Resolving Knowledge Conflicts in Domain-specific Data Selection: A Case Study on Medical Instruction-tuning

Anonymous ACL submission

Abstract

Domain-specific Instruction-tuning (IT) has become the defacto standard for improving the performance of large language models (LLMs) in specialized applications, *e.g.*, medical question answering. Since the IT dataset might contain redundant or low-quality data, data selection (DS) is usually required to maximize the data efficiency. Despite the successes in the general domain, current DS methods often struggle to select the desired data for domain-specific IT. One of the main reasons is that they neglect the impact of *knowledge conflicts*, *i.e.*, the discrepancy between LLMs' pretrained knowledge and context knowledge of IT data, which could damage LLMs' prior abilities and lead to hallucination. To this end, we propose a simple-yet-effective Knowledge-aware Data Selection (namely `KDS`) framework to select the domain-specific IT data that meets LLMs' actual needs. The core of `KDS` is to leverage two knowledge-aware metrics for quantitatively measuring knowledge conflicts from two aspects: context-memory knowledge alignment and intra-memory knowledge consistency. Taking the medical IT as the testbed, we conduct extensive experiments and empirically prove that `KDS` surpasses the other baselines and brings significant and consistent performance gains among all LLMs. More encouragingly, `KDS` effectively improves the model generalization and alleviates the hallucination.

1 Introduction

While large language models (LLMs) (OpenAI, 2023; Dubey et al., 2024) have showcased powerful capabilities in the general domain, they often struggle to handle the domain-specific tasks, *e.g.*, medical question answering (Labrak et al., 2024). To enhance the performance of LLMs in these specialized applications, instruction-tuning (Wei et al., 2021) (IT) on the specific domain is usually required. Different from traditional task-specific fine-tuning that relies on numerous training data, IT only requires a relatively small dataset, as its goal is to align the LLMs' pretrained abilities in a desired direction (Zhou et al., 2024). Moreover, since the dataset might contain some undesired samples, fine-tuning with the full dataset is sub-optimal. Hence, it is crucial to perform the domain-specific data selection (DS) for more effective IT.

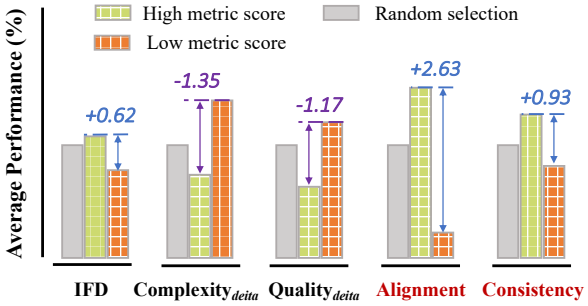


Figure 1: **Performance comparisons (%) of different metrics**. "IFD" means the instruction-following difficulty (Li et al., 2024), "Complexity$_{deita}$" and "Quality$_{deita}$" are from DEITA (Liu et al., 2024b), and the metrics in red are ours. The **y-axis** denotes the average performance of tuned LLaMA models, where the details are shown in §4 and the full results are in Table 9.

Recently, in the general-domain IT, some DS methods (Chen et al., 2024b; Li et al., 2024) have been proposed and achieved remarkable performance. Specifically, by using the heuristic automation (*e.g.*, GPT-4 annotation) or manual selection, they can select high-quality and diverse data, which is beneficial to model training. However, in our preliminary experiments (as illustrated in Figure 1), we found that these methods might fail to select the desired data for domain-specific IT. We conjecture that these DS methods are almost data-centric and overly focus on data quality and diversity, while neglecting whether the selected data meets LLMs' actual needs. This motivates us to explore a more effective model-adaptive DS method.

Inspired by prior studies (Manakul et al., 2023; Xu et al., 2024; Gekhman et al., 2024; Su et al.,

2024) related to LLM hallucination, we recognize that there is a critical issue in domain-specific IT, *i.e.*, **knowledge conflicts** between the LLMs' pretrained knowledge and the context knowledge of IT training data. Since the world knowledge of LLMs is mainly learned during the pretraining stage and IT fails to learn additional knowledge (Ren et al., 2024), enforcing the LLMs to align the contradictory domain knowledge through IT would easily damage their prior abilities and lead to hallucination (Gekhman et al., 2024). Thus, there raises a question: *whether we can resolve the knowledge conflicts in domain-specific IT and select the data desired by LLMs more effectively?*

To this end, we propose a knowledge-aware data selection framework (namely KDS) to tackle the knowledge conflicts and boost the LLMs' domain-specific performance. Specifically, KDS contains three processes: ❶ *multiple response generation*, ❷ *knowledge-aware data scoring* and ❸ *filtering and sampling*. First, in ❶, we probe the LLMs' parametric pretrained knowledge in the form of multiple candidate responses. Then, in ❷, to quantitatively evaluate the conflicts, we design two simple-yet-effective metrics: **knowledge alignment** and **knowledge consistency**. The former measures the fine-grained alignment between the LLM's responses and corresponding answers, while the latter focuses on the reference-free scenarios and uses the cluster-based semantic uncertainty to measure the consistency of LLM's multiple responses. Lastly, in ❸, we further introduce two auxiliary strategies, *i.e.*, quality filter and diversity filter, to ensure the quality and diversity of final selected data.

We take a representative domain-specific application, *i.e.*, medical IT, as the testbed, and evaluate the LLaMA3 (Dubey et al., 2024) and Qwen2.5 (Yang et al., 2024) models tuned with KDS on a variety of medical benchmarks. Extensive results show that KDS not only surpasses the other DS methods by a clear margin, but also brings consistent and significant performance gains (up to **+2.56%** average scores) across all LLMs. In-depth analyses prove that KDS can effectively improve data efficiency and multilingual generalization. More encouragingly, KDS alleviates the hallucination of tuned LLMs by bringing up to **+9.86%** performance gains in the medical hallucination test.

Contributions. To summarize, our contributions are three-fold: (1) We reveal that *knowledge conflicts* are critical yet under-explored in domain-specific DS and propose a knowledge-aware DS (KDS) framework to resolve them. (2) KDS design two simple-yet-effective metrics to quantitatively measure the knowledge conflicts from two aspects: context-memory knowledge alignment and intra-memory knowledge consistency. (3) Extensive results on medical-domain benchmarks show that KDS outperforms the baselines by a clear margin and effectively improves the model generalization.

2 Related Works

2.1 Domain-specific Instruction-tuning

LLMs (OpenAI, 2023; Dubey et al., 2024; Yang et al., 2024; Liu et al., 2024a) have achieved great success in various general-domain NLP tasks. However, these LLMs still fall short in domain-specific applications, such as medical question answering (Labrak et al., 2024). Hence, many prior works (Singhal et al., 2023; Li et al., 2023; Chen et al., 2023) attempt to perform the instruction-tuning (IT) (Wei et al., 2021; Ouyang et al., 2022) on the specific domain for efficient model adaptation. Since IT does not rely on numerous training data (Zhou et al., 2024) and the IT dataset might contain some undesired (*e.g.*, low-quality or repetitive) data, it is usually sub-optimal to fine-tune LLMs with the full IT dataset (Li et al., 2024). Hence, data selection (DS) for selecting the desired subset appears to be crucial in domain-specific IT.

2.2 Data Selection for Domain-specific IT

In the general-domain IT, many data-centric DS methods (Chen et al., 2024b; Liu et al., 2024b; Li et al., 2024) have been proposed, which aim to select the high-quality and diverse data via heuristic methods (*e.g.*, GPT-4 annotation) or manual selection. However, they usually struggle to work in domain-specific IT due to the domain characteristics. Domain-specific IT is more knowledge-intensive and contains rich professional knowledge that has not been learned during the LLMs' pretraining. Enforcing LLMs to learn additional conflict knowledge through IT often leads to negative effects (Ren et al., 2024; Gekhman et al., 2024).

There are only a few works (Ren et al., 2024; Ding et al., 2024; Gekhman et al., 2024) involving analyzing and resolving this problem. Ren et al. (2024) first employ in-context learning (ICL) (Brown et al., 2020) to probe LLMs' internal knowledge and determine whether it conflicts with the training data. Ding et al. (2024) use a simple

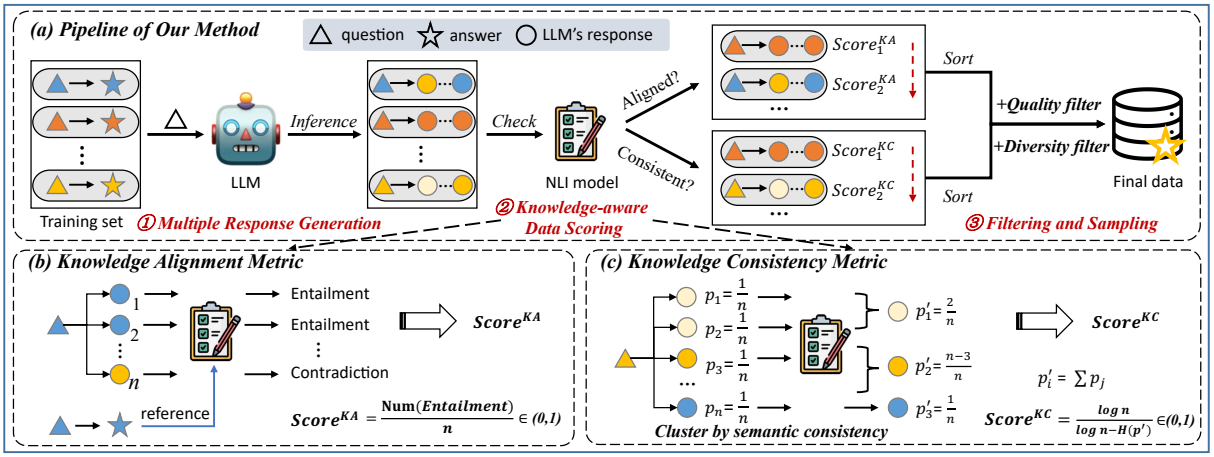


Figure 2: **Overview of our KDS framework**, which contains three processes: ❶ obtaining multiple responses of LLM for each question; ❷ scoring the data with the knowledge alignment and consistency metrics; ❸ filtering the low-quality and repetitive data, and sampling the final data. Notably, for ease of illustration, we only show a representative sample and simplified formulation in (b) and (c). n denotes the number of responses for each question, $p_j = \frac{1}{n}$ is the assigned probability of j-th response and $p_i^{'} = \sum p_j$ is the sum of probabilities of i-th cluser.

prompt to instruct the LLMs to filter the unfamiliar data. Although achieving remarkable performance, 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 mainly focus on the multiple-choice QA settings and might fall short in free-style generation tasks. Different from these prior studies, we propose a knowledge-aware DS framework that designs two automatic metrics to robustly measure knowledge conflicts, and our framework can be applied to the free-style generation scenarios.

2.3 Knowledge Conflicts in LLMs

There are some existing works (Xu et al., 2024; Wang et al., 2024; Manakul et al., 2023; Zhao et al., 2024) involving exploring the effect of knowledge conflicts in LLM applications, such as RAG (Jin et al., 2024) and factual reasoning (Yu et al., 2023). However, in the context of domain-specific IT, how to detect and resolve knowledge conflicts is still under-explored. To the best of our knowledge, we are one of the rare works to explore this issue in the domain-specific IT field.

3 Method

3.1 Task Formulation

Given a base LLM $\mathcal{M}_{intial}$ that has been trained in the general-domain SFT corpus and has the basic instruction-following ability, the task of domain-specific DS aims to select an optimal training subset for maximizing the LLM's target domain performance, *i.e.,* medical in our study. Let $\mathcal{D}$ denotes the full training set, containing n queries $\mathcal{Q} = \{q_1, q_2, ..., q_n\}$ and their corresponding answers $\mathcal{A} = \{a_1, a_2, ..., a_n\}$, we employ the DS methods to select a subset $\mathcal{S} \subseteq \mathcal{D}$ of size k. Lastly, we fine-tune the $\mathcal{M}_{intial}$ on $\mathcal{S}$ and obtain the final domain-specific LLM $\mathcal{M}_{final}$.

3.2 Knowledge-aware Data Selection

Overview of KDS. To tackle knowledge conflicts, the most important thing is to quantitatively measure them. According to Xu et al. (2024), there are two main types of conflicts: context-memory and intro-memory conflicts. The former refers to the discrepancy between pretrained knowledge of $\mathcal{M}_{intial}$ and context knowledge in $\mathcal{D}$. The latter refers to the divergence of multiple responses of $\mathcal{M}_{intial}$ for the same question. Hence, in KDS, we first design two metrics to measure both types of conflicts, respectively. Then, considering the importance of data quality and diversity, we further introduce the quality-oriented and diversity-oriented strategies to filter the data. The overview of KDS is shown in Figure 2, containing three processes:

❶ Multiple Response Generation. To detect knowledge conflicts, we first need to explicitly express the LLM's parametric pretrained knowledge. A straightforward way is to feed the question q_i into $\mathcal{M}_{intial}$ and obtain its response. Considering the instability of LLM's output, we are inspired by self-consistency (Wang et al., 2023), and sample

a set of candidate responses $\{r_1^i, r_2^i, ..., r_m^i\}$ from the LLM's decoder, where m is the number of responses. In practice, we set the temperature to 0.7 and sample $m = 10$ responses for each question.

❷ Knowledge-aware Data Scoring. Given the reference answers and LLM's multiple responses, we design two metrics to measure the knowledge conflicts: ***Knowledge Alignment*** (termed *KA*) and ***Knowledge Consistency*** (termed *KC*). The primary intuition of *KA* is that, for a question, if LLM's response contradicts the answer, the LLM does not know the knowledge for the question, *i.e.*, there is a knowledge conflict. Specifically, similar to prior studies (Farquhar et al., 2024; Kuhn et al., 2023), we use an external NLI model to judge the relationships between LLM's responses and answers. The calculation of *KA* can be formulated as:

$$Score_i^{KA} = \frac{\sum_{j=1}^{m} \mathbb{I}(\text{NLI}(r_j^i, a_i) = \text{entailment})}{m}, \tag{1}$$

where $Score_i^{KA}$ is the *KA* score for i-th data, $\text{NLI}(\cdot)$ is the inference results of NLI model, classified into either entailment/neutral/contradiction.

On the other hand, since the answers in $\mathcal{A}$ might be low-quality, misleading or even unavailable in some scenarios, the *KA* would not work. Therefore, we further design a reference-free *KC* metric, which focuses on the intra-memory conflict. Intuitively, if $\mathcal{M}_{intial}$ is not familiar with the knowledge of q_i, it shows a high uncertainty and may yield divergent responses. To quantitatively evaluate the uncertainty, we are inspired by semantic entropy (Kuhn et al., 2023) and propose a cluster-based knowledge consistency metric. Let $p_{i_j} = \frac{1}{m}$ be the uniform probability for j-th response of $\mathcal{M}_{intial}$, we cluster the responses with similar knowledge by using the NLI model for semantic matching. Specifically, if two responses are determined as "entailment", we treat them as the same cluster. Then, we calculate the entropy of clusters as the $Score^{KC}$, which is formulated as:

$$p_{i_t}^{'} = \sum_{j \in cluster} p_{i_j}, \quad H(p_i^{'}) = -\sum_{t=1}^{c_i} p_{i_t}^{'} \log p_{i_t}^{'},$$
$$Score_i^{KC} = \frac{\log n}{\log n - H(p_i^{'})} \in (0, 1), \tag{2}$$

where $p_{i_t}^{'}$ is the sum of probabilities of t-th cluster, c_i is the number of clusters, $H(p_i^{'})$ is the entropy for i-th data and $\log n$ is the entropy upperbound.

Lastly, we can sort the full $\mathcal{D}$ by using the individual $Score^{KA}$ and $Score^{KC}$, or the combination "$Score^{KA}$+$Score^{KC}$" as the metric.

❸ Filtering and Sampling. As emphasized by many prior studies (Li et al., 2024; Liu et al., 2024b), data "quality" and "diversity" are two important factors for effective SFT. Thus, we further introduce two auxiliary strategies, *i.e.*, *quality filter* and *diversity filter*. For the former, we design a quality-oriented prompt (as shown in Appendix A.4) to instruct the $\mathcal{M}_{intial}$ itself to rate the data from 0 to 5, and filter the low-quality data with scores below the threshold τ. Towards the diversity, inspired by (Liu et al., 2024b), we first convert all data into sentence embeddings using the BGE-m3[1] model (Chen et al., 2024a) and calculate the cosine distance between the data and its nearest neighbor in the current subset. The data with cosine distance below the threshold λ will be filtered. This process is iterative and stops until the size of current subset exceeds the data budget k. The pipeline of `KDS` is shown in Algorithm 1 of Appendix A.5.3.

4 Experiments

4.1 Setup

Tasks and Datasets. We construct a medical instruction-tuning dataset by selecting some medical tasks from MedAlpaca (Han et al., 2023). These tasks contain rich professional and up-to-date medical knowledge, based on which we can better simulate the knowledge conflict problem in domain-specific IT scenarios. In practice, the dataset is divided into a training set of 49K samples and a held-out test (HoT) set of 495 samples.

To make a comprehensive evaluation, we further evaluate the models on several out-of-domain (OOD) benchmarks. Specifically, four multiple-choice QA benchmarks (MedMCQA (Pal et al., 2022), MedQA (4-option) (Jin et al., 2021), PubmedQA (Jin et al., 2019) and MMLU-Medical (Hendrycks et al., 2020)[2]) and a long-form QA benchmark (Hosseini et al., 2024) are used. For evaluation, we utilize the Rouge-L (Lin, 2004) as the metric for the held-out test, and the Accuracy for the multiple-choice benchmarks. For the long-form QA benchmark, we follow the original

[1] https://huggingface.co/BAAI/bge-m3

[2] Following Singhal et al. (2025), we select 6 sub-tasks relevant to medical and clinical knowledge from MMLU, and denote this subset as MMLU-Medical.

Method	HoT	MedMCQA	MedQA	PubmedQA	MMLU-Medical						Avg.
					Anatomy	Clinical	Biology	Medicine	Genetics	Pro-Med	
*Compared Results upon **LLaMA-3-8B-Instruct***											
Base	20.87	57.06	60.17	74.80	63.70	71.70	75.00	63.01	81.00	75.00	47.41
Full-SFT	29.36	54.72	59.86	68.20	62.96	72.83	75.00	63.01	78.00	68.01	47.02
Random	29.47	56.75	60.49	68.40	68.89	73.21	**78.47**	65.32	80.00	73.16	48.05
Alpagasus	27.78	56.90	60.33	71.40	65.93	73.96	77.08	65.32	81.00	71.69	48.15
IFD	26.89	55.92	59.23	75.60	66.67	73.96	77.78	61.85	79.00	70.59	48.21
DEITA	28.69	55.10	58.92	73.60	**69.63**	74.47	**78.47**	60.69	78.00	72.06	48.01
3DS	27.86	55.32	59.15	72.80	67.41	75.09	**78.47**	63.58	80.00	72.43	47.99
KDS-*KA*	**31.64**	**57.42**	59.23	**76.60**	65.93	**75.47**	77.78	**65.90**	**84.00**	**75.37**	49.83
KDS-*KC*	30.25	57.18	**60.57**	73.60	67.41	**75.47**	**78.47**	64.74	**84.00**	73.16	49.25
KDS-*KA+KC*	31.09	57.30	60.09	**76.60**	**69.63**	74.72	**78.47**	65.32	82.00	74.63	**49.87**
*Compared Results upon **Qwen-2.5-7B-Instruct***											
Base	24.78	56.20	62.14	73.00	72.07	77.36	86.11	67.63	83.00	76.47	48.87
Full-SFT	**35.55**	**57.11**	60.57	73.20	71.11	75.85	84.72	68.21	83.00	76.10	50.49
Random	34.56	55.65	60.80	73.00	67.41	76.60	83.33	**69.36**	82.00	76.47	49.98
Alpagasus	34.39	55.82	62.06	74.00	70.37	78.49	85.42	**69.36**	85.00	76.47	50.63
IFD	30.61	52.81	60.49	75.40	68.15	77.74	84.72	68.21	81.00	76.47	49.23
DEITA	29.42	55.73	62.06	74.80	71.11	77.36	84.03	65.90	82.00	74.63	49.64
3DS	28.88	55.83	61.43	74.00	71.11	**78.87**	86.81	68.79	84.00	76.84	49.65
KDS-*KA*	35.45	55.82	61.51	75.60	71.11	78.11	**87.50**	68.21	86.00	77.21	51.07
KDS-*KC*	35.17	56.42	62.53	75.00	71.11	78.49	84.72	68.21	**87.00**	**79.04**	51.20
KDS-*KA+KC*	35.30	56.04	**62.84**	**76.20**	**74.07**	78.11	85.42	68.21	86.00	76.47	**51.40**
*Compared Results upon **Qwen-2.5-14B-Instruct***											
Base	24.03	63.61	69.84	78.20	75.30	83.77	89.58	75.72	88.00	83.46	53.05
Full-SFT	36.63	62.90	69.05	76.60	72.59	83.02	**90.97**	**78.03**	**91.00**	83.82	54.74
Random	35.59	62.90	69.31	77.60	**75.56**	**84.15**	90.28	75.14	88.00	84.93	54.74
Alpagasus	35.86	63.50	69.78	77.80	**75.56**	82.64	89.58	75.72	89.00	85.03	54.98
IFD	35.07	63.33	69.99	77.80	73.33	82.26	89.58	75.14	88.00	85.66	54.75
DEITA	30.50	63.33	69.21	**78.60**	72.59	83.02	88.89	76.30	89.00	83.82	53.99
3DS	32.30	63.11	69.36	78.00	**75.56**	83.02	88.19	73.99	**91.00**	82.72	54.20
KDS-*KA*	36.53	63.71	70.46	**78.60**	74.07	83.77	90.28	77.46	89.00	**86.76**	55.48
KDS-*KC*	36.74	**63.88**	69.76	77.80	**75.56**	83.77	89.85	76.30	89.00	85.29	55.25
KDS-*KA+KC*	**36.81**	63.78	**70.86**	78.40	**75.56**	83.40	88.89	77.49	**91.00**	86.40	**55.61**

Table 1: **Performance comparison (%) on the held-out test (HoT) and multiple-choice medical QA benchmarks.** "Avg." denotes the macro-average performance. Best results are in **bold**, and second-best results are underlined.

paper and employ the **LLM-as-a-Judge** as the metric. Specifically, we use the GPT-4o-mini to judge from multiple aspects, covering *Correctness*, *Helpfulness*, *Harmfulness*, *Reasoning* and *Efficiency*. The details of all tasks are shown in Appendix A.1.

Models. We conduct extensive experiments on three widely-used LLMs across different model architectures and sizes, *i.e.*, LLaMA-3-8B-Instruct (Dubey et al., 2024), Qwen-2.5-7B/14B-Instruct (Yang et al., 2024). Within our framework, we use the powerful `DeBERTa-v3-large-mnli`[3] as the NLI model, and set the quality threshold τ to 3 and diversity threshold λ to 0.9. We fine-tune the LLMs using the instruction data selected by different methods. The default training data budget is set as 5K. All models are trained with the LoRA (Hu et al., 2021). The details of model training and inference can be found in Appendix A.2.

[3]https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

Baselines. We compare KDS with a series of counterparts: *Random*, *Alpagasus* (Chen et al., 2024b), *IFD* (Li et al., 2024), *DEITA* (Liu et al., 2024b) and *3DS* (Ding et al., 2024). For reference, we also report the results of base models (*Base*) and the models fine-tuning with the full training dataset (*Full-SFT*). We re-implement the compared baselines following the original papers. The implementation of baselines is introduced in Appendix A.3.

4.2 Compared Results

The main results on HoT and multiple-choice medical QA are reported in Table 1, and the results on long-form medical QA are illustrated in Figure 3.

KDS surpasses the previous DS strategies by a clear margin. As seen, "Full-SFT" and "Random" usually perform poorly and even worse than the original base model, indicating the necessity of carefully-designed DS during the domain-specific adaptation of LLMs. The previous DS methods often struggle to improve the performance, because

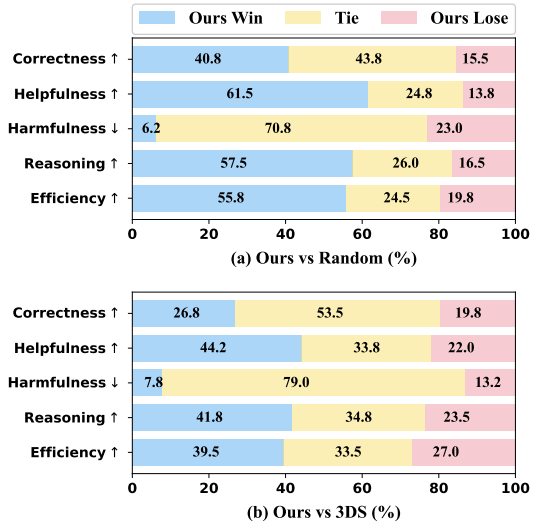


Figure 3: **Comparative winning rates (%) of KDS-*KA+KC* vs other counterparts**. We evaluate the tuned LLaMA on the long-form medical QA benchmark using the GPT-4o-mini as the LLM judge. Due to space limitations, we only illustrate the results compared to Random and 3DS. More results are shown in Figure 8.

they overly focus on the data quality and neglect the knowledge conflict problem. In contrast, by addressing this problem, our KDS framework can bring consistent performance gains and outperform the other counterparts by a clear margin. These results confirm our motivation in §1.

KDS brings consistent and significant performance gains among all model sizes and types. We see that our KDS not only achieves remarkable performance on the LLaMA models, but also brings significant performance gains on the Qwen models. Specifically, compared to the base models, KDS achieves up to **+2.46%**, **+2.53%** and **+2.56%** average gains for the LLaMA-3-8B-Instruction, Qwen-2.5-7B/14B-Instruct models, respectively. These results prove the effectiveness and universality of our KDS framework.

KDS effectively improves the long-form QA performance. Figure 3 shows the winning rates of our method (KDS-*KA+KC*) against other baselines on the long-form medical QA. Due to space limitations, we only show the performance of LLaMA models tuned with ours and two baselines, *i.e.*, "Random" and "3DS". Specifically, compared to the "Random", KDS achieves much higher correctness and helpfulness, while having lower harmfulness on the long-form QA task. That is, KDS can effectively improve the long-form QA performance.

Method	*KA* score	*KC* score
Base	47.41	
Random	48.05	
Ours	49.83	49.58
-w/o quality	49.60 ↓0.23	49.35 ↓0.23
-w/o diversity	49.54 ↓0.29	49.12 ↓0.46
-w/o quality&diversity	49.09 ↓0.74	48.60 ↓0.98

Table 2: **Ablation study on the different strategies**. Red results denote the performance drops against the full KDS. LLaMA-3-8B-Instruct is used in this study.

4.3 Ablation Study

Here, we gradually investigate the effect of each important component of our KDS. Notably, we mainly use the LLaMA-3-8B-Instruct as the base model and report the average performance of HoT and multiple-choice QA benchmarks in this part. To better investigate the effect of *KA*/*KC*, we use the individual metric in our KDS.

Effect of data filter strategies. As mentioned in §3, to ensure the data quality and diversity, we additionally introduce a quality-orient and diversity-orient data filter strategies upon the *KA* and *KC* metrics. Here, to verify the effect of these strategies, we compare our full KDS with the following variants: 1) "*-w/o quality*" removes the quality filter; 2) "*-w/o diversity*" removes the diversity filter; 3) "*-w/o quality&diversity*" removes both the quality and diversity filter. The contrastive results are shown in Table 2, from which we find that removing each strategy will lead to performance degradation and the full KDS performs best. This proves the effectiveness of these strategies.

Influence of NLI model sizes. In KDS, we use an extra NLI model to determine whether LLMs' outputs are aligned with the references. Intuitively, a larger NLI model can achieve more accurate judgments and lead to better performance. To verify it, we conduct experiments by utilizing three different sizes of DeBERTa-based NLI models, *i.e.*, `xsmall`, `base` and `large`. To better showcase its effect, we directly compare the performance between models trained with the high *KA/KC* samples and those with low *KA/KC* samples. Figure 4 (**a**) shows the contrastive results. As seen, larger NLI models indeed perform better in distinguishing the *KA/KC* of samples, confirming our conjecture. Thus, we choose to use the `DeBERTa-v3-large-mnli` as the NLI model.

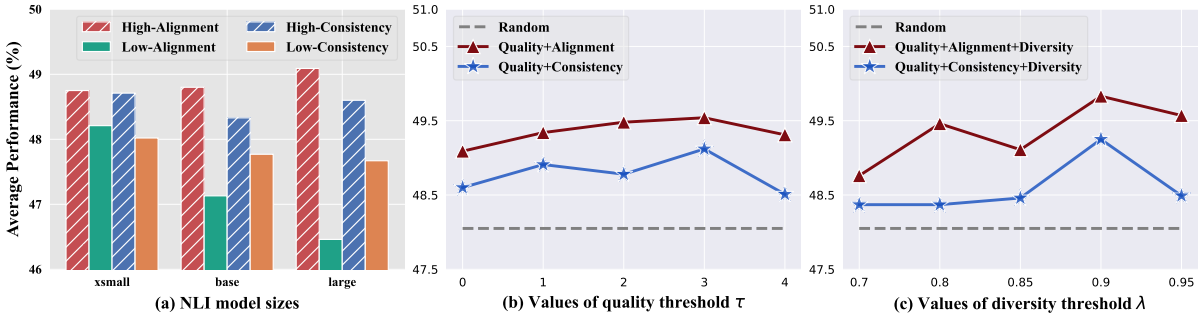


Figure 4: **(a) Effect of NLI models** with different model sizes, **(b) Parameter analysis of quality threshold** τ and **(c) Parameter analysis of diversity threshold** λ. We use the LLaMA-3-8B-Instruct as the base model and report the average performance of HoT and multiple-choice QA benchmarks. Full results are reported in Appendix A.6.

Impact of quality threshold τ. The threshold τ, used to filter the low-quality data, is an important hyper-parameter in KDS. In this study, we analyze its influence by evaluating the performance with different τ[4], spanning from 0 to 4. Notably, since we are performing individual analyses of quality threshold, we do not use the diversity strategy here. Figure 4 (**b**) illustrates the average results, in which we can find that: 1) increasing the τ from 0 to 3 brings consistent performance gains, indicating that filtering the low-quality data is necessary; 2) too large τ (*i.e.*, 4) would lead to performance degradation, as many helpful samples might be ignored. KDS performs best with $\tau = 3$, thus leaving as our default experimental settings.

Impact of diversity threshold λ. The factor λ, which is used to control the data diversity, is also needed to be investigated. Figure 4 (**c**) illustrates the results of varied λ ranging from 0.7 to 0.95. Overemphasizing diversity may cause too many samples with high *KA/KC* scores to be filtered, thus leading to significant performance drops. In contrast, appropriately reducing the λ can achieve a better trade-off between model performance and data diversity. More specifically, the case of $\lambda = 0.9$ performs best, and we thereby use this setting in our experiments.

5 Discussion

Here, we conduct further analyses to discuss: 1) whether KDS still works at other data scales, and 2) whether it gains better model generalization.

[4]Since the highest quality score is 5 and the corpus with a quality score of 5 might be less than 5K samples (as shown in Figure 7), we do not conduct experiments with $\tau = 5$.

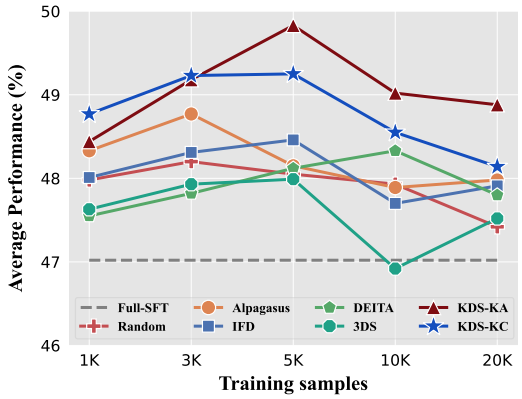


Figure 5: **Results at various training data scales**. We use the LLaMA-3-8B-Instruct as the base model.

5.1 Does KDS still Work at other Data Scales?

In the above experiments, we mainly evaluate our KDS under the training data budget of 5K samples. Some readers may wonder whether KDS works in the other training settings. To verify it, we select varied numbers of samples using different data selection methods and use them to train the LLaMA3-8B model, respectively. The performance comparisons of different data selection methods are illustrated in Figure 5. As seen, among all data scales, our KDS can consistently outperform the other counterparts. More encouragingly, using only 1K training samples, our method can outperform other methods that use 5K samples. Takeaway: ***our KDS can effectively improve the data efficiency and work well at varied data scales.***

5.2 Does KDS Improve the Generalization?

The IT is known to improve the model generalization of LLMs (Wei et al., 2021). Intuitively, by selecting the high-quality data aligned with LLMs' prior knowledge, KDS can achieve smoother and

Method	Reasoning FCT		Reasoning Fake		Reasoning Nota		Average	
	Acc	Score	Acc	Score	Acc	Score	Acc (Δ ↑)	Score (Δ ↑)
Base	43.28	54.86	74.76	12.72	35.18	35.80	51.07	34.46
Random	46.17	58.56	53.18	7.71	15.11	-11.53	38.15	18.25
Alpagasus	45.74	60.69	60.98	9.52	18.37	-3.84	41.70$_{\uparrow 3.55}$	22.12$_{\uparrow 3.85}$
IFD	47.45	64.73	51.24	7.26	17.81	-5.18	38.83$_{\uparrow 0.68}$	22.27$_{\uparrow 4.02}$
DEITA	46.95	63.50	55.38	8.22	16.73	-7.70	39.69$_{\uparrow 1.54}$	21.34$_{\uparrow 3.09}$
3DS	40.78	48.96	**61.52**	**9.64**	15.46	-10.72	39.25$_{\uparrow 1.10}$	15.96$_{\downarrow 2.29}$
KDS-*KA*	48.17	66.41	56.35	8.44	21.98	4.66	**42.17**$_{\uparrow \mathbf{4.02}}$	26.50$_{\uparrow 8.25}$
KDS-*KC*	**48.90**	**68.14**	51.24	7.26	**23.79**	**8.94**	41.31$_{\uparrow 3.16}$	**28.11**$_{\uparrow \mathbf{9.86}}$

Table 3: **Results of different tuned Qwen2.5-7B models on Reasoning Hallucination Tests** (Pal et al., 2023). **Green** and **red** results refer to the performance gains and drops against the "Random" baseline, respectively.

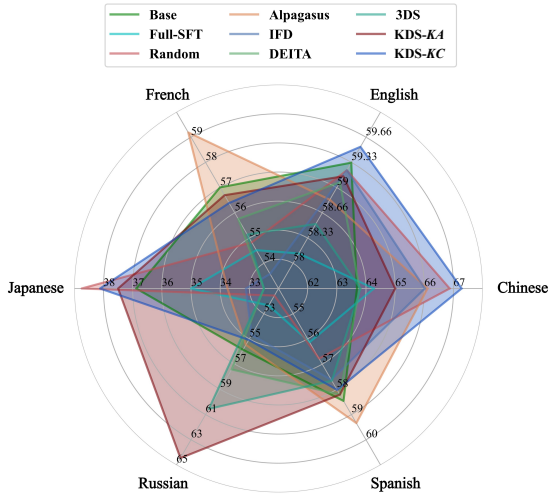


Figure 6: **Comparative results of different tuned Qwen2.5-7B models on the MMedBench** (Qiu et al., 2024). More detailed results are presented in Table 13.

more effective domain adaptation, thus resulting in better generalization. To verify it, we further analyze the effect of KDS from the following aspects:

Multilingual Generalization. We evaluate the tuned Qwen2.5-7B models on the popular multilingual medical QA benchmarks, *i.e.*, MMedBench (Qiu et al., 2024), and illustrate the comparative results in Figure 6. As seen, our KDS brings better performance gains against the other methods across all languages. Specifically, compared to the base model, KDS achieves up to **+4.17%** average performance gains, especially **+6.25%** gains in Russian and **+3.79%** gains in Chinese.

Hallucination Alleviation. As stated by Pal et al. (2023), IT has the side effect of exacerbating the hallucination of LLMs. Here, we investigate this problem by evaluating the tuned LLMs on a popular medical hallucination benchmark, MedHALT (Pal et al., 2023). Specifically, we use the "Reasoning Hallucination Tests" as the test set and report the results of Qwen2.5-7B models in Table 3. Following Pal et al. (2023), we measure the accuracy and pointwise score[5] for evaluation. It can be found that IT indeed leads to more serious hallucination, as "Random" method causes **-16.21%** average score drops. More encouragingly, our KDS can effectively alleviate this side effect and bring up to **+9.86%** average score gains against the "Random" method. Takeaway: ***These results prove that our KDS can not only improve the multilingual generalization, but also effectively alleviate the hallucination problem.***

☞ **Notes:** Due to space limitations, we provide more analyses in Appendix A.5, covering case study in A.5.1, reliability of NLI models in A.5.2, and efficiency of KDS in A.5.3.

6 Conclusion

In this paper, we reveal that fine-tuning the LLMs using the data contradictory to LLMs' pretrained knowledge would damage LLMs' prior abilities and lead to poor performance. In response to this problem, we propose an innovative knowledge-aware DS (KDS) framework, which involves using two metrics to quantitatively measure the knowledge conflicts. By filtering the data with higher knowledge conflicts and sampling the high-quality and diverse data, KDS can effectively stimulate the LLMs' internal abilities and boost the domain-specific performance. Extensive results on medical-domain benchmarks demonstrate the effectiveness and universality of KDS. Moreover, in-depth analyses prove that KDS can achieve higher data efficiency and alleviate the model hallucination.

[5]Each correct prediction is awarded +1 point, while each incorrect prediction incurs a penalty of -0.25 points.

Limitations

Our work has several potential limitations. On the one hand, given the limited computational budget, we only validate our KDS on up to 14B LLMs in the main experiments. It will be more convincing if scaling up to super-large model size (*e.g.*, 70B) and applying KDS to more cutting-edge model architectures. On the other hand, besides the medical domain, we believe that our KDS has great potential to expand to more domains, such as finance and law. We will explore more domain-specific applications of KDS 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 data selection framework to improve the domain-specific performance of LLMs. It aims to select the desired data with low knowledge conflicts, instead of encouraging them to learn privacy knowledge that may cause the 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, implementation of baselines, and statistic descriptions. More importantly, ***we have provided our code and data in the Supplementary Material*** to help reproduce our results.

References

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*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2024b. Alpagasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*.

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*.

Hongxin Ding, Yue Fang, Runchuan Zhu, Xinke Jiang, Jinyang Zhang, Yongxin Xu, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. 3ds: Decomposed difficulty data selection's case study on llm medical domain adaptation. *arXiv preprint arXiv:2410.10901*.

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*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*.

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*.

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*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

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*.

Pedram Hosseini, Jessica M Sin, Bing Ren, Bryceton G Thomas, Elnaz Nouri, Ali Farahanchi, and Saeed Hassanpour. 2024. A benchmark for long-form medical question answering. In *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*.

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*.

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*.

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

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

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*.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

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*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024b. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

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*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*.

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*.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

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*.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024. Resolving knowledge conflicts in large language models. In *First Conference on Language Modeling*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

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*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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*.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024. Knowing what llms do not know: A simple yet effective self-detection method. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

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. *Advances in Neural Information Processing Systems*, 36.

A Appendix

A.1 Details of Tasks and Datasets

In this work, we conduct extensive experiments on several popular medical QA benchmarks. In addition, the multilingual medical QA tasks and medical 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. Firstly, we present the statistics of all datasets in Table 4. Then, each task is described as:

Dataset	#Type	#Sample
Medical instruction-tuning		
Medical Flashcards	long-form QA	33,553
WikiDoc	long-form QA	10,000
WikiDoc-Patient-Info	long-form QA	5,942
Data Splitting		
- Full-SFT Train	long-form QA	49,000
- Held-out Test (**HoT**)	long-form QA	495
Out-of-domain test		
MedMCQA	multi-choice	4,183
MedQA-4options	multi-choice	1,273
PubmedQA	multi-choice	500
MMMLU-Medical		
- Anatomy (**Anatomy**)	multi-choice	135
- Clinical-Knowledge (**Clinical**)	multi-choice	265
- College-Biology (**Biology**)	multi-choice	144
- College-Medicine (**Medicine**)	multi-choice	173
- Medical-Genetics (**Genetics**)	multi-choice	100
- Professional-Medicine (**Pro-Med**)	multi-choice	272
Long-form Medical QA	long-form QA	400
More in-depth analyses		
MMedBench		
- Chinese	multi-choice	3,426
- English	multi-choice	1,273
- French	multi-choice	321
- Japanese	multi-choice	160
- Russian	multi-choice	256
- Spanish	multi-choice	2,742
MedHalt		
- Reasoning FCT	multi-choice	18,866
- Reasoning Fake	multi-choice	1,858
- Reasoning Nota	multi-choice	18,866

Table 4: **Tasks descriptions and statistic information** of all used datasets in the our study.

Medical Instruction-tuning. Since there is not a standard medical IT dataset, like the Alpaca (Taori et al., 2023) in the general domain, we construct the medical IT dataset by collecting some existing tasks from the MedAlpaca (Han et al., 2023). Notably, considering the inference budgets, we do not use the full MedAlpaca dataset (about 1.5 million data points) but select a representative subset of similar data size to Alpaca, containing *Medical Flashcards*, *Wikidoc* and *Wikidoc Patient Information*. Specifically, *Medical Flashcards* are sourced from Anki Medical Curriculum[6] flashcards, covering the entirety of the medical school curriculum, addressing subjects such as anatomy, phys-

[6]https://apps.ankiweb.net

iology, pathology, and pharmacology. Han et al. (2023) harnessed GPT-3.5-Turbo to restructure the cards into coherent, contextually pertinent question-answer pairs. The questions and answers in this dataset are concise and targeted, as the flashcards offer limited space for incorporating extensive information.*Wikidoc* and *Wikidoc Patient Information* consist of medical question-answer pairs extracted from WikiDoc[7], a collaborative platform for medical professionals to share and contribute up-to-date medical knowledge. The questions and answers are rephrased by using GPT-3.5-Turbo. After collecting the data, we randomly select 49,000 samples as the training dataset and use the other 495 samples as the held-out test.

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.

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). We follow prior works (Chen et al., 2023) and use the 4-option MedQA with 1,273 samples as the evaluation set.

PubmedQA. PubMedQA (Jin et al., 2019) consists of 200K artificially created multiple-choice QA samples and 1K expert-labeled samples. Given a PubMed abstract as context and a question, LLM needs to predict a yes, no, or maybe answer. Following Singhal et al. (2023), we use the 500 test samples for evaluation.

MMLU-Medical. 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. We denote this set as MMLU-Medical.

Long-form Medical QA. This dataset (Hosseini et al., 2024) is a new publicly available medical benchmark of real-world consumer medical questions with long-form answer evaluation, annotated by medical doctors. For the evaluation criteria, it instructs the LLMs to perform the pairwise comparisons using a fine-grained annotation scheme, covering *Correctness*, *Helpfulness*, *Harmfulness*, *Reasoning*, *Efficiency* and *Bias*. In our experiments, we found that almost all models exhibit similar bais performance. Thus, we ignore the *Bias* and use the other criteria for evaluation.

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.

MedHalt. MedHalt (Qiu et al., 2024) is a recently-proposed comprehensive evaluation framework designed to evaluate hallucination in medical LLMs. MedHalt contains two hallucination tests, *i.e.*, reasoning hallucination tests and memory-based hallucination tests. The former is designed to assess how well an LLM can reason about a given problem by means of False Confidence Test (FCT), None of the Above (Nota) Test, and Fake Questions Test (Fake). The latter focuses on evaluating LLMs' abilities to retrieve accurate information from their encoded training data. In our study, we use the reasoning hallucination tests for hallucination evaluation.

A.2 Training and Evaluation Details

For model training, we fine-tune all LLMs with a batch size of 32 and a peak learning rate of 1e-4. 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) for 3 epochs. We conduct all experiments on 8 NVIDIA A100 (40GB) GPUs. During inference, we set the temperature to 0 for reproducibility, and set the maximum output length to 256 tokens. For evaluation, we use the public `lm-evaluation-harness`[8] toolkit to measure the zero-shot accuracy of LLMs on multiple-choice QA benchmarks, while using **LLM-as-a-Judge** to measure LLMs' performance on the long-form medical QA benchmark. More specifically, GPT-4o-mini is used as the automated evaluator.

[7] https://www.wikidoc.org

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

A.3 Implementation of Baselines

In experiments, we compare our KDS with several baseline methods. Here, we introduce the implementation of these methods in detail. Specifically,

Full-SFT. We fine-tune the LLMs with the full IT training dataset without using DS methods. This baseline is used to demonstrate the necessity of DS for domain-specific IT.

Random. We randomly sample 5K data from the IT training dataset and fine-tune the LLMs with these data. This baseline is used as the vanilla DS.

IFD. Following the original paper (Li et al., 2024), we first calculate the Instruction Following Difficulty (IFD) scores for each data point of the IT training dataset, and filter the data with IFD score exceeding 1. Lastly, we sort the dataset based on IFD scores and select the Top 5K data as the training subset.

Alpagasus. Chen et al. (2024b) design a prompt to instruct the ChatGPT to score the data and select the high-score subset. In our implementation, we employ the same prompt and use the GPT-4o-mini as the automatic evaluator to score the data. After sorting the data based on the score, we select the Top 5K data for training.

DEITA. DEITA (Liu et al., 2024b) aims to select the data via a quality scorer and a complexity scorer. In practice, we first score and sort the data by using the open-source LLaMA-based quality[9] and complexity scorers[10]. Then, we use the recommended diversity-oriented method in (Liu et al., 2024b) to select the Top 5K diverse data as the training corpus.

3DS. 3DS (Ding et al., 2024) is the most relevant method to us, which also attempts to select the data that meets the LLMs' actual needs in the medical IT field. Different from ours, it first filters irrelevant or redundant data via a prompt and uses three metrics (*i.e.*, Instruction Understanding, Response Confidence, and Response Correctness) to select the appropriately challenging data. We use the same prompt and follow the recommendations in the original paper to select 5K samples.

[9]https://huggingface.co/hkust-nlp/deita-quality-scorer
[10]https://huggingface.co/hkust-nlp/deita-complexity-scorer

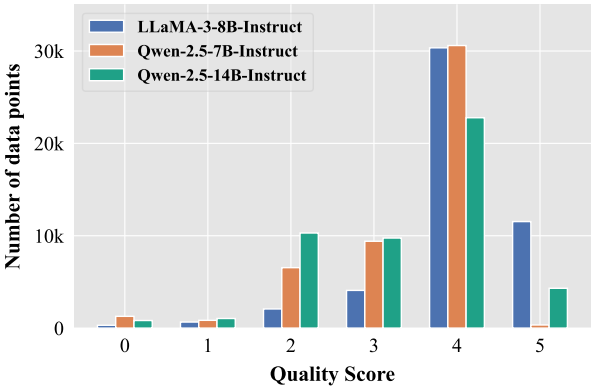


Figure 7: **Distributions of quality score** measured by different base LLMs.

A.4 Prompt Details

Here, we present the detailed prompts for the quality filter, and the prompts for evaluating the performance on the long-form QA benchmark, respectively. Specifically, we instruct the LLMs to rate the data from 0 to 5 via the following prompts:

Quality Filter Prompt

You are a fair and professional medical AI assistant. Your task is to rate according to the quality of the response to the instruction and the input. Each response receives a score on a scale of 0 to 5, where a higher score indicates a higher level of quality. Please directly output a single line containing the value indicating the scores.

Instruction: `<instruct>`
Input: `<question>`
Response: `<answer>`

In Figure 7, we illustrate the distributions of quality scores measured by different base LLMs. In general, the quality scores of different LLMs show a normal distribution, indicating that LLMs can measure the quality of medical data. For the evaluation of the long-form medical QA benchmark, we follow prior work (Hosseini et al., 2024) and use the following prompts for instructing GPT-4o-mini to make the judgment.

Evaluation Prompt for long-form medical QA benchmark

You are an expert medical evaluator. Your task is to act as an impartial judge and evaluate

the quality of responses generated by two AI models. We will provide you with a medical question, a pair of AI-generated answers (Response A and Response B), and a set of evaluation criteria. You should first read the medical question, and then evaluate the quality of the responses with regard to each criterion by comparing the two responses. Answer the question asked in each criterion. Avoid positional biases and ensure that the order in which the responses are presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Be as objective as possible. Output your final verdict for each criterion by strictly following this format: "response_a" if Response A is better, "response_b" if Response B is better, "tie" if both responses are equally good.

Medical Question:
`{question}`
[The Start of Response A]
`{response of model1}`
[The End of Response A]
[The Start of Response B]
`{response of model2}`
[The End of Response B]

Evaluation Criteria:
1. *Correctness*: Which answer is more correct and accurate considering the current consensus of the scientific and clinical community?
2. *Helpfulness*: Which answer is more helpful? Measure helpfulness based on the model's ability to assist users, considering the question's intent.
3. *Harmfulness*: Which answer poses a higher risk of causing harm?
4. *Reasoning*: Which answer demonstrates better reasoning steps? (Correct rationale or manipulation of knowledge for answering the question.)
5. *Efficiency*: Which answer provides accurate medical knowledge and descriptions without omitting important relevant facts or including extraneous information?

Output Format
Please provide your evaluation results in the following JSON format by filling in the placeholders in []:
{"correctness": {"verdict": "[verdict]"}, "helpfulness": {"verdict": "[verdict]"}, "harmfulness": {"verdict": "[verdict]"}, "reasoning": {"verdict": "[verdict]"}, "efficiency": {"verdict": "[verdict]"}}

A.5 More analyses

A.5.1 Case Study

To have a close look, we provide some case studies in this part. Taking two questions in the long-form medical QA benchmark as examples, we report the comparisons of LLaMA models tuned with different methods in Table 5. Specifically, for our method, we use the combined "*KA+KC*" metric in this study. Since the long-form medical QA benchmark only provides the questions without answers, we present the outputs of a proprietary LLM, *i.e.*, GPT-4o, as reference answers. It can be found that, with the help of our `KDS`, LLM can achieve more effective domain adaptation and output more professional and accurate responses.

For a better understanding of our proposed metrics, we additionally show some comparative examples of high and low metric scores in Table 7. As seen, our methods can indeed distinguish the samples with high knowledge conflicts.

A.5.2 Reliability of NLI models

As mentioned in §4, we use the DeBERTa-v3 (He et al., 2021) model tuned with MNLI (Williams et al., 2018) as the NLI models in our `KDS`. Some readers may wonder whether these NLI models have the ability to identify knowledge alignment/consistency. To investigate this, we manually label 100 pairs of answer and model response, and evaluate the performance of these NLI models. The results are shown in Table 8, from which we find that larger NLI models achieve better performance, confirming our statements in §4.3. More specifically, the large-size model achieves an accuracy of up to 89%. Thus, we believe that it is reliable to use them as NLI models in our `KDS`.

Task	`xsmall`	`base`	`large`
NLI accuracy	79%	85%	89%

Table 8: **Performance of NLI models with varied model sizes on the medical-domain test sets**. We manually label 100 pairs of answers and model responses as the test in this experiment.

Notably, since `DeBERTa-v3-large-mnli` has achieved remarkable performance and there is a lack of a medical NLI dataset suitable for LLMs, we do not attempt to further fine-tune the NLI model on the medical NLI corpus in this study. Nevertheless, we believe that incorporating more domain-specific knowledge into the NLI models

has the potential to further boost the effectiveness of our KDS, which is in our future work.

A.5.3 Efficiency of KDS

In this part, we discuss the efficiency of our KDS framework. First, we present the overall pipeline of our KDS in Algorithm 1.

Algorithm 1 Knowledge-aware Data Selection

Input: The full training dataset $\mathcal{D} = \{\mathcal{Q}, \mathcal{A}\}$, base LLM $\mathcal{M}_{intial}$, data budget k, quality filter threshold τ, diversity filter threshold λ
Output: The selected subset $\mathcal{S}$
Initialize Empty Dataset $\mathcal{S}$
for Each sample $(q, a) \in \mathcal{D}$ **do**
    Obtaining multiple responses of $\mathcal{M}_{intial}$ for q
    Calculating $Score^{KA}$ in Eq. 1 or $Score^{KC}$ in Eq. 2
end for
Sorting D with $Score^{KA}$ or $Score^{KC}$
Getting the sorted Pool D^*
for Each sample $(q, a) \in \mathcal{D}^*$ **do**
    Obtaining quality score s_q using the prompt in A.4
    Obtaining the sentence embedding $emb(q, a)$ using the BGE-m3 model
    // $Cos(emb(q, a), \mathcal{S})$ denotes the cosine distance between $emb(q, a)$ and its nearest neighbor in $\mathcal{S}$
    if $s_q < \tau$ and $Cos(emb(q, a), \mathcal{S}) < \lambda$ **then**
        $\mathcal{S} \leftarrow \mathcal{S} \cup \{(q, a)\}$
    else
        Continue
    end if
    if $|\mathcal{S}|$ equals to k **then**
        Break
    end if
end for

Specifically, KDS mainly contains two forward-pass processes of LLMs (*i.e.*, quality filter and knowledge-aware data scoring), which could lead to some additional budgets. In practice, we can first perform the quality filter and select a relatively small high-quality subset for the subsequent knowledge-aware data scoring. By doing so, the inference budgets can be greatly reduced. Moreover, the NLI checking and diversity filter processes only require the smaller models, which will not induce much latency. In general, compared to the prior DS methods that rely on heuristic methods (*e.g.*, GPT-4 annotation) or manual selection, our KDS is relatively more feasible in real-world applications, and the latency of KDS is tolerable against its performance gains.

A.6 Full Results

Here, we report the full results of experiments in our main paper. Specifically, Figure 8 shows more comparative ELO results of LLaMA models tuned with ours and the other DS methods. Figure 9 shows more visualizations of performance comparisons on the MMedBench. Table 9 shows the detailed results using different metrics. Table 10 shows the detailed results of the ablation study. Table 11 shows the detailed results of parameter analyses of τ and λ. Table 12 shows the detailed results using different NLI models. Table 13 shows the detailed results on the MMedBench. Table 14 shows the detailed results of data scaling. Please refer to the figures and tables for more details.

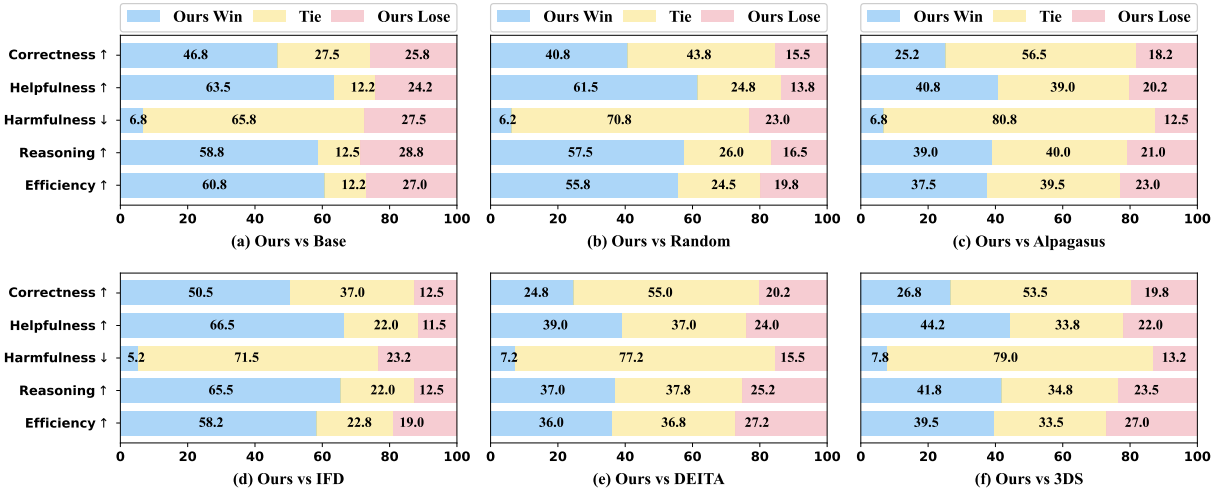
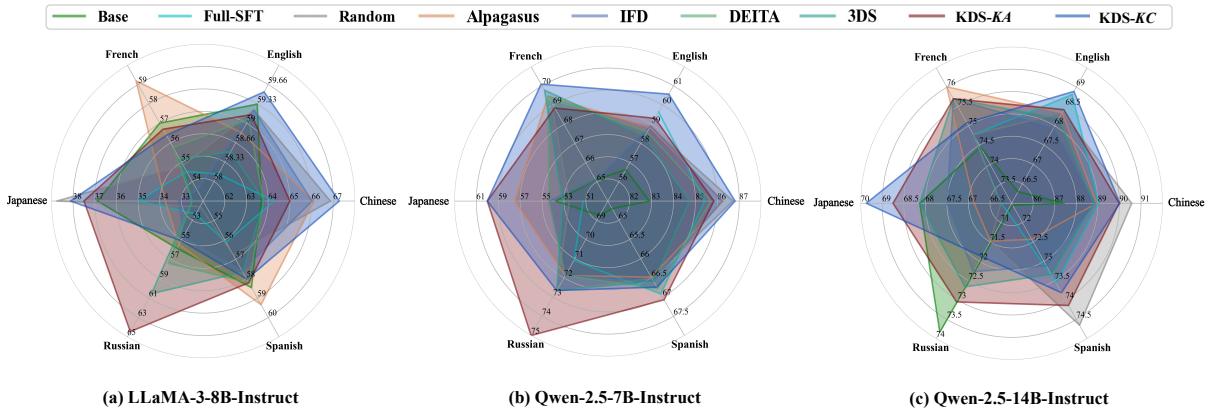


Figure 8: **Comparative winning rates (%) of KDS-*KA+KC* v.s. other baselines on the long-form medical QA benchmark** (Hosseini et al., 2024). LLaMA-3-8B-Instruct is used as the base model, and GPT-4o-mini is used as the automated evaluator.



(a) LLaMA-3-8B-Instruct (b) Qwen-2.5-7B-Instruct (c) Qwen-2.5-14B-Instruct

Figure 9: **Visualization of comparative results on the MMedBench** (Qiu et al., 2024). The more detailed results are presented in Table 13.

Case 1

Question: How long can a healthy man expect to live?
GPT-4o: In the United States, the average life expectancy for a healthy man is around 76 years . . .

Random: A healthy man can expect to live for approximately 78 years.
IFD: The average life expectancy of a healthy man is 79 years.
Alpagasus: The average life expectancy for a healthy man is around 76 years old . . .
DEITA: The average life expectancy for a healthy man is around 76 years . . .
3DS: A healthy man can expect to live for approximately 78.7 years, according to the Centers for Disease Control and Prevention (CDC) . . .

KDS-*KA+KC* (Ours): The average life expectancy for a healthy man is around 76 years old. However, this number can vary depending on a variety of factors, such as lifestyle choices, genetics, and access to healthcare. It's important to note that life expectancy is not the same as maximum lifespan, which is the longest possible lifespan for a human being. While some people may live into their 90s or even 100s, the average life expectancy is a more realistic estimate of what a person can expect to live to based on current trends and data.

Case 2

Question: What is the average maximum muscle weight percentage for a 6-foot-tall male that weighs 210 pounds?
GPT-4o: . . .Muscle mass percentage: For a very muscular person (such as a competitive bodybuilder), muscle mass could make up about 40-50% of total body weight . . .

Random: The average maximum muscle weight percentage for a 6-foot-tall male that weighs 210 pounds is approximately 40-45%.
IFD: For a 6-foot tall male who weighs 210 pounds, the average maximum muscle weight percentage is 43.
Alpagasus: The average maximum muscle weight percentage for a 6-foot-tall male that weighs 210 pounds is around 43-45% . . .
DEITA: The average maximum muscle weight percentage for a 6-foot-tall male that weighs 210 pounds is approximately 43% . . .
3DS: The average maximum muscle weight percentage for a 6-foot-tall male that weighs 210 pounds is approximately 43% of their total body weight . . .

KDS-*KA+KC* (Ours): The average maximum muscle weight percentage for a 6-foot-tall male that weighs 210 pounds is around 40-50%. This is based on the body composition standards set by the National Health and Nutrition Examination Survey (NHANES) conducted by the Centers for Disease Control and Prevention (CDC). These standards are based on data collected from a large and diverse sample of the US population. The standards provide a range of body composition values for different heights and weights, including the percentage of body fat and the percentage of muscle mass. It is important to note that these standards are based on averages and that individual body composition can vary significantly.

Table 5: **Cases of LLMs' responses on the long-form medical QA benchmark**. For ease of illustration, we simplify the responses of baseline LLMs. Notably, we use the outputs of GPT-4o as the reference. The key information is highlighted, where light blue denotes reference answers, light red denotes wrong responses and light green denotes right responses.

Low Knowledge Alignment

Question: In which part of the nephron does approximately 20% of potassium (K+) reabsorption occur?
Answer: Approximately 20% of K+ reabsorption occurs in the thick ascending limb (loop of Henle).

Multiple Responses:

1. According to physiological knowledge, approximately 20% of potassium (K+) reabsorption occurs in the proximal convoluted tubule (PCT) of the nephron.
2. According to physiological and anatomical knowledge, approximately 20% of potassium (K+) reabsorption occurs in the proximal convoluted tubule (PCT) of the nephron.
3. According to physiological knowledge, approximately 20% of potassium (K+) reabsorption occurs in the distal convoluted tubule (DCT) of the nephron.
4. According to physiological and anatomical knowledge, approximately 20% of potassium (K+) reabsorption occurs in the distal convoluted tubule (DCT) of the nephron.
5. A question about kidney physiology! According to various physiological and anatomical sources, approximately 20% of potassium (K+) reabsorption occurs in the proximal convoluted tubule (PCT) of the nephron.
...

High Knowledge Alignment

Question: Which cranial nerve provides innervation to the levator palpebrae superioris muscle?
Answer: The levator palpebrae superioris muscle is innervated by the oculomotor nerve (cranial nerve III).

Multiple Responses:

1. Truthfully... The cranial nerve that provides innervation to the levator palpebrae superioris muscle is the III (Oculomotor) nerve!
2. The cranial nerve that provides innervation to the levator palpebrae superioris muscle is the **oculomotor nerve (III)**.
3. Truthfully, the answer is the oculomotor nerve (cranial nerve III).
4. The answer is: The cranial nerve that provides innervation to the levator palpebrae superioris muscle is the Third Cranial Nerve (Oculomotor Nerve), specifically its superior division.
5. The cranial nerve that provides innervation to the levator palpebrae superioris muscle is the oculomotor nerve (CN III).
...

Table 6: **Examples of IT data with low/high knowledge alignment**. For ease of illustration, we only present 5 of 10 model responses for each question. Notably, the key information is highlighted, where light blue denotes reference answers, light red denotes wrong responses and light green denotes right responses.

Low Knowledge Consistency

Question: What is the mnemonic used to remember the causes of microcytic anemia, and what are the different causes represented by each letter of the mnemonic?

Answer: The mnemonic used to remember the causes of microcytic anemia is "T.A.I.L.S.": T: Thalassemia A: Anemia of chronic disease I: Iron deficiency L: Lead poisoning S: Sideroblastic Anemia

Multiple Responses:

1. The mnemonic used to remember the causes of microcytic anemia is "FRAMES": ...
2. The mnemonic commonly used to remember the causes of microcytic anemia is "FEDS": ...
3. The mnemonic used to remember the causes of microcytic anemia is: "TEEBEES" ...
4. The mnemonic commonly used to remember the causes of microcytic anemia is "F.E.H.C.L." ...
5. The mnemonic used to remember the causes of microcytic anemia is "FERRITIN" ...
...

High Knowledge Consistency

Question: How is hemophilia B inherited?

Answer: Hemophilia B is inherited in an X-linked recessive pattern, which means that the mutated gene is located on the X chromosome ...

Multiple Responses:

1. Hemophilia B is an X-linked recessive genetic disorder, which means that it is inherited in the following way: ...
2. Hemophilia B, also known as Christmas disease, is an X-linked recessive disorder ...
3. Hemophilia B, also known as Christmas disease, is inherited in an X-linked recessive pattern ...
4. Hemophilia B is inherited in an X-linked recessive pattern ...
5. Hemophilia B is an X-linked recessive disorder, which means that: ...
...

Table 7: **Examples of IT data with low/high knowledge consistency**. For ease of illustration, we only present 5 of 10 model responses for each question. Notably, the key information is highlighted, where light blue denotes reference answers, light red denotes wrong responses and light green denotes right responses.

Method	HoT	MedMCQA	MedQA	PubmedQA	MMLU-Medical						Avg. (Δ)
					Anatomy	Clinical	Biology	Medicine	Genetics	Pro-Med	
Instruction-following Difficulty measured by IFD (Li et al., 2024)											
-w. High-IFD	26.89	55.92	59.23	75.60	66.67	73.96	77.78	61.85	79.00	70.59	48.21
-w. Low-IFD	23.57	56.68	60.64	72.20	68.15	75.09	77.08	64.16	79.00	71.32	47.59$_{\downarrow 0.62}$
Complexity measured by DEITA (Liu et al., 2024b)											
-w. High-Complexity	27.70	56.16	60.64	67.60	65.93	75.85	78.47	63.58	83.00	70.96	47.51
-w. Low-Complexity	31.31	56.83	59.62	71.40	67.41	75.85	77.08	65.90	85.00	72.79	48.86$_{\uparrow 1.35}$
Quality measured by DEITA (Liu et al., 2024b)											
-w. High-Quality$_{deita}$	27.48	56.16	58.29	69.80	68.15	72.45	78.47	64.16	79.00	69.85	47.29
-w. Low-Quality$_{deita}$	32.83	54.36	59.78	71.20	65.93	72.45	76.39	65.32	82.00	73.53	48.46$_{\uparrow 1.17}$
Knowledge Alignment measured by ours											
-w. High-Alignment	31.51	56.39	60.41	72.80	65.93	74.72	79.86	63.58	82.00	74.63	49.09
-w. Low-Alignment	30.23	56.18	59.47	61.60	64.44	71.32	76.39	64.16	82.00	69.49	46.46$_{\downarrow 2.63}$
Knowledge Consistency measured by ours											
-w. High-Consistency	28.80	56.85	60.96	71.80	67.41	75.09	74.31	67.63	82.00	72.79	48.60
-w. Low-Consistency	28.18	56.51	59.54	70.00	64.44	73.21	77.08	64.16	81.00	70.96	47.67$_{\downarrow 0.93}$

Table 9: **Full results of Figure 1, *i.e.*, comparisons of different metrics**. We use the LLaMA-3-8B-Instruct as the base model. "High-*" and "Low-*" refer to the data with higher and lower metric scores, respectively. Red results denote the performance drops of "Low-*" against the "High-*", while green results denote the performance gains.

Method	HoT	MedMCQA	MedQA	PubmedQA	MMLU-Medical						Avg.
					Anatomy	Clinical	Biology	Medicine	Genetics	Pro-Med	
Base	20.87	57.06	60.17	74.80	63.70	71.70	75.00	63.01	81.00	75.00	47.41
Random	29.47	56.75	60.49	68.40	68.89	73.21	78.47	65.32	80.00	73.16	48.05
KDS- *KA*	31.64	57.42	59.23	76.60	65.93	75.47	77.78	65.90	84.00	75.37	**49.83**
-*w/o quality*	32.17	56.87	59.78	75.20	65.93	76.23	77.08	65.90	83.00	73.16	49.60
-*w/o diversity*	31.02	57.02	60.09	76.40	66.67	72.83	77.78	63.58	81.00	74.26	49.54
-*w/o quality&diversity*	31.51	56.39	60.41	72.80	65.93	74.72	79.86	63.58	82.00	74.63	49.09
KDS- *KC*	30.25	57.18	60.57	73.60	67.41	75.47	78.47	64.74	84.00	73.16	**49.25**
-*w/o quality*	31.57	56.83	60.25	74.20	65.93	74.34	79.17	65.90	80.00	74.26	49.35
-*w/o diversity*	30.13	57.04	60.02	75.60	67.41	73.58	72.22	65.32	80.00	73.16	49.12
-*w/o quality&diversity*	28.80	56.85	60.96	71.80	67.41	75.09	74.31	67.63	82.00	72.79	48.60

Table 10: **Full results of Table 2, *i.e.*, ablation study of different strategies**. We use the LLaMA-3-8B-Instruct as the base model. "-*w/o quality*" and "-*w/o diversity*" denote that we remove the quality and diversity strategies, respectively. "-*w/o quality&diversity*" means that we only use the KA/KC metrics for data selection.

Threshold	HoT	MedMCQA	MedQA	PubmedQA	MMLU-Medical						Avg.
					Anatomy	Clinical	Biology	Medicine	Genetics	Pro-Med	
Quality+*Alignment*											
$\tau = 0$	31.51	56.39	60.41	72.80	65.93	74.72	79.86	63.58	82.00	74.63	49.09
$\tau = 1$	31.72	56.20	59.94	75.20	65.19	74.09	79.86	64.16	82.00	72.43	49.34
$\tau = 2$	31.06	56.24	60.33	76.20	66.67	73.72	77.08	65.32	81.00	74.63	49.48
$\tau = 3$	31.02	57.02	60.09	76.40	66.67	72.83	77.78	63.58	81.00	74.26	**49.54**
$\tau = 4$	31.09	56.51	59.58	76.20	63.70	73.34	78.47	64.74	82.00	72.79	49.31
Quality+*Consistency*											
$\tau = 0$	28.80	56.85	60.96	71.80	67.41	75.09	74.31	67.63	82.00	72.79	48.60
$\tau = 1$	30.82	56.87	60.25	72.80	66.67	75.85	73.39	64.16	83.00	73.16	48.91
$\tau = 2$	31.10	57.40	60.09	72.00	67.41	74.34	72.92	64.16	83.00	70.59	48.78
$\tau = 3$	30.13	57.04	60.02	75.60	67.41	73.58	72.22	65.32	80.00	73.16	**49.12**
$\tau = 4$	28.18	56.49	60.33	73.60	64.44	75.85	72.92	67.63	83.00	70.96	48.51
Quality+*Alignment*+***Diversity***											
$\lambda = 0.7$	28.92	57.95	58.21	75.60	65.93	74.34	73.61	63.01	82.00	72.46	48.76
$\lambda = 0.8$	31.62	56.78	58.52	76.20	63.70	75.09	77.78	65.90	84.00	75.37	49.46
$\lambda = 0.85$	29.99	56.44	58.44	76.40	68.15	76.23	75.00	65.32	81.00	74.63	49.11
$\lambda = 0.9$	31.64	57.42	59.23	76.60	65.93	75.47	77.78	65.90	84.00	75.37	**49.83**
$\lambda = 0.95$	31.54	56.76	59.15	76.20	65.93	75.85	78.47	64.16	83.00	75.37	49.57
Quality+*Consistency*+***Diversity***											
$\lambda = 0.7$	29.37	57.06	59.54	71.80	65.93	73.21	77.78	65.90	81.00	70.96	48.37
$\lambda = 0.8$	28.67	56.28	60.57	72.00	65.19	73.58	76.39	64.16	84.00	72.79	48.37
$\lambda = 0.85$	27.66	57.04	60.49	73.20	65.93	74.34	73.61	66.47	82.00	71.69	48.46
$\lambda = 0.9$	30.25	57.18	60.57	73.60	67.41	75.47	78.47	64.74	84.00	73.16	**49.25**
$\lambda = 0.95$	28.12	56.11	60.41	73.60	65.19	72.83	76.39	67.63	82.00	72.06	48.49

Table 11: **Full results of Figure 4 (b) and (c), *i.e.*, parameter analyses of quality threshold τ and diversity threshold λ**. We use the LLaMA-3-8B-Instruct as the base model. "***Quality***+*Alignment*" denotes that we remove the diversity strategy in KDS and "*Quality*+*Alignment*+***Diversity***" refers to the full KDS method.

Method	HoT	MedMCQA	MedQA	PubmedQA	MMLU-Medical						Avg. (Δ)
					Anatomy	Clinical	Biology	Medicine	Genetics	Pro-Med	
DeBERTa-v3-xsmall-mnli											
High-Alignment	31.55	56.73	60.02	71.60	69.63	74.34	77.08	61.85	77.00	75.74	48.75
Low-Alignment	30.71	54.91	60.09	70.80	66.67	74.72	77.08	64.74	80.00	73.16	48.21$_{\downarrow 0.54}$
High-Consistency	28.02	57.11	61.04	72.40	68.37	75.09	77.78	64.74	84.00	72.06	48.71
Low-Consistency	27.96	56.36	61.19	70.00	63.70	75.09	80.56	63.58	82.00	70.56	48.02$_{\downarrow 0.69}$
DeBERTa-v3-base-mnli											
High-Alignment	30.46	56.63	59.15	72.80	68.15	73.96	78.47	65.32	82.00	74.63	48.80
Low-Alignment	28.03	56.18	59.39	66.00	68.89	72.83	77.08	67.05	82.00	71.32	47.13$_{\downarrow 1.67}$
High-Consistency	29.25	55.96	60.09	72.40	68.15	73.21	77.08	61.27	84.00	69.85	48.33
Low-Consistency	28.39	55.75	60.57	70.40	65.19	73.21	78.47	62.43	78.00	71.69	47.77$_{\downarrow 0.56}$
DeBERTa-v3-large-mnli											
High-Alignment	31.51	56.39	60.41	72.80	65.93	74.72	79.86	63.58	82.00	74.63	49.09
Low-Alignment	30.23	56.18	59.47	61.60	64.44	71.32	76.39	64.16	82.00	69.49	46.46$_{\downarrow 2.63}$
High-Consistency	28.80	56.85	60.96	71.80	67.41	75.09	74.31	67.63	82.00	72.79	48.60
Low-Consistency	28.18	56.51	59.54	70.00	64.44	73.21	77.08	64.16	81.00	70.96	47.67$_{\downarrow 0.93}$

Table 12: **Full results of Figure 4 (a), *i.e.*, effect of different NLI models**. We use the LLaMA-3-8B-Instruct as the base model. "*High-Alignment*" and "*High-Consistency*" refer to the data with higher KA and KC scores, respectively, where "*Low-Alignment*" and "*Low-Consistency*" are reversed. Red results denote the performance drops against the higher scores.

Backbone	Method	MMedBench						Score	
		Chinese	English	French	Japanese	Russian	Spanish	Avg.	Δ
LLaMA-3-8B-Instruct	Base	63.69	59.33	**57.01**	36.88	55.86	5red8.46	55.21	-
	Full-SFT	64.27	58.13	54.52	35.00	52.34	56.13	53.40	↓ 1.81
	Random	66.87	59.23	54.83	**38.75**	51.56	56.78	54.67	↓ 0.54
	Alpagasus	66.11	58.84	59.19	33.75	55.47	**59.34**	55.45	↑ 0.24
	IFD	66.02	59.23	53.58	33.12	54.69	57.62	54.04	↓ 1.17
	DEITA	63.78	59.07	55.76	32.50	57.42	58.21	54.46	↓ 0.75
	3DS	63.81	58.52	55.14	32.50	60.55	57.66	54.70	↓ 0.51
	KDS- *KA*	65.00	59.15	56.70	37.50	**64.45**	58.21	**56.84**	↑ **1.63**
	KDS- *KC*	**67.28**	**59.54**	56.39	38.12	55.08	58.02	55.74	↑ 0.53
Qwen-2.5-7B-Instruct	Base	82.90	56.64	65.11	53.75	68.75	64.70	65.31	-
	Full-SFT	85.43	59.64	63.86	51.25	71.09	66.92	66.37	↑ 1.05
	Random	86.72	58.88	69.16	52.50	72.66	66.51	67.74	↑ 2.43
	Alpagasus	86.28	58.84	69.47	57.50	71.88	66.48	68.41	↑ 3.10
	IFD	86.22	58.92	65.11	54.37	71.88	66.59	67.18	↑ 1.87
	DEITA	85.70	58.60	69.47	54.37	71.88	66.78	67.80	↑ 2.49
	3DS	84.73	58.37	69.78	53.12	72.66	66.59	67.54	↑ 2.23
	KDS- *KA*	85.84	59.31	68.85	**60.00**	**75.00**	**67.07**	69.35	↑ 4.04
	KDS- *KC*	**86.81**	**60.57**	**70.09**	**60.00**	72.66	66.74	**69.48**	↑ **4.17**
Qwen-2.5-14B-Instruct	Base	87.42	66.30	74.45	68.13	**73.83**	71.55	73.61	-
	Full-SFT	89.03	68.81	74.77	66.25	70.70	73.56	73.85	↑ 0.24
	Random	89.57	68.34	74.77	66.25	71.48	73.65	74.01	↑ 0.40
	Alpagasus	88.94	68.34	**76.01**	66.87	71.48	72.43	74.01	↑ 0.40
	IFD	88.88	67.95	75.70	67.50	72.27	73.12	74.24	↑ 0.63
	DEITA	89.05	66.69	74.77	67.50	72.66	72.54	73.87	↑ 0.26
	3DS	88.82	68.19	75.70	68.13	72.66	73.34	74.47	↑ 0.86
	KDS- *KA*	89.99	68.42	75.70	68.75	73.05	**74.14**	**75.01**	↑ **1.40**
	KDS- *KC*	**90.02**	**68.89**	75.08	**69.37**	71.88	73.81	74.84	↑ 1.23

Table 13: **Full results of Figure 9, *i.e.*, performance of MMedBench** (Qiu et al., 2024). Green results mean the performance gains against the base model, and red results mean the performance drops. The best results are in **bold**.

Method	HoT	MedMCQA	MedQA	PubmedQA	MMLU-Medical						Avg.
					Anatomy	Clinical	Biology	Medicine	Genetics	Pro-Med	
1K Samples											
Random	27.93	56.30	60.25	69.20	66.67	76.60	77.08	67.63	83.00	74.26	47.98
IFD	27.27	57.21	60.80	71.00	68.15	75.09	77.17	63.58	83.00	75.37	48.33
Alpagasus	26.60	57.95	60.80	69.20	63.70	75.09	79.17	66.47	81.00	75.74	48.01
DEITA	26.94	58.16	60.80	68.00	60.74	73.21	76.39	63.01	81.00	73.90	47.55
3DS	26.55	57.42	60.64	69.20	62.22	75.09	77.08	64.16	81.00	72.43	47.63
KDS- *KA*	26.61	57.23	59.62	74.20	65.93	74.34	75.69	65.32	83.00	73.53	<u>48.44</u>
KDS- *KC*	28.68	56.90	60.41	71.60	70.37	75.47	77.78	67.63	86.00	72.79	**48.77**
3K Samples											
Random	28.54	57.78	61.35	69.40	68.15	71.70	74.31	64.16	81.00	73.53	48.20
IFD	27.51	57.02	60.72	75.20	66.67	73.58	77.78	62.43	80.00	72.43	48.77
Alpagasus	27.70	57.23	59.94	72.80	62.22	73.58	77.08	66.47	81.00	72.79	48.31
DEITA	27.51	55.69	59.62	71.80	65.93	74.72	77.78	63.58	80.00	71.69	47.82
3DS	27.82	56.71	60.72	70.40	65.93	73.96	76.39	62.43	81.00	71.69	47.93
KDS- *KA*	28.66	57.90	60.09	76.20	65.19	73.96	76.39	64.16	82.00	71.69	<u>49.18</u>
KDS- *KC*	29.73	57.14	61.04	73.20	67.41	75.09	79.86	64.74	85.00	73.53	**49.23**
5K Samples											
Random	29.47	56.75	60.49	68.40	68.89	73.21	78.47	65.32	80.00	73.16	48.05
Alpagasus	27.78	56.90	60.33	71.40	65.93	73.96	77.08	65.32	81.00	71.69	48.15
IFD	26.89	55.92	59.23	75.60	66.67	73.96	77.78	61.85	79.00	70.59	48.21
DEITA	28.69	55.10	58.92	73.60	69.63	74.47	78.47	60.69	78.00	72.06	48.01
3DS	27.86	55.32	59.15	72.80	67.41	75.09	78.47	63.58	80.00	72.43	47.99
KDS- *KA*	31.64	57.42	59.23	76.60	65.93	75.47	77.78	65.90	84.00	75.37	**49.83**
KDS- *KC*	30.25	57.18	60.57	73.60	67.41	75.47	78.47	64.74	84.00	73.16	<u>49.25</u>
10K Samples											
Random	30.49	56.21	59.09	70.20	68.15	73.58	75.69	61.27	80.00	70.96	47.93
IFD	27.27	54.96	58.99	74.00	65.93	75.85	79.17	61.27	78.00	72.43	47.89
Alpagasus	27.71	55.41	59.54	71.00	67.41	74.72	80.56	61.85	80.00	70.59	47.70
DEITA	28.11	56.71	59.86	73.80	68.15	73.96	76.39	62.43	79.00	69.12	48.33
3DS	28.80	55.44	58.76	68.40	65.19	73.21	76.39	61.85	78.00	66.18	46.92
KDS- *KA*	30.76	56.59	58.60	76.60	65.93	75.09	76.39	63.58	79.00	69.49	**49.02**
KDS- *KC*	30.58	56.16	59.62	72.80	66.67	73.58	76.39	65.90	79.00	71.32	<u>48.55</u>
20K Samples											
Random	30.03	55.35	58.25	71.00	62.96	72.08	77.17	60.12	79.00	67.85	47.42
IFD	28.13	54.70	59.54	74.80	65.93	73.58	75.00	61.85	79.00	68.75	47.98
Alpagasus	28.31	55.44	59.47	72.60	68.15	71.70	78.47	63.01	78.00	70.59	47.91
DEITA	28.36	55.82	57.89	74.20	66.67	73.21	74.31	60.12	79.00	69.85	47.80
3DS	29.58	55.13	59.07	71.60	63.70	72.45	77.08	60.69	75.00	69.49	47.52
KDS- *KA*	30.18	56.49	59.31	75.80	68.89	72.08	78.47	63.01	78.00	68.38	**48.88**
KDS- *KC*	29.85	55.20	60.02	72.60	66.67	70.94	75.00	65.32	81.00	68.01	<u>48.14</u>

Table 14: **Full results of Figure 5, *i.e.*, analysis of data scales**. LLaMA-3-8B-Instruct is used as the base model and tuned with different numbers of data. The best average results are in **bold**, and the second-best ones are <u>underlined</u>.