

Multilingual Generalization. We evaluate the tuned models on a popular multilingual medical QA benchmark, *i.e.*, MMedBench (Qiu et al., 2024), consisting of six languages: Chinese, English, French, Japanese, Russian, and Spanish. The comparative results of tuned LLaMA3-8B models are illustrated in Figure 5. As seen, our KaFT brings better performance gains than the other methods across most languages. Specifically, compared to the base model, KaFT achieves **+4.94%** average performance gains, especially **+6.25%** gains in Japanese and **+7.81%** gains in Russian.

Hallucination Alleviation. As stated in the prior work (Gekhman et al., 2024), fine-tuning with conflict data increases the LLMs' tendency to hallucinate. Here, we investigate this problem by evaluating the tuned models on a popular hallucination detection benchmark, HaluEval (Li et al., 2023a). Specifically, the models are required to classify whether a sample contains hallucinated contents from three tasks, *i.e.*, question answering (QA), knowledge-grounded dialogue (Dialogue), and text summarization (Summarization). The results of Mistral-7B and LLaMA3-8B models are reported in Table 4. For references, we also report the results of directly fine-tuning on the `wrong` subset, denoted as "`Wrong`-only". It can be found that enforcing LLMs to learn the new knowledge from conflict data indeed causes serious hallucination, as "`Wrong`-only" and "Vanilla SFT" cause up to **-9.24%** and **-0.44%** average score drops, respectively. More encouragingly, our KaFT can effectively alleviate this side effect and bring up to **+5.69%** average score gains against base models. Takeaway: ***These results prove that our KaFT can indeed bring better model generalization.***

5.2 Does KaFT still work in other scenarios?

Although our KaFT is mainly evaluated in the medical domain, we believe that it has great potential to expand to more domain-specific applications. To verify it, we conduct additional experiments from three domains: history, engineering, and law. Following Ren et al. (2024), we use the corresponding domain-specific training and test sets, collected from Xiezhi Benchmark (Gu et al., 2024). The

Method	History →				Engineering →				Law →			
	History	Engineering	Law	Avg.	History	Engineering	Law	Avg.	History	Engineering	Law	Avg.
Base	49.60	59.20	51.60	53.47	49.60	59.20	51.60	53.47	49.60	59.20	51.60	53.47
Vanilla SFT	64.40	66.00	67.20	65.87	57.60	65.20	**56.80**	59.87	64.40	64.40	60.80	63.20
No-conflict	60.00	**66.80**	65.20	64.00	56.80	65.20	56.00	59.33	61.60	64.00	60.00	61.87
Self-aligning	56.80	66.00	60.40	61.07	52.80	63.60	55.20	57.20	57.60	64.40	58.00	60.00
KaFT (Ours)	**66.40**	66.00	**67.60**	**66.67**	**58.80**	**67.20**	**56.80**	**60.93**	**66.00**	**65.60**	**61.20**	**64.27**

Table 5: **Performance comparison (%) on more domain-specific QA applications.** Notably, we fine-tune the LLaMA3-8B with the individual domain-specific training set (*i.e.*, History, Engineering, and Law) and evaluate them on all domains' test sets. "Avg." denotes the average performance, and the best results are in **bold**.

data statistics are provided in Appendix A.1. We fine-tune the LLMs with the individual domain-specific training set and evaluate them on the test sets of all domains. Results of tuned LLaMA3-8B models are reported in Table 5, from which we observe that KaFT performs best and brings consistent and significant performance gains among all domains. Takeaway: ***KaFT not only works well in medical QA, but also can be applied to more domain-specific scenarios***.

6 Related Works

LLMs (Ouyang et al., 2022; OpenAI, 2023; Dubey et al., 2024; Liu et al., 2024a) have showcased powerful general-purpose capabilities. However, they might fall short in domain-specific applications, such as medical QA (Labrak et al., 2024). To this end, many prior works (Singhal et al., 2023; Li et al., 2023b; Chen et al., 2023; He et al., 2025) attempt to perform SFT on the domain-specific QA dataset for facilitating domain adaptation.

Despite achieving remarkable performance, SFT often faces a critical challenge, *i.e.*, knowledge conflicts. Specifically, since domain-specific SFT is more knowledge-intensive and contains rich professional knowledge that has not been learned during the LLMs' pretraining, there is usually a discrepancy between the LLMs' internal knowledge and the context knowledge of the SFT corpus. More recently, Ren et al. (2024) reveal that SFT fails to learn additional knowledge and Gekhman et al. (2024) find that enforcing LLMs to learn new knowledge through SFT would easily damage their prior abilities and lead to hallucination. Thus, it is suboptimal to directly fine-tune LLMs using the full SFT training samples equally.

To address this problem, there are few existing works (Ren et al., 2024; Gekhman et al., 2024; Ye et al., 2024). However, they still have some shortcomings and struggle to tackle this problem effectively. On the one hand, their conflict detection methods highly rely on ICL (Brown et al., 2020), which is sensitive to the few-shot examples (Min et al., 2022). On the other hand, after detecting the conflict data, they mitigate its negative effect by either using early-stopping or filtering out it from the training dataset, while neglecting how to make full use of these conflict data.

Different from these prior studies, we first design a query diversification strategy to robustly detect the conflict and then propose a knowledge-aware SFT (KaFT) to make full use of all training data. The main idea of KaFT is to use sample-adaptive rewards for better guiding the learning of LLMs, which is somewhat similar to prior adaptive-learning methods (Wang et al., 2024a; Li et al., 2020). The way to obtain the sample-adaptive rewards is innovative, and we believe that our approach has great potential to unleash the power of LLMs in real-world applications.

7 Conclusion

In this paper, we focus on the knowledge conflict problem in the domain-specific SFT, which is critical yet under-explored. Specifically, we propose a query diversification strategy to robustly detect the conflict. Based on it, we conduct a series of preliminary analyses and reveal that different training samples contribute differently, where those with more conflicts would dynamically damage LLMs' abilities. To this end, we further propose a knowledge-aware SFT approach (KaFT). In short, KaFT utilizes sample-adaptive rewards to suppress the negative effect of conflict data and encourage LLMs to activate more relevant knowledge. Extensive results on medical QA benchmarks demonstrate the effectiveness and universality of KaFT. More encouragingly, in-depth analyses prove that KaFT can achieve better model generalization and alleviate the model hallucination effectively.

Limitations

Our work has several potential limitations. First, given the limited computational budget, we only validate our KaFT on up to 8B LLMs in the main experiments. It will be more convincing if scaling up to super-large model sizes (*e.g.*, 70B) and applying KaFT to more cutting-edge model architectures. On the other hand, to better probe LLMs' internal knowledge, we follow the prior studies (Ren et al., 2024; Ye et al., 2024) and mainly focus on multiple-choice QA tasks. We will expand our methods to the long-form QA scenarios in future work.

Ethics and Reproducibility Statements

Ethics We take ethical considerations very seriously and strictly adhere to the ACL Ethics Policy. This paper proposes a knowledge-aware fine-tuning framework to improve the domain-specific QA performance of LLMs. It aims to activate LLMs' internal domain-specific knowledge, *e.g.*, medical, instead of encouraging them to learn privacy knowledge that may cause an ethical problem. Moreover, all training and evaluation datasets used in this paper are publicly available and have been widely adopted by researchers. Thus, we believe that this research will not pose ethical issues.

Reproducibility In this paper, we discuss the detailed experimental setup, such as training hyperparameters and statistic descriptions. More importantly, ***we have provided our code and data in the Supplementary Material*** to help reproduce the experimental results of this paper.

References

David Paul Ausubel, Joseph Donald Novak, Helen Hanesian, et al. 1978. Educational psychology: A cognitive view.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, et al. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. 2020. Boosting few-shot learning with adaptive margin loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024b. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. In *Advances in Neural Information Processing Systems*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*.

Mengjie Ren, Boxi Cao, Hongyu Lin, Cao Liu, Xianpei Han, Ke Zeng, Guanglu Wan, Xunliang Cai, and Le Sun. 2024. Learning or self-aligning? rethinking instruction fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024a. Openchat: Advancing open-source language models with mixed-quality data. In *The Twelfth International Conference on Learning Representations*.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, et al. 2024b. Cmb: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Junjie Ye, Yuming Yang, Qi Zhang, Tao Gui, Xuanjing Huang, Peng Wang, Zhongchao Shi, and Jianping Fan. 2024. Empirical insights on fine-tuning large language models for question-answering. *arXiv preprint arXiv:2409.15825*.

Liang Zhang, Katherine Jijo, Spurthi Setty, Eden Chung, Fatima Javid, Natan Vidra, and Tommy Clifford. 2024. Enhancing large language model performance to answer questions and extract information more accurately. *arXiv preprint arXiv:2402.01722*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*.

Dataset	#Training	#Test
Medical QA		
MedQA (Jin et al., 2021)	10,178	1,273
MedMCQA (Pal et al., 2022)	-	4,183
MMLU* (Hendrycks et al., 2020)	-	1,089
- Anatomy	-	135
- Clinical Knowledge	-	265
- College Biology	-	144
- College Medicine	-	173
- Medical Genetics	-	100
- Professional Medicine	-	272
CMB (Wang et al., 2024b)	-	9,998
CMExam (Liu et al., 2024b)	-	6,607
CMMLU* (Li et al., 2024)	-	1,140
- Anatomy	-	148
- Clinical Knowledge	-	237
- College Medical Statistics	-	106
- College Medicine	-	273
- Professional Medicine	-	376
Other domain-specific QA		
History (Gu et al., 2024)	8,605	250
Engineering (Gu et al., 2024)	4,805	250
Law (Gu et al., 2024)	6,510	250
More in-depth analyses		
MMedBench (Qiu et al., 2024)	-	8,178
- Chinese	-	3,426
- English	-	1,273
- French	-	321
- Japanese	-	160
- Russian	-	256
- Spanish	-	2,742
HaluEval (Li et al., 2023a)	-	30,000
- question answering	-	10,000
- knowledge-grounded dialogue	-	10,000
- text summarization	-	10,000

Table 6: **Statistic information** of all used datasets in our study. "#Training" and "#Test" denote the number of training and test samples, respectively.

A Appendix

A.1 Details of Tasks and Datasets

In this work, to investigate the effectiveness and universality of our KaFT, we conduct extensive experiments on four domain-specific QA applications, covering medical, history, and law. In addition, the multilingual medical QA tasks and hallucination detection tasks are used to reveal the underlying mechanism of our method. Here, we introduce the descriptions of these tasks and datasets in detail. First, we present the statistics of all datasets in Table 6. Then, each task is described as:

MedQA. MedQA (Jin et al., 2021) consists of questions and corresponding 4-option or 5-option answers in the style of the US Medical License Exam (USMLE). Since it consists of diverse medical knowledge, MedQA is a challenging benchmark and is thus used as our training corpus. Specifically, the training set consists of 10,178 samples, and the test set has 1273 questions.

MedMCQA. MedMCQA (Pal et al., 2022) consists of 4-option multiple-choice QA samples from the Indian medical entrance examinations (AIIMS/NEET). This dataset covers 2.4K healthcare topics and 21 medical subjects. We use the validation set with 4,183 questions for evaluation.

MMLU*. MMLU (Hendrycks et al., 2020) is a comprehensive benchmark, including exam questions from 57 subjects (*e.g.*, STEM and social sciences). Each MMLU subject contains 4-option multiple-choice QA samples. Similar to prior works (Singhal et al., 2025), we select 6 subjects that are most relevant to medical and clinical knowledge: Anatomy, Clinical Knowledge, College Biology, College Medicine, Medical Genetics and Professional Medicine. For convenience, we denote this subset as MMLU*.

CMB. CMB (Wang et al., 2024b) is a comprehensive medical benchmark in Chinese, designed and rooted entirely within the native Chinese linguistic and cultural framework. It consists of two parts: CMB-Exam, featuring multiple-choice questions from qualification exams, and CMB-Clin, including complex clinical diagnostic questions derived from real case studies. In our experiments, we evaluate the models on the samples with single answers from the test set of CMB-Exam.

CMExam. CMExam (Liu et al., 2024b) is sourced from authentic medical licensing exams, containing more than 60K questions. It can reflect the comprehensive coverage of medical knowledge and reasoning required in clinical practice, covering Traditional Medicine Disease Patterns, Digestive System Diseases, Certain Infectious, etc. For evaluation, we select the data with single-choice answers from the test set.

CMMLU*. CMMLU (Li et al., 2024) is a comprehensive Chinese benchmark that covers various subjects, including natural sciences, social sciences, engineering, and the humanities. Similar to MMLU-Medical, we also select the subjects that are most relevant to medical and clinical knowledge as the medical QA benchmarks, covering Anatomy, Clinical Knowledge, College Medical Statistics,

College Medicine, and Professional Medicine. For convenience, we refer to this subset as CMMLU* in the main experiments.

Other domain-specific QA. In addition to the medical QA, we also evaluate our method in the other domains, covering history, engineering, and law. Specifically, we follow Ren et al. (2024) and procure the relevant items from the Xiezhi (Gu et al., 2024) Benchmark for each domain. Xiezhi contains 249587 questions with 516 disciplines, ranging from 13 different categories. Since Ren et al. (2024) have publicly released the collected dataset, we directly reuse the corresponding training and test sets in our experiments.

MMedBench. MMedBench (Qiu et al., 2024) is a multilingual medical multiple-choice QA benchmark across six primary languages: English, Chinese, Japanese, French, Russian, and Spanish. The entire test set of MMedBench comprises 8,518 QA pairs. For a unified evaluation, we remove the samples with multiple answers and use the filtered 8,178 samples as the evaluation set.

HaluEval. HaluEval (Li et al., 2023a) is a large collection of generated and human-annotated hallucinated samples for evaluating the performance of LLMs in recognizing hallucination. It includes 5,000 general user queries with ChatGPT responses and 30,000 task-specific examples from three tasks, *i.e.*, question answering, knowledge-grounded dialogue, and text summarization. In the evaluation, it randomly samples a ground-truth or a hallucinated output for each data. If the text is a hallucinated answer, the LLM should recognize the hallucination and output "Yes", which means the text contains hallucinations. If the text is a ground-truth answer, the LLM should output "No" indicating that there is no hallucination. The accuracy can evaluate the hallucination, where a larger value means less hallucination. In our study, we use task-specific examples from HaluEval for hallucination evaluation.

A.2 Training and Evaluation Details

For model training, we fine-tune each LLM with a batch size of 16 and a peak learning rate of 1e-4, except 2e-4 for LLaMA3-3B. The warm-up ratio is 0.1 and the maximum tokenizer length is 2,048. All models are trained with LoRA (Hu et al., 2021) method for 1 epoch. We conduct all experiments on 8 NVIDIA A100 (40GB) GPUs. For conflict detection in KaFT, we set the temperature to 0.7 and sample 10 responses for each query. During evaluation, we set the temperature to 0 for reproducibility. Specifically, we use the widely-used `lm-evaluation-harness`[2] toolkit to measure the zero-shot accuracy of LLMs on multiple-choice QA benchmarks.

A.3 Prompt Details

As mentioned in §2, we use the ICL-based method to probe the LLMs' internal domain-specific knowledge for each query. Specifically, we randomly select three samples from the training set as the few-shot examples, and use them to guide the output format of LLMs. Taking the medical QA as an example, we present a case as follows:

Probing for LLMs' internal knowledge

For the following medical question, select one correct answer from A to D.
Question: A 3900-g (8.6-lb) male infant is delivered at 39 weeks' gestation via spontaneous vaginal delivery. Pregnancy and delivery were uncomplicated but a prenatal ultrasound at 20 weeks showed a defect in the pleuroperitoneal membrane. Further evaluation of this patient is most likely to show which of the following findings?
Options:
A. Gastric fundus in the thorax
B. Pancreatic ring around the duodenum
C. Hypertrophy of the gastric pylorus
D. Large bowel in the inguinal canal
Answer: A

. . . (the other two examples)

For the following medical question, select one correct answer from A to D.
Question: `<question>`
Options:
A. `<option_a>`
B. `<option_b>`
C. `<option_c>`
D. `<option_d>`
Answer: `[output]`

where `<question>` and `<option>` denote the input question and answer options, `[output]` denotes the corresponding model response.

[2]https://github.com/EleutherAI/lm-evaluation-harness

Backbone	Subset	English Medical Benchmark			Chinese Medical Benchmark			Avg.
		MedQA	MedMCQA	MMLU*	CMB	CMExam	CMMLU*	
Mistral-7B	Random	55.85	50.32	66.67	36.24	36.10	35.30	46.75
	`right`	53.10	47.72	66.55	37.48	36.69	36.18	46.29
	`might-right`	57.11	50.42	67.69	36.85	35.92	36.45	47.41
	`might-wrong`	56.56	50.25	64.71	35.09	34.43	34.29	45.89
	`wrong`	27.89	32.27	22.58	18.47	19.95	24.95	24.35
Qwen1.5-7B	Random	50.82	50.35	62.06	75.05	76.62	70.27	64.19
	`right`	50.04	49.63	60.78	74.95	77.02	70.95	63.89
	`might-right`	51.37	50.35	62.87	75.61	77.37	70.32	64.65
	`might-wrong`	45.48	45.04	46.32	68.41	70.70	65.08	56.84
	`wrong`	15.63	29.69	14.72	22.79	20.13	20.87	20.64
LLaMA3-8B	Random	60.80	55.42	71.68	46.84	47.15	44.42	54.38
	`right`	59.31	56.59	72.20	47.37	47.54	45.64	54.77
	`might-right`	60.49	55.58	73.15	48.13	48.01	46.96	55.39
	`might-wrong`	61.67	55.49	72.58	45.76	46.12	43.75	54.23
	`wrong`	23.72	30.43	37.96	23.65	22.16	25.04	27.16
LLaMA3-3B	Random	53.10	49.03	62.32	38.26	38.40	37.93	46.51
	`right`	51.30	50.08	62.07	40.39	40.29	37.73	46.98
	`might-right`	54.20	48.29	61.09	37.09	38.14	37.48	46.05
	`might-wrong`	50.75	46.14	61.11	30.26	30.92	32.95	42.02
	`wrong`	19.25	25.32	19.80	21.02	20.63	24.58	21.77

Table 7: **Full results of Figure 2 (b), *i.e.*, comparison of different subsets**. For reference, we also present the results of SFT on the randomly selected samples. Note that all subsets hold the same number of training samples.

A.4 Full Results

Here, we report the full results of experiments in our main paper. Specifically, Table 7 shows the detailed results of different subsets. Table 8 shows the detailed results of varied `wrong` data. Table 9 and Table 10 show the ablation study of our proposed conflict detection method and KaFT method, respectively. Table 11 shows the detailed results of parameter analyses of α and β. Table 12 shows the detailed results on the MMedBench. Please refer to the tables for more details.

Backbone	Ratio	English Medical Benchmark			Chinese Medical Benchmark			Avg.
		MedQA	MedMCQA	MMLU*	CMB	CMExam	CMMLU*	
Mistral-7B	0%	52.95	49.20	63.76	38.91	39.78	37.27	46.98
	25%	53.73	49.15	64.52	38.86	39.14	37.47	47.15
	50%	54.91	49.94	62.04	37.81	38.70	36.32	46.62
	75%	55.85	49.32	62.93	38.17	38.99	38.37	47.27
	100%	54.99	50.08	62.94	38.00	39.17	38.13	47.22
LLaMA3-3B	0%	58.37	51.11	68.69	38.09	36.92	36.17	48.22
	25%	58.29	51.23	69.25	37.54	37.17	39.25	48.79
	50%	60.02	50.32	69.69	36.75	36.73	39.15	48.78
	75%	61.19	51.66	68.54	35.87	35.51	37.29	48.34
	100%	59.86	43.75	68.06	36.25	36.25	35.48	46.61

Table 8: **Full results of Figure 2 (c), *i.e.*, analysis of ratio of `wrong` data**. Notably, we randomly select varied samples from the `wrong` subset and merge them with the other three subsets. We set three different random seeds for data sampling and report the average results in this table.

Backbone	Method	English Medical Benchmark			Chinese Medical Benchmark			Avg.
		MedQA	MedMCQA	MMLU*	CMB	CMExam	CMMLU*	
LLaMA3-8B	Random	60.80	55.42	71.68	46.84	47.15	44.42	54.38
	Ours	23.72	30.43	37.96	23.65	22.16	25.04	27.16
	-w/o diverse query	41.32	40.04	61.80	29.93	29.60	31.05	38.96
	-w/o response sampling	29.38	26.63	45.03	27.46	27.27	25.16	30.15
	-w/o both	55.22	52.88	69.61	38.10	39.14	39.07	49.00

Table 9: **Full results of Table 3, *i.e.*, ablation of our conflict detection method**. LLaMA3-8B is used as the base model. Notably, we use different conflict detection to select the `wrong` subset for training. The worse results mean that the method can detect the conflict data more accurately, *i.e.*, worse results refer to better performance.

Backbone	Method	English Medical Benchmark			Chinese Medical Benchmark			Avg.
		MedQA	MedMCQA	MMLU*	CMB	CMExam	CMMLU*	
Mistral-7B	**KaFT (Ours)**	59.54	49.87	68.47	38.11	37.35	40.73	**49.01**
	-w. constant	59.86	43.75	68.06	36.25	36.25	35.48	46.61
	-w. auto-adapt	59.07	51.09	68.53	37.72	36.92	39.56	48.82
Qwen1.5-7B	**KaFT (Ours)**	53.57	49.82	63.23	75.57	77.27	72.07	**65.25**
	-w. constant	52.71	50.20	61.60	74.30	76.15	70.16	64.19
	-w. auto-adapt	52.55	50.06	63.04	75.18	77.04	70.73	64.77
LLaMA3-8B	**KaFT (Ours)**	64.10	56.94	74.01	47.39	47.89	45.44	**55.96**
	-w. constant	61.82	55.75	73.34	45.99	45.85	44.95	54.62
	-w. auto-adapt	61.35	56.49	73.19	47.91	48.45	46.26	55.61
LLaMA3-3B	**KaFT (Ours)**	54.52	50.51	64.93	40.19	39.93	39.70	**48.30**
	-w. constant	54.99	50.08	62.94	38.00	39.17	38.13	47.22
	-w. auto-adapt	53.57	50.30	63.65	39.15	39.55	38.17	47.40

Table 10: **Full results of Figure 3, *i.e.*, performance comparison (%) between different reward strategies in KaFT**. The best average results are in **bold**.

Backbone	Method	English Medical Benchmark			Chinese Medical Benchmark			Avg.
		MedQA	MedMCQA	MMLU*	CMB	CMExam	CMMLU*	
Mistral-7B	α=0.1, β=0.1	57.03	48.10	68.84	37.67	37.79	40.54	48.33
	α=0.1, β=0.5	59.54	49.87	68.47	38.11	37.35	40.73	**49.01**
	α=0.1, β=1.0	58.68	50.39	68.35	37.99	37.63	39.29	48.72
	α=0.5, β=0.1	58.76	48.82	67.98	37.38	36.76	40.05	48.29
	α=0.5, β=0.5	59.78	51.21	68.06	37.30	36.72	39.27	48.72
	α=0.5, β=1.0	59.31	49.70	69.05	37.99	36.87	40.18	48.85
	α=1.0, β=0.1	56.32	46.12	67.67	36.83	36.87	40.04	47.31
	α=1.0, β=0.5	60.09	50.18	68.41	36.31	36.39	38.77	48.36
	α=1.0, β=1.0	59.86	43.75	68.06	36.25	36.25	35.48	46.61
Qwen1.5-7B	α=0.1, β=0.1	53.57	50.13	62.45	75.57	77.27	71.55	65.09
	α=0.1, β=0.5	53.57	49.82	63.23	75.57	77.27	72.07	**65.25**
	α=0.1, β=1.0	52.47	50.18	62.53	75.25	76.89	71.46	64.80
	α=0.5, β=0.1	52.95	49.65	61.96	75.14	76.99	71.94	64.77
	α=0.5, β=0.5	53.97	50.04	61.24	75.09	76.69	71.30	64.72
	α=0.5, β=1.0	53.42	50.27	60.82	74.59	76.19	70.90	64.37
	α=1.0, β=0.1	52.87	50.59	62.35	75.17	76.48	70.45	64.65
	α=1.0, β=0.5	51.53	50.08	61.73	74.86	76.13	70.49	64.14
	α=1.0, β=1.0	52.71	50.20	61.60	74.30	76.15	70.16	64.19

Table 11: **Full results of Figure 4, *i.e.*, parameter analyses of α and β**. The best average results are in **bold**.

Backbone	Method	MMedBench						Avg.
		Chinese	English	French	Japanese	Russian	Spanish	
Mistral-7B	Base	35.76	51.06	41.74	26.88	48.05	49.27	42.13
	Vanilla SFT	41.07	58.99	48.29	30.00	60.94	58.02	49.55
	No-conflict	41.77	56.56	49.84	36.25	62.50	56.20	50.52
	Self-aligning	41.04	54.91	46.11	32.50	61.33	55.87	48.63
	KaFT (Ours)	41.36	58.37	48.29	38.12	62.11	57.22	**50.91**
Qwen1.5-7B	Base	82.25	46.19	41.12	35.62	55.08	49.02	51.55
	Vanilla SFT	79.16	46.82	45.48	36.25	61.72	49.12	53.09
	No-conflict	83.07	50.90	48.60	44.38	57.81	52.88	56.27
	Self-aligning	82.11	47.29	46.11	38.75	58.59	51.50	54.06
	KaFT (Ours)	82.81	51.61	47.66	40.62	63.67	52.63	**56.50**
LLaMA3-8B	Base	56.98	58.68	53.58	40.00	55.86	59.48	54.10
	Vanilla SFT	58.20	61.74	57.01	42.38	63.67	61.93	57.49
	No-conflict	59.40	60.72	58.57	46.25	62.50	61.42	58.14
	Self-aligning	59.78	60.49	57.63	45.62	62.50	62.31	58.09
	KaFT (Ours)	60.19	63.24	58.88	46.25	63.67	62.00	**59.04**
LLaMA3-3B	Base	46.15	49.65	40.81	28.12	50.78	49.31	44.14
	Vanilla SFT	47.14	53.10	40.81	33.75	51.95	51.79	46.42
	No-conflict	48.63	53.57	42.06	35.00	51.17	51.17	46.93
	Self-aligning	47.72	52.40	39.56	33.12	51.17	50.95	45.82
	KaFT (Ours)	48.22	53.42	46.11	33.12	51.95	52.81	**47.61**

Table 12: **Full results of Figure 5, *i.e.*, performance of MMedBench** (Qiu et al., 2024). In addition to LLaMA3-8B models, we also report the results of other LLMs. The best average results are in **bold**.