3DS: Medical Domain Adaptation of LLMs via Decomposed Difficulty-based Data Selection

Anonymous ACL submission

Abstract

Large Language Models (LLMs) excel in general language tasks, motivating their adaptation to specialized domains such as healthcare. Effective domain adaptation typically involves supervised fine-tuning (SFT) on carefully selected instruction-tuning data. Current data selection methods adopt a **data-centric** approach, relying on external annotations and heuristics to identify external defined high-quality and challenging data. Our exploratory experiments highlight this approach *fails to improve model's domain performance, due to misalignment between selected data and the model's knowledge distribution*. To tackle this, we propose Decomposed Difficulty-based Data Selection (3**DS**), a two-stage **model-centric** data selection framework that aligns data selection with the model's distribution. 3DS employs a *Prompt-Driven Data Selection* to filter out noisy data based on the model's knowledge via explicit alignment in Stage#1, then adopts a *Decomposed Difficulty-based Data Selection* to guide selection via three novel data difficulty metrics, including *Instruction Understanding*, *Response Confidence*, and *Response Correctness* in Stage#2. These metrics are enhanced by an *attention-based importance weighting mechanism* for accurate calibration. Extensive experiments in the healthcare domain show 3DS outperforms existing methods by over 2.97% accuracy, with additional validation in the law domain confirming its generalization ability. Our dataset and code are open-sourced at `https://anonymous.4open.science/r/3DS-E67F`.

1 Introduction

Large Language Models (LLMs) such as proprietary GPT-4 (OpenAI, 2023), open-sourced LLaMA (Touvron et al., 2023) and Qwen (Bai et al., 2023), have demonstrated remarkable capabilities in language understanding and generation. Encouraged by their successes, there is growing interest in leveraging LLMs in specialized domains like healthcare, where domain-specific abilities are required (Sanaei et al., 2023; Harris, 2023; Waisberg et al., 2023) for essential tasks like diagnosis (Panagoulias et al., 2024; Ullah et al., 2024) and treatment recommendations (Wilhelm et al., 2023; Nwachukwu et al., 2024). To address this, many existing works (Wang et al., 2023a; Zhang et al., 2023; Yang et al., 2023b; Zhu et al., 2023a; Pal and Sankarasubbu, 2023) have tried to adapt LLMs to the medical domain by training on large-scale healthcare-specific datasets.

An essential step in adapting general LLMs to specialized domains is Supervised Fine-Tuning (SFT) on domain instruction-tuning datasets. However, large-scale, unfiltered domain datasets aggregated from multiple sources often include *noise*. Directly utilizing such data can disrupt learning (Wang et al., 2023d, 2024a), hinder the identification of knowledge gaps (Havrilla and Iyer, 2024), and increase the risk of overfitting (Budach et al., 2022; Wang et al., 2024b), yielding poor performance. Recent findings (Zhou et al., 2024) suggest that a *small but carefully selected high-quality* dataset can effectively enhance model's alignment with human instructions and elicit its abilities in the desired direction, highlighting the necessity of rigorous data selection for domain adaptation fine-tuning. This presents a critical challenge in fine-tuning general LLMs to specialized domains:

How to identify and select domain instruction-tuning data that is most suitable for the target LLM to optimally elicit its domain-specific abilities?

Previous data selection methods predominantly adopt a **data-centric** perspective, typically focusing on two dimensions: *quality* and *difficulty*. For quality, existing methods rely on powerful external models or manual rules to identify "high-quality" samples (Liu et al., 2023; Ji et al., 2023; Song et al., 2024). They treat quality as a model-agnostic, intrinsic data property, assuming the assessments are universally applicable. However, LLMs differ substantially in architectures and training corpora, which shape their distinct internal knowledge distributions. External "high-quality" data may still introduce redundancy or conflicting information that impede learning. For difficulty, methods typically prioritize the most challenging samples based on heuristic metrics (Li et al., 2024b,a). However, recent studies (Gekhman et al., 2024; Ren et al., 2024) have revealed that fine-tuning LLMs on data beyond their pre-trained knowledge distribution, particularly unfamiliar content, can lead to severe hallucinations, which underscores the potential risk of selecting hardest samples. A common limitation of these methods is their lack of consideration for model-specific compatibility, both external "high-quality" data or most challenging data could be misaligned with the model's distribution and lead to suboptimal results.

Motivated by this gap, we propose a new hypothesis: *data selection should be* ***model-centric****, tailored to align with the model's knowledge distribution*.

To validate this hypothesis, we conduct a pilot study guided by two research questions: **RQ#1.** Is model-centric quality selection more effective than external quality scoring? **RQ#2.** Is model-centric difficulty selection more effective than prioritizing the hardest samples? The results demonstrate that model-centric data selection, which relies on the target model's own assessment of data quality and selection of appropriately difficult data, consistently outperforms selection guided by external criteria.

While these findings highlight the importance of model-centric data selection, its practical application still faces substantial challenges:

❶ **Challenge#1. How to identify high-quality data based on the model's knowledge distribution?** Redundant knowledge that the model already possesses and conflicting information that goes against the model's knowledge hinders learning (Ren et al., 2024; Gekhman et al., 2024). Selecting high-quality data based on the model's knowledge distribution is thus necessary, but inherently challenging due to the complexity and opacity of LLMs

❷ **Challenge#2. How to properly balance the selected data difficulty with the model's learning capacity?** Overly simplistic data wastes training resources and may cause overfitting, while excessively complex data can overwhelm the model, impeding effective learning (Kang et al., 2024; Lin et al.). Accurately assessing difficulty based on the model's distribution to guide selection is thus crucial. However, there isn't a effective metric to comprehensively measure the model's knowledge state and its ability to handle complex data.

To tackle these challenges, we propose **D**ecomposed **D**ifficulty-based **D**ata **S**election (3**DS**), a two-stage **model-centric** data selection framework which aligns data selection with the model's distribution to optimize domain fine-tuning. For **Challenge#1**, we propose *Prompt-Driven Data Selection via Explicit Alignment*, leveraging the target model's own evaluations to explicitly select high-quality data, ensuring that the remaining data lies within the model's knowledge distribution. For **Challenge#2**, inspired by the general human problem-solving process (Polya and Pólya, 2014; OECD, 2014)—understanding the problem, building confidence, and producing a solution, we propose novel *Decomposed Difficulty-based Data Selection via Implicit Alignment*, extending traditional perplexity (PPL) measures with three difficulty metrics: Instruction Understanding Difficulty, Response Confidence Difficulty, and Response Correctness Difficulty. Furthermore, an *attention-based importance weighting mechanism* captures token-level importance and calibrates difficulty calculations. In summary, our contributions are:

- We introduce 3DS, a two-stage model-centric data selection framework, aligning training data with the model's knowledge distribution, optimizing domain adaptation fine-tuning.
- We propose a novel difficulty decomposition strategy, employing fine-grained metrics: Instruction Understanding, Response Confidence, and Response Correctness, for accurate data difficulty quantification tailored to domain-specific fine-tuning.
- Comprehensive experiments on Chinese medical datasets demonstrate that 3DS outperforms existing methods, significantly boosting LLMs performance. Additional experiments on law domain also showcase 3DS's generalization ability.
- We have open-sourced a carefully curated Chinese medical dataset, including medical dialogues and domain-specific instructions, to support further research in healthcare-oriented LLM.

2 Importance of Model-Centric Selection

In this section, we empirically investigate the importance of model-centric data selection by studying the following two research questions:

- **RQ#1**. *Is model-centric quality selection more effective than external quality scoring?*
- **RQ#2**. *Is model-centric difficulty selection more effective than prioritizing the objectively hardest samples?*

2.1 Experimental Setup

In both investigations, we utilized two models: DeepSeek-R1 (Guo et al., 2025), an external model regarded as strong and capable, which is expected to provide reliable data evaluation, and LLaMA-3-8B-Instruct (Grattafiori et al., 2024), the target model intended for domain fine-tuning. We utilized a large-scale Chinese medical instruction-tuning dataset and designed tailored prompts to assess data quality and difficulty (see Appendix J.1 and J.2).

2.2 Model-Centric *vs.* External Quality Selection

To answer **RQ#1**, we prompted both the external model and the target model to assess data quality based on their knowledge. From data scored above a predefined threshold by each model, we randomly selected 5K samples and fine-tuned LLaMA-3-8B-Instruct on each subset. Performance evaluated on two Chinese medical multiple-choice question benchmarks (Zeng, 2023; Wang et al., 2023c) is shown in Table 1.
Surprisingly, fine-tuning on high-quality data selected by the strong DeepSeek-R1 led to performance degradation of LLaMA-3-8B-Instruct, while data selected by LLaMA-3-8B-Instruct itself significantly improved its performance. This discrepancy likely stems from a misalignment between the external quality assessment and the target model's inherent knowledge distribution. Based on this, we derive our first key observation:

Observation I: *Model-centric quality selection yields better performance than external quality scoring.*

2.3 Model-Centric *vs.* External Difficulty Selection

To answer **RQ#2**, we evaluated the commonly held assumption that training on the most challenging data improves model abilities. Similar to the previous investigation, we prompted DeepSeek-R1 and LLaMA-3-8B-Instruct to score data difficulty based on their knowledge. The dataset was partitioned into Easy, Medium, and Hard subsets, according to difficulty scores from each model. We then fine-tuned LLaMA-3-8B-Instruct on randomly selected 5k samples from each subset, and compared their performance across medical benchmarks, with results shown in Table 2.

Across all experiments, fine-tuning on *Easy* and *Medium* subsets consistently outperformed training on *Hard* subset, with Medium subset yielding more stable improvements, indicating that overly difficult data, likely exceeding model's knowledge, adversely impacts learning, while overly simple data also fails to sufficiently benefit fine-tuning. Additionally, difficulty assessments from LLaMA-3-8B-Instruct itself consistently led to better results compared to external evaluations by DeepSeek-R1, which validates the necessity of model-centric difficulty evaluation and selection. This motivates our second and third key observations:

Observation II: *Difficulty scoring based on the target model yields more reliable performance than scores provided by an external model.*

Observation III: *Moderately difficult data leads to more stable and effective performance improvements.*

Data	Annotator	CMB-Exam	MMCU-Med
Original	N/A	41.72	46.47
High-quality	DeepSeek-R1 LLaMA3-8B	39.70 **43.71**	42.46 **47.57**

Table 1: High-quality Data Selection Results (%). Improvements over the original model are in **bold**.

Data	Annotator	CMB-Exam	MMCU-Med
Original	N/A	41.72	46.47
Easy	DeepSeek-R1 LLaMA3-8B	41.03 41.53	45.76 **48.00**
Medium	DeepSeek-R1 LLaMA3-8B	**41.76** **41.75**	45.26 **46.72**
Hard	DeepSeek-R1 LLaMA3-8B	40.50 40.62	44.06 45.23

Table 2: Difficult Data Selection Results (%).

2.4 Conclusion and Motivation

Both investigations lead to a key conclusion: effective data selection for domain adaptation fine-tuning requires alignment with the target model's knowledge distribution. External assessed high-quality data may not suit the target model, and excessively difficult data may introduce unfamiliar, out-of-distribution content, causing suboptimal outcomes.

Motivated by these observations, we propose to shift from conventional **data-centric** selection strategies toward a **model-centric** approach. Specifically, data selection should be guided by the target model, ensuring that the selected data are considered as high-quality(addressing **Observation I**) and appropriately challenging by the target model(addressing **Observation II and III**), thus achieving close alignment with its knowledge distribution and learning capacity. Building on this insight, we propose our novel model-centric framework 3DS in the following sections.

3 Methodology

Task Formulation We formally define the Data Selection for Domain Adaptation Fine-tuning task. Let:

- M_θ denotes the target model to be fine-tuned, which is a pre-trained and generally fine-tuned LLM (e.g., LLaMA-chat) parameterized by θ.
- $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$ denotes the full domain-specific dataset where each sample $x^{(i)} =< Q^{(i)}, A^{(i)} >$ consists of instruction $Q^{(i)} = \{q_1^{(i)}, q_2^{(i)}, \ldots, q_m^{(i)}\}$, and response $A^{(i)} = \{a_1^{(i)}, a_2^{(i)}, \ldots, a_n^{(i)}\}$. Here $q_m^{(i)}, a_n^{(i)}$ denote individual tokens within the instruction and response sets, respectively.
- $k \in \mathbb{N}^+$ denotes a fixed data budget, where $k \ll |\mathcal{X}|$.

The task is to identify an optimal subset $\mathcal{S}^* \subseteq \mathcal{X}$ that maximizes the target domain performance of the fine-tuned model M'_θ, formally:

$$\mathcal{S}^* = \underset{S\subseteq\mathcal{X},|S|=k}{\arg\max}\ \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{test}}}\left[\mathcal{P}(M_{\theta'}(x;\mathcal{S}),y)\right], \quad (1)$$

where $\mathcal{D}_{test}$ is the target domain test distribution containing diverse multiple domain tasks; $\mathcal{P} : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ is the performance metric (e.g., accuracy, BLEU, ROUGE), and $M_{\theta'}$ is M_θ fine-tuned on $\mathcal{S}$, i.e., $\theta' = \theta - \eta\nabla_\theta \sum_{x\in S} \mathcal{L}(M_\theta(x), x)$, with learning rate η and loss function $\mathcal{L}$.

3.1 Stage#1: Prompt-Driven Data Selection via Explicit Alignment

The first stage of 3DS is to identify high-quality data based on the model's knowledge. As illustrated in Figure 1, a quality-rating prompt, detailed in Appendix J.1, is used to instruct M_θ to score data quality based on its inner knowledge to explicitly align data, filtering out noise from the original large-scale dataset to avoid conflicting information. After obtaining model-generated scores, samples with scores exceeding a predefined threshold δ are retained for the next selection.

3.2 Stage#2: Decomposed Difficulty-based Data Selection via Implicit Alignment

The second stage of 3DS is to analyze data difficulty via implicit distribution modeling of M_θ, thereby balancing the selected data difficulty with the model's learning capacity. To achieve this, we employ a fine-grained evaluation for data difficulty.

Inspired by the general problem-solving process (Polya and Pólya, 2014; OECD, 2014)—understanding the problem, building confidence, and producing a solution—we decompose data difficulty into three key components to reflect the model's understanding: (1) *Instruction Understanding Difficulty* measures whether the model comprehends the instruction. (2) *Response Confidence Difficulty* measures the model's

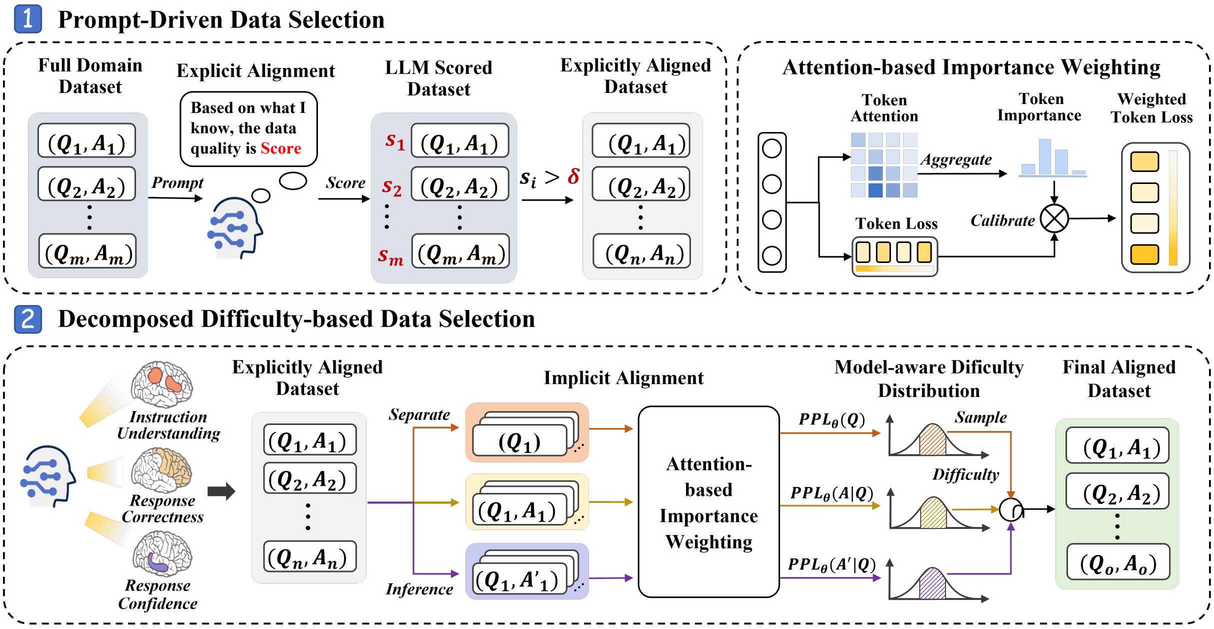


Figure 1: 3**DS framework**. **Stage#1:** Prompt-Driven Data Selection select high-quality data via explicit alignment. **Stage#2:** Decomposed Difficulty-based Data Selection decomposes data difficulty via modeling LLM's implicit distribution and filters data. Attention-based importance weighting calibrates difficulty calculation.

confidence in its response. (3) *Response Correctness Difficulty* measures whether the model can generate a response that accurately matches the reference answer. To enhance the precision of difficulty calculations, we incorporate an *attention-based importance weighting mechanism* that calibrates difficulty by accounting for the varying semantic significance of output tokens. We now detail the quantification of these decomposed difficulties and the corresponding selection strategy.

(1) Instruction Understanding Difficulty. Challenging data often comes with complex instructions. In specialized domains like healthcare, instructions may contain intricate terminologies, making instruction comprehension a key factor of data difficulty. To capture this, we introduce Instruction Understanding Difficulty. Previous research (Gonen et al., 2023) shows that lower model perplexity over a prompt correlates with better understanding and performance. Building on this insight, we further recognize that perplexity inherently captures the predictive uncertainty from model's distribution. Consequently, we employ model perplexity as a measure to quantify data difficulty from the model's perspective. Formally, for a model M_θ, given a data sample $x =< Q, A >$ with instruction $Q = \{q_1, q_2, \dots q_m\}$, its Instruction Understanding Difficulty is defined as:

$$\begin{aligned}\mathrm{D1}_\theta(x) &= \mathrm{PPL}_\theta(Q)\\ &= \exp\left(-\frac{1}{m}\sum_{i=1}^{m}\log P_\theta(q_i|q_1, q_2, \dots, q_{i-1})\right),\end{aligned}\tag{2}$$

where $P_\theta(q_i|q_1, q_2, \dots, q_{i-1})$ represents the probability M_θ generates the i-th token in instruction Q given the preceding tokens. Higher perplexity indicates greater difficulty for the model to comprehend the instruction.

(2) Response Confidence Difficulty. When encountering challenging data, models often struggle to provide a confident response. This uncertainty arises from its inability to handle the task and determine the most appropriate response, similar to human learners (Preheim et al., 2023), which indicates high data difficulty. To quantify this difficulty, we introduce Response Confidence Difficulty, measured by the model's conditional perplexity when generating a response given the instruction. Formally, for a model M_θ, given a data sample $x =< Q, A >$ with instruction Q is and model-generated response $A' = \{a'_1, a'_2 \dots, a'_{n'}\}$ based on Q, its Response Confidence Difficulty is defined as:

$$\begin{aligned}\mathrm{D2}_\theta(x) &= \mathrm{PPL}_\theta(A'|Q)\\ &= \exp\left(-\frac{1}{n'}\sum_{j=1}^{n'}\log P_\theta(a'_j|a'_1, a'_2, \dots, a'_{j-1}, Q)\right).\end{aligned}\tag{3}$$

Higher conditional perplexity indicates greater uncertainty in the model's distribution and greater difficulty for the model to provide a confident answer.

(3) Response Correctness Difficulty. For instruction-tuning data with reference answers, it is essential to assess the model's ability to generate correct responses to assess data difficulty. We introduce Response Correctness Difficulty, measured by the model's conditional perplexity when generating the reference answer $A = \{a_1, a_2 \dots, a_n\}$ given instruction Q.

$$\begin{aligned}\mathrm{D3}_\theta(x) &= \mathrm{PPL}_\theta(A|Q)\\ &= \exp\left(-\frac{1}{n}\sum_{j=1}^{n}\log P_\theta(a_j|a_1, a_2, \dots, a_{j-1}, Q)\right).\end{aligned}\tag{4}$$

Higher conditional perplexity indicates greater difficulty in producing the correct response, suggesting the sample poses more challenge for the model.

Attention-based importance weighting mechanism. Response Confidence and Response Correctness Difficulties rely on evaluating the uncertainty inherent in the model's generation process. While conditional perplexity serves as an effective proxy, it treats all tokens equally, disregarding their varying semantic importance. While key tokens significantly influence the meaning and correctness of a response, trivial tokens like conjunctions may exhibit high uncertainty without substantially influencing semantics. This can lead to inaccurate data difficulty assessments. To address this, inspired by Su et al. (2024), we introduce an attention-based importance weighting mechanism that adjusts token's uncertainty contributions by weighting based on their semantic importance. We argue that critical tokens are those playing a pivotal role in guiding subsequent generations. Therefore, we derive importance scores from the model's internal attention mechanism. Specifically, for a token sequence $s = \{t_1, t_2, \ldots, t_i, \ldots, t_n\}$, when a transformer-based LLM generates token $t_j (i < j)$, it computes the attention weight A_{ji} by applying a softmax function to the dot product of the query vector q_j and the key vector k_i:

$$A_{ji} = (q_j \cdot k_i)/\sqrt{d_k}, \tag{5}$$

where d_k is the dimension of k_i. A_{ji} represents the attention the model pays to token t_i when generating token t_j, reflecting the importance of t_i. We define the importance score of token t_i as the aggregated attention weight it receives from all subsequent tokens:

$$\mathrm{I}(t_i) = \underset{j>i}{\texttt{Aggregate}}\ (A_{ji}). \tag{6}$$

We use mean aggregation to compute token importance scores. Using these scores, Response Confidence and Response Correctness Difficulties are refined as:

$$\begin{aligned} \texttt{Atten-D2}_\theta(x) &= \texttt{weightedPPL}_\theta(A'|Q) \\ &= \exp\left(-\frac{\sum_{j=1}^{n'} \mathrm{I}(t_j)\cdot\phi}{\sum_{j=1}^{n'} \mathrm{I}(t_j)}\right), \\ \phi &= \log P_\theta(a'_j|a'_1, a'_2, \ldots, a'_{j-1}, Q), \end{aligned} \tag{7}$$

$$\begin{aligned} \texttt{Atten-D3}_\theta(x) &= \texttt{weightedPPL}_\theta(A|Q) \\ &= \exp\left(-\frac{\sum_{j=1}^{n} \mathrm{I}(t_j)\cdot\phi'}{\sum_{j=1}^{n} \mathrm{I}(t_j)}\right), \\ \phi' &= \log P_\theta(a_j|a_1, a_2, \ldots, a_{j-1}, Q). \end{aligned} \tag{8}$$

By integrating attention-based importance weights, this mechanism prioritizes tokens crucial for semantic correctness and clarity, offering a more accurate estimation of model uncertainty and data difficulty.

Selection Strategy based on Decomposed Difficulty. Based on the decomposed data difficulties, 3DS identifies samples whose difficulty metrics fall within a predefined middle range, discarding either trivially easy or overly complex data, focusing on moderately challenging samples that match the model's learning capacity. K-Center sampling (introduced in Appendix C) based on instruction embeddings is then applied on this subset to enhance data diversity, reducing the risk of overfitting on highly similar samples.

3.3 Model-Centric Data Selection Framework

The overall architecture of our model-centric data selection framework is illustrated in Figure 1. Pseudo codes of the process are shown in Appendix A.

4 Main Experiments

4.1 Experimental Setup

Training dataset. For medical domain adaptation fine-tuning, we construct a comprehensive medical instruction-tuning dataset of diversity and abundance. The dataset comprises over 1.9 M samples, with its statistics provided in Table 7 and data construction details introduced in Appendix B. We have released this complete training dataset to support further research.

Evaluation datasets. We assess fine-tuned models on diverse medical test datasets: two multi-task, multiple-choice datasets, MMCU-Med (Zeng, 2023) and CMB-Exam (Wang et al., 2023c), and an open Q&A dataset, CMB-Clin (Wang et al., 2023c). Data statistics are provided in Table 8. MMCU-Medical and CMB-Exam, consisting of medical exam questions, assess the model's reasoning and medical knowledge application abilities with accuracy as the metric. CMB-clin, comprising of patient record analysis tasks, assesses the model's complex medical analysis ability, with BLEU-1, BLEU-4 and ROUGE as the metric (detailed in Appendix F). Together, these datasets provide a comprehensive evaluation of the model's proficiency in the medical domain.

Models. Experiments are conducted on instruct models of varying architectures and parameter sizes: Baichuan2-13B-Chat (Yang et al., 2023a), Qwen1.5-7B-Instruct, Qwen2.5-7B-Instruct (Bai et al., 2023) and LLaMA3-8B-Instruct (Touvron et al., 2023).

Baselines. We compare 3DS with a series of LLM fine-tuning data selection strategies. (1) **Base** directly tests instruct models without further fine-tuning. (2) **Full-Sft** fine-tunes models on the full training dataset. (3) **Random Selection** randomly selects data. (4) **Alpagasus** (Chen et al., 2023a) utilizes GPT-4 to identify high-quality data. (5) **DEITA** (Liu et al., 2023) trains quality and complexity scorers and selects data according to their judgments (6) **MoDS** (Du et al., 2023) filters high-quality data via a reward model, and selects data necessary for model learning through training and inference processes. (7) **IFD** (Li et al., 2024a,b) designs instruction following difficulty metric based on the ground truth output loss with or without inputs. (8) **LESS** (Xia et al., 2024) searches for training data

Method	LLM Turbo	Baichuan2-13B-Chat		Qwen1.5-7B-Instruct		Qwen2.5-7B-Instruct		LLaMA3-8B-Instruct		Avg.
	Dataset	CMB-Exam	MMCU-Med	CMB-Exam	MMCU-Med	CMB-Exam	MMCU-Med	CMB-Exam	MMCU-Med	
Baselines	Base	46.67	47.11	59.80	64.24	78.28	83.43	41.72	46.47	58.47
	Full-Sft	40.38	37.90	48.05	47.53	71.04	75.49	40.85	46.72	51.00
	Random	44.07	47.61	61.81	65.10	75.92	82.41	41.54	45.23	57.96
	Alpagasus	42.24	43.56	55.67	58.74	69.90	78.08	41.60	45.26	54.38
	DEITA	46.78	49.88	45.33	44.09	74.07	81.59	41.31	45.80	53.60
	MoDS	47.25	50.37	61.09	64.67	76.31	82.23	39.25	42.53	57.96
	IFD	46.44	50.08	**62.06**	65.37	78.17	84.57	38.25	40.48	58.18
	LESS	45.79	51.01	60.74	64.85	78.83	83.20	41.80	44.63	58.86
Ours	**3DS**	**47.37**	**51.08**	61.96	**66.09**	**79.06**	**85.70**	**43.95**	**49.70**	**60.61**
***Performance Gain ↑**		0.70	3.97	2.16	1.85	0.78	2.27	2.23	3.23	2.14

Table 3: Performance comparison (%) on *CMB-Exam*, *MMCU-Medical* of EM score. The best performance is highlighted in **bold**. Performance gains are measured against the base model.

Method	LLM Turbo	Baichuan2-13B-Chat			Qwen1.5-7B-Instruct			Qwen2.5-7B-Instruct			LLaMA3-8B-Instruct			Avg.
	Metric	BLEU-1	BLEU-4	ROUGE	BLEU-1	BLEU-4	ROUGE	BLEU-1	BLEU-4	ROUGE	BLEU-1	BLEU-4	ROUGE	
Baselines	Base	11.15	21.02	14.08	16.17	32.03	16.31	21.87	64.11	36.74	5.06	35.09	10.40	23.67
	Full-Sft	7.19	16.33	11.70	6.68	16.61	9.62	16.72	36.52	19.84	2.80	6.87	6.58	13.12
	Random	12.14	25.95	14.75	16.09	34.45	16.19	16.49	33.68	17.89	9.01	25.49	12.14	19.52
	Alpagasus	10.16	20.42	12.58	14.48	31.63	14.77	16.85	35.74	18.77	8.66	22.51	12.36	18.24
	DEITA	19.42	42.07	19.32	18.92	42.93	20.32	21.71	49.33	23.40	9.91	23.33	13.86	25.38
	MoDS	22.43	51.02	22.85	17.61	39.19	19.93	18.83	41.31	21.45	12.38	29.74	15.33	26.01
	IFD	21.44	51.73	24.94	19.24	43.10	21.08	18.07	39.16	20.28	10.59	29.32	14.83	26.15
	LESS	13.27	29.20	16.40	17.48	38.88	17.58	19.08	45.20	22.42	11.82	31.98	15.55	23.24
Ours	**3DS**	**24.15**	**63.51**	**31.50**	**24.40**	**60.32**	**28.07**	**22.05**	**64.95**	**37.11**	**12.52**	**36.88**	**17.09**	**35.21**
***Performance Gain ↑**		13.00	42.49	17.42	9.45	29.49	11.92	0.18	0.84	0.37	7.46	1.79	6.69	11.54

Table 4: Performance comparison (%) on *CMB-Clin*.

similar to target task examples through low-rank gradient similarity. The implementation details are introduced in Appendix D.

Implementations. The selection data budget is 5K samples. In 3DS, the Prompt-Driven Data Selection stage retains samples with a quality score ≥ 90. In the subsequent Decomposed Difficulty-based Data Selection stage, difficulty thresholds are determined via experiments on the CMB hold-out validation set. Specifically, for Baichuan2-13B-Chat, the thresholds are set to 15% and 65%; for Qwen1.5-7B-Chat and Qwen2.5-7B-Chat, 25% and 75%; and 40% and 90% for LLaMA3-8B-Instruct. More implementation details are introduced in Appendix E.

4.2 Main Results

Experiment results are shown in Table 3 and Table 4. We summarize our findings below.

Data selection is necessary for LLM domain adaptation fine-tuning. We observe that fine-tuning LLMs on the full 1.9 million dataset (Full-SFT) leads to drastic performance drops across three benchmarks. This suggests that domain datasets directly collected from public resources contain significant noise that hinders model learning, highlighting the necessity of data selection.

3DS effectively enhances LLM's diverse domain abilities, significantly outperforming baselines. As shown in Table 3 and Table 4, across various benchmarks and LLM backbones, 3DS generally achieves the highest accuracy, outperforming the backbones and strong data selection baselines. On medical exam datasets, it improves base model performance by up to 8.43% (on MMCU-Med for `Baichuan2-13B-Chat`), and exceeds the best baseline an average of 2.97%, greatly enhancing the model's medical knowledge application abilities. On the open Q&A CMB-clin, models fine-tuned with 3DS significantly outperforms all baselines by a large margin, exhibiting superior medical analysis ability. To more comprehensively analyze model's domain performance, for CMB-Clin, we also conduct a pair-wise comparison using GPT-4o as the judge, detailed in Appendix G.1. Both the quantitative and qualitative evidence demonstrate that 3DS boosts the model's multi-faceted domain abilities.

In contrast, methods relying on external, model-agnostic data evaluations, such as Alpagasus and DEITA, often lead to performance declines, especially on Qwen models. This further validates our previous conclusion that misalignment between selected data and the model hinders learning. Baselines MoDS and IFD show relatively strong results due to their considerations of model distribution and data difficulty. However, their selection on the most challenging data also proves inefficient, as they only bring marginal improvements across tasks and even underperform the backbone and random selection on LLaMA3-8B-Instruct. Baseline LESS, which aims to enhance performance on one specific downstream task, fails to generalize to domain adaptation fine-tuning where diverse abilities need to be improved, leading to performance degradation on MMCU-Medical for Baichuan2-13B-Chat.

3DS exhibits strong generalization ability. 3DS's consistent performance gains across backbones and benchmarks highlight its generalization ability to adapt to different models and domain tasks. To further validate the practicality of 3DS, we compare models fine-tuned using 3DS-with existing medical LLMs, with results shown in Appendix G.2.

4.3 Ablation Studies

3DS is composed of two stages, and in Stage#2, three difficulty metrics are proposed. To validate the effectiveness of each component, we conduct comprehensive ablation studies. Without loss of generality, experiments are done on Baichuan2-13B-Chat and Qwen1.5-7B-Instruct. Main results are shown in Table 5, and additional metrics are in Appendix G.3.

Ablation on stages. To evaluate the contributions of each stage in 3DS, we compare: **(1) removing Stage#1**, where 70K samples are randomly sampled from the complete training dataset for subsequent difficulty-based selection, and **(2) removing Stage#2**, where K-Center sampling is directly applied to the high-quality samples identified in stage#1. Additionally, to validate the necessity of decomposed difficulty calculation based on model perplexity, we investigate **(3) collapsing Stage#2 into Stage#1**, where the model is prompted to verbalize its assessments of the three data difficulties (corresponding prompts are shown in Appendix J.3), bypassing the original difficulty calculation.

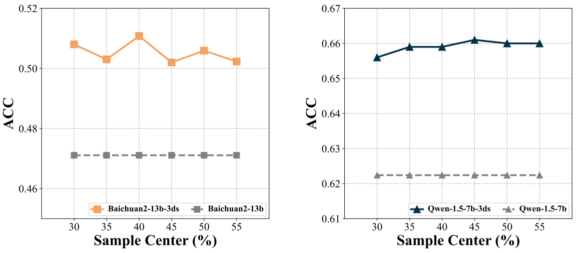
The results show a consistent pattern: **each modification leads to a decrease in performance compared to the original method**, emphasizing the necessity of quality controls and selecting appropriately difficult data. When Stage#2 is collapsed into Stage#1 via difficulty evaluation prompts, performance also degrades. During experiments, we observe that LLMs struggle to provide fine-grained assessments of data difficulty, often generating coarse-grained scores such as 0.5, 0.8, and 1. This lack of granularity makes it challenging to identify nuanced differences in data difficulty and select targeted data with desired moderate difficulties.

While results on multiple-choice benchmarks do not indicate which stage is more important, the analysis performance on CMB-Clin reveals a clearer trend: removing Stage#1 leads to the poorest performance, followed by removing Stage#2 and collapsing Stage#2 into Stage#1. This pattern highlights the crucial role of quality control for the model to provide coherent and high-quality answers. Difficulty-based selection is also essential, as even coarser-grained difficulty measurements by model verbalization yield better results than ignoring difficulty at all. This progressive improvement further reinforces the two-stage design of 3DS.

Ablation on difficulty metrics. We remove each of the three metrics—Instruction Understanding, Response Confidence, and Response Correctness Difficulties and run 3DS without any other modifications. The results in Table 5 demonstrate that, in general, removing any single component results in noticeable performance drops, indicating a decline in certain aspects of the model's medical abilities. These observations validate the necessity of each difficulty metric in identifying beneficial data samples for enhancing LLM's domain abilities. Plus, removing the attention-based importance weighting mechanism also brings performance declines, which validates its effectiveness.

Additional ablation studies on data budgets are introduced in Appendix G.4.

4.4 Impact of Difficulty Thresholds



(a) Baichuan2-13b-chat

(b) Qwen-1.5-7b-chat

Figure 2: Impact of Difficulty Thresholds on Model Performance. This figure illustrates how varying difficulty thresholds affect the accuracy (ACC) of the models Baichuan2-13b-chat and Qwen-1.5-7b-chat across different difficulty sample centers (%).

We conduct sliding-window experiments, varying the selection difficulty ranges ($\sigma \pm 25\%$), to investigate how training data difficulty affects the model's medical domain fine-tuning. As shown in Figure 2, the model's performance improves as difficulties increase, reaching a peak before declining. This pattern further highlights the importance of selecting data that best suits the model's learning capacity. Training on overly simple data limits improvements, while training on excessively difficult data impedes effective learning.

5 Generalization on Law Domain

While our pilot study and main experiments focus on adapting LLMs to the medical domain using Chinese-language data, we note that our 3DS is intrinsically domain-agnostic. To validate its generalization ability, we conduct additional experiments on the law domain using an English-language dataset CaseHOLD (Zheng et al., 2021). Details of experiment setups are introduced in Appendix I. We compare 3DS with random selection and a strong baseline IFD. The results in Table 6 demonstrate that 3DS consistently outperforms baselines in terms of accuracy, achieving an average accuracy of 76.13% with low variance. These results suggest that our model-centric data selection 3DS is adaptable to other specialized domains, supporting its applicability beyond the medical setting.

6 Related Work

Data Selection for LLM Training Data selection for LLM training has been explored by various works.

LLM Turbo	**Baichuan2-13B-Chat**			**Qwen1.5-7B-Instruct**			**Avg.**
Benchmark	CMB-Exam	MMCU-Med	CMB-Clin	CMB-Exam	MMCU-Med	CMB-Clin	
w/o Stage#1	44.64	48.06	16.19	60.37	64.03	15.88	41.53
w/o Stage#2	47.09	50.83	21.83	61.59	65.91	21.55	44.80
Stage#2 Collapsed into Stage#1	47.28	51.01	22.69	60.56	63.99	23.41	44.82
w/o D1	47.35	50.59	23.99	61.47	65.80	24.68	45.65
w/o D2	47.34	47.18	23.54	**62.00**	66.05	23.84	44.99
w/o D3	47.07	50.59	23.08	61.64	65.73	23.83	45.32
w/o Atten	47.10	50.19	29.58	61.79	65.84	27.69	47.03
3DS	**47.37**	**51.08**	**31.50**	61.96	**66.09**	**28.07**	**47.68**

Table 5: Performance comparisons (%) on *CMB-Exam*, *MMCU-Medical* and *CMB-Clin* of ablation studies on stages and difficulty metrics of 3DS. For *CMB-Clin*, the ROUGE score is reported.

Model	**Accuracy**	**Std. Dev.**
Original	57.77	0.25
+Random	73.40	0.80
+IFD	60.30	2.62
+3DS	**76.13**	0.80

Table 6: Accuracy (%) comparison on law domain.

Some works (Das and Khetan, 2023) focus on diversity via statistical clustering or core-set selection, but often overlook data quality, potentially incorporating noisy samples that hinder training. To address quality concerns, some works employ external models like proprietary LLMs (Chen et al., 2023a; Liu et al., 2023; Wettig et al., 2024) or reward models (Du et al., 2023) to evaluate and select high-quality data. However, due to distributional gaps between external evaluators and the model to be trained, data labeled as high-quality may still contain redundant or conflicting information for the target model, limiting its effectiveness. Another line of work leverages internal signals from the target model, such as perplexity (Marion et al., 2023), gradients (Xia et al., 2024), or derived metrics like data learnability (Zhou et al., 2023) and instruction following difficulty (Li et al., 2024b,a). While these signals provide more direct insights into the model's understanding of data, they typically offer only coarse estimates of data difficulty, failing to capture different aspects of data complexity or account for the model's generation behavior. Their selection of the most difficult data also risks overwhelming the model. Though related to active learning (Yoo and Kweon, 2019; Karamcheti et al., 2021; Mindermann et al., 2022) in challenges and insights, the purpose and workflow of LLM data selection are distinct. In this work, we focus exclusively on data selection tailored for training LLMs. We note that existing data selection methods for LLMs mainly focus on pre-training, general instruction-tuning (transforming a base model into a chat model), or task-specific fine-tuning. In contrast, data selection for domain adaptation fine-tuning remains underexplored, where unique challenges lie in selecting data that best elicit the model's diverse domain abilities. To bridge this gap and overcome the limitations of current methods, we introduce a novel model-centric data selection framework and provide fine-grained analysis of data difficulty, enabling better aligned data selection for LLM domain adaptation fine-tuning.

Data Learnability in LLM SFT LLMs encounter significant challenges when learning unfamiliar or complex knowledge during supervised fine-tuning, particularly when the data was not encountered during pretraining, which can impede domain adaptation fine-tuning. Gekhman et al. (2024) found that models acquire new factual knowledge slowly during SFT, especially when the information diverges from their preexisting understanding, leading to a higher risk of hallucinations. Ren et al. (2024) further shows that when the knowledge introduced during Instruction Finetuning significantly differs from what was learned in pre-training, the model struggles to integrate it, causing performance degradation. This highlights the difficulty models face in using pre-training knowledge to understand new concepts. Kang et al. (2024) also emphasizes that unfamiliar examples during fine-tuning increase the likelihood of hallucinations, suggesting that high-difficulty data can destabilize the model and negatively impact its ability to adapt to new domains. Together, these findings underscore the risks associated with fine-tuning on excessively difficult data, which can undermine model performance in domain-specific tasks.

7 Conclusion

In this paper, we highlight the importance of selecting data aligned with the model's distribution for LLM domain adaptation fine-tuning through a pilot study. To this end, we propose a two-stage model-centric data selection framework 3DS. The Stage#1 explicitly aligns data with the LLM's knowledge through prompt-driven selection. The Stage#2 implicitly aligns data via difficulty decomposition. Leveraging Instruction Understanding, Response Confidence, and Response Correctness difficulties calibrated by attention-based importance weighting, 3DS effectively models the LLM's implicit distribution and selects data well-matched to its learning capacity. Extensive experiments on multiple medical and legal tasks show significant performance gains, demonstrating 3DS 's effectiveness and generalization ability. Overall, we offer a path toward more efficient LLM domain adaptation fine-tuning. Future work will explore extending the framework to more domains and refining training strategies based on difficulty metrics for broader applications.

Limitations

Due to time and resource constraints, we have only validated our method in the medical and legal domains. The results show that 3DS is domain-agnostic and adaptable to other fields. However, further experiments may still be needed to fully verify its generalization. 3DS requires the model to rate the entire training set and perform inference on the selected subset. Although in experiments, we utilize VLLM to accelerate the process, it still involves certain computational costs. 3DS performs data selection prior to fine-tuning. Considering that the model's evaluation of data difficulty may evolve during training, future research should explore dynamic selection that adapts to the model's changing state. Additionally, data filtered out is currently discarded. Future work should consider integrating mechanisms such as human-in-the-loop validation or strategies to recover potentially relevant and valuable data from the discarded pool. Finally, considerations for social bias and fairness issues are discussed in Appendix K.

References

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*.

Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. 2022. The effects of data quality on machine learning performance. *arXiv preprint arXiv:2207.14529*.

Minh Duc Bui and Katharina Von Der Wense. 2024. The trade-off between performance, efficiency, and fairness in adapter modules for text classification. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 40–50, Mexico City, Mexico. Association for Computational Linguistics.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Huatuogpt-ii, one-stage training for medical adaption of llms. *Preprint*, arXiv:2311.09774.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023a. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023b. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Devleena Das and Vivek Khetan. 2023. Deft: Data efficient fine-tuning for large language models via unsupervised core-set selection. *arXiv preprint arXiv:2310.16776*.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*.

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? *arXiv preprint arXiv:2405.05904*.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Emily Harris. 2023. Large language models answer medical questions accurately, but can't match clinicians' knowledge. *JAMA*.

Alex Havrilla and Maia Iyer. 2024. Understanding the effect of noise in llm training data with algorithmic chains of thought. *arXiv preprint arXiv:2402.04004*.

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

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. Hykge: A hypothesis knowledge graph enhanced framework for accurate and reliable medical llms responses. *Preprint*, arXiv:2312.15883.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar finetuning examples control how language models hallucinate. *arXiv preprint arXiv:2403.05612*.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher D Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. In *Proceedings of the 59th Annual Meeting of the Association for*

Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7265–7281.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Preprint*, arXiv:2004.04906.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26m, a large-scale chinese medical qa dataset. *Preprint*, arXiv:2305.01526.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024a. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14255–14273, Bangkok, Thailand. Association for Computational Linguistics.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024b. 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)*, pages 7595–7628.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. Flame: Factuality-aware alignment for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. 2022. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pages 15630–15649. PMLR.

Benedict U Nwachukwu, Nathan H Varady, Answorth A Allen, Joshua S Dines, David W Altchek, Riley J Williams III, and Kyle N Kunze. 2024. Currently available large language models do not provide musculoskeletal treatment recommendations that are concordant with evidence-based clinical practice guidelines. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*.

OECD. 2014. *PISA 2012 Results: Creative Problem Solving (Volume V)*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Ankit Pal and Malaikannan Sankarasubbu. 2023. Gemini goes to med school: Exploring the capabilities of multimodal large language models on medical challenge problems and hallucinations.

Dimitrios P Panagoulias, Maria Virvou, and George A Tsihrintzis. 2024. Evaluating llm–generated multimodal diagnosis from medical images and symptom analysis. *arXiv preprint arXiv:2402.01730*.

George Polya and George Pólya. 2014. *How to solve it: A new aspect of mathematical method*, volume 34. Princeton university press.

Michael Preheim, Josef Dorfmeister, and Ethan Snow. 2023. Assessing confidence and certainty of students in an undergraduate linear algebra course. *Journal for STEM Education Research*, 6(1):159–180.

Mengjie Ren, Boxi Cao, Hongyu Lin, Liu Cao, Xianpei Han, Ke Zeng, Guanglu Wan, Xunliang Cai, and Le Sun. 2024. Learning or self-aligning? rethinking instruction fine-tuning. *arXiv preprint arXiv:2402.18243*.

Mohammad-Javad Sanaei, Mehrnaz Sadat Ravari, and Hassan Abolghasemi. 2023. Chatgpt in medicine: Opportunity and challenges. *Iranian Journal of Blood and Cancer*, 15(3):60–67.

Inhwa Song, Sachin R. Pendse, Neha Kumar, and Munmun De Choudhury. 2024. The typing cure: Experiences with large language model chatbots for mental health support. *Preprint*, arXiv:2401.14362.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. *arXiv preprint arXiv:2403.10081*.

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

Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. 2024. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1):43.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. 2023. Gpt-4 and medical image analysis: Strengths, weaknesses and future directions. *Journal of Medical Artificial Intelligence*, 6.

Bin Wang, Chengwei Wei, Zhengyuan Liu, Geyu Lin, and Nancy F Chen. 2024a. Resilience of large language models for noisy instructions. *arXiv preprint arXiv:2404.09754*.

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

Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. 2024b. A survey on data selection for llm instruction tuning. *arXiv preprint arXiv:2402.05123*.

Jun Wang, Changyu Hou, Pengyong Li, Jingjing Gong, Chen Song, Qi Shen, and Guotong Xie. 2023b. Awesome dataset for medical llm: A curated list of popular datasets, models and papers for llms in medical/healthcare. `https://github.com/onejune2018/Awesome-Medical-Healthcare-Dataset-For-LLM`.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. 2023c. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023d. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36:74764–74786.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training lanugage models. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Theresa Isabelle Wilhelm, Jonas Roos, and Robert Kaczmarczyk. 2023. Large language models for therapy recommendations across 3 clinical specialties: comparative study. *Journal of medical Internet research*, 25:e49324.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*.

Zhichao Xu. 2023. Context-aware decoding reduces hallucination in query-focused summarization. *Preprint*, arXiv:2312.14335.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023a. Baichuan 2: Open large-scale language models. *Preprint*, arXiv:2309.10305.

Songhua Yang, Hanjia Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2023b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue.

Donggeun Yoo and In So Kweon. 2019. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102.

Hui Zeng. 2023. Measuring massive multitask chinese understanding. *arXiv preprint arXiv:2304.12986*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, and Benyou Wang. 2023. Huatuogpt, towards taming language model to be a doctor.

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

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

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. SafetyBench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553, Bangkok, Thailand. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International*

Conference on Artificial Intelligence and Law, ICAIL '21, page 159–168, New York, NY, USA. Association for Computing Machinery.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

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.

Haotian Zhou, Tingkai Liu, Qianli Ma, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2023. Lobass: Gauging learnability in supervised fine-tuning data. *arXiv preprint arXiv:2310.13008*.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *Preprint*, arXiv:2101.00774.

He Zhu, Ren Togo, Takahiro Ogawa, and Miki Haseyama. 2023a. A medical domain visual question generation model via large language model.

Wei Zhu, Xiaoling Wang, Huanran Zheng, Mosha Chen, and Buzhou Tang. 2023b. Promptcblue: a chinese prompt tuning benchmark for the medical domain. *arXiv preprint arXiv:2310.14151*.

A Pseudo Codes of 3DS

We provide the pseudo codes of 3DS in Algorithm 1.

Algorithm 1: Model-Centric Data Selection Framework

```
Input: Full dataset $\mathcal{X}$, model $M$, scoring
       threshold $\theta$, difficulty calculation
       functions D1, D2, D3, percentage
       thresholds $p_1, p_2, p_3$, sampling budget $k$
Output: Selected data subset $\mathcal{S}$
Stage#1: Prompt-Driven Data Selection
Initialize $\mathcal{X}_1 \leftarrow \emptyset$
foreach $x \in \mathcal{X}$ do
    Get score $s_x \leftarrow M(\text{prompt}, x)$
    if $s_x \geq \theta$ then
        Add $x$ to $\mathcal{X}_1$
    end
end
Stage#2: Decomposed Difficulty-based Data
  Selection
Initialize $\mathcal{S} \leftarrow \emptyset$
Compute $\mathrm{D1}(x), \mathrm{D2}(x), \mathrm{D3}(x)$ for all $x \in \mathcal{X}_1$
Set $\tau_1, \tau_2, \tau_3$ based on percentiles $p_1, p_2, p_3$ of
  D1, D2, D3
foreach $x \in \mathcal{X}_1$ do
    if $\tau_1^{low} \leq D1(x) \leq \tau_1^{high}$ and
      $\tau_2^{low} \leq D2(x) \leq \tau_2^{high}$ and
      $\tau_3^{low} \leq D3(x) \leq \tau_3^{high}$ then
        Add $x$ to intermediate set $\mathcal{S}_{\text{mid}}$
    end
end
Apply K-Center sampling on $\mathcal{S}_{\text{mid}}$ to select $k$
  diverse data points
Return final selected subset $\mathcal{S}$
```

B Datasheet for Medical Domain Adaptation Fine-Tuning Dataset

Data statistics

The statistics of the training dataset and the test dataset are shown below. The use of the test datasets complies with their respective licenses.

What is the primary purpose of creating this dataset?

This dataset was created to construct a large-scale medical domain instruction-following fine-tuning dataset. The primary purpose is to support the adaptation of large language models (LLMs) to the medical domain by providing diverse and comprehensive training instances. By integrating heterogeneous data sources, including doctor-patient dialogues, medical knowledge bases, and various medical tasks formulated into the instruction-output format, the dataset aims to enhance the ability of LLMs to perform effectively across a wide range of real-world medical scenarios. It is designed to address the unique challenges of the medical domain, such as specialized terminology, complex reasoning,

Dataset	Size (K)
medtalk_singleround	177
medknowledge_KG	796
medknowledge_webqa	360
medtask_promptcblue	82
qa_website	490
Total	**1905**

Table 7: Training Dataset Statistics

Dataset	Type	Size
CMB-Exam	multiple-choice	11200
MMCU-medical	multiple-choice	2819
CMB-Clin	open Q&A	208

Table 8: Test Dataset Statistics

and context-sensitive responses, thereby enabling LLMs to better meet the demands of healthcare applications.

What are the specific components of the dataset, and how were they constructed or sourced?

Our dataset integrates multiple open-sourced medical instruction fine-tuning datasets from diverse sources, along with doctor-patient dialogue data extracted from medical consultation websites and a variety of medical tasks reformulated into the instruction-output format, as detailed in Table 7. **Medtalk_singleround** originates from open-sourced doctor-patient question-and-answer datasets, including CMedQA2 (Zhang et al., 2018) and Health-Care-Magic[1]. **Medknowledge_KG** is built from the Online Medical Knowledge-Based Data in Huatuo26M (Li et al., 2023), which is derived from the extensive medical literature data provided by the Chinese Medical Association. **Medknowledge_webqa** includes knowledge-driven, open-ended question-and-answer pairs in the medical domain, sourced from (Wang et al., 2023b). **Medtask_promptcblue** combines the promptCBLUE dataset (Zhu et al., 2023b) with additional data converted into the instruction-output format from the CBLUE benchmark (Zhang et al., 2022). **QA_website** contains authentic doctor-patient dialogue data collected from the online platform of a collaborating hospital. Examples from these datasets are shown in Table 9.

Are the data sources legal? How are privacy and ethical considerations addressed?

The dataset is derived from carefully selected sources, including publicly available datasets and data crawled from the website of a collaborating hospital. Explicit permission was obtained from the collaborating hospital for the use of the crawled data, and all data have been anonymized to ensure that no personal information is exposed. Additionally, the hospital's website provides open-access data, complying with relevant legal and ethical standards. This ensures the legality and security of the data while addressing privacy and ethical concerns.

What are the potential risks and limitations of this dataset?

The dataset has certain inherent risks and limitations that should be acknowledged. First, as the data is collected from diverse sources, it may contain noise or inconsistencies, which could affect the quality and reliability of downstream applications. Additionally, since the dataset is derived from Chinese text corpora, including medical advice and Q&A exchanges, its content may be culturally and regionally specific, making it more suitable for East Asian populations. As a result, the medical recommendations and insights in the dataset may not generalize well to other demographic or cultural contexts.

To address these issues, users should carefully evaluate the dataset's suitability for their intended applications and, if necessary, consider adapting the data to align with broader use cases. Moreover, noise reduction and validation techniques can be employed to improve data quality and reliability in specific tasks.

What is the usage case for this dataset?

This dataset is primarily intended for instruction fine-tuning of large language models (LLMs), as already utilized in this study. Practitioners can use it to fine-tune LLMs to adapt to the medical domain, as well as to enhance its medical abilities in general fine-tuning. Additionally, the dataset may be useful for more specific tasks, such as fine-tuning for sub-tasks in the dataset.

What is the distribution method and maintenance plan for this dataset?

The dataset is distributed as an open-source resource at https://drive.google.com/drive/folders/1SfrwQkDrQJ8i_EIqfc2Di0Xa5Y5pzY9H, allowing researchers and developers to access and utilize it freely under the specified license. We are committed to the ongoing maintenance of the dataset. If any errors or inaccuracies are identified, particularly those related to medical knowledge, we will promptly update the dataset to correct such issues, removing erroneous data as necessary. Additionally, we will continue to provide updated documentation to ensure the dataset's effective use. While the dataset is stable at present, users are encouraged to provide feedback or suggest improvements, and we will consider updates based on user input or evolving needs in the field. This ensures that the dataset remains reliable and beneficial for the community.

C K-Center Sampling Algorithm

In our data selection framework, K-Center sampling is employed to ensure diversity within the selected instruction fine-tuning data. After filtering based on difficulty levels, we obtain an intermediate set $\mathcal{S}_{\text{mid}}$, composed of data points within a moderate difficulty range. The K-Center sampling is then applied on $\mathcal{S}_{\text{mid}}$. Specifically, the process works as follows:

1. Embedding Generation: For each data sample,

[1] https://www.kaggle.com/datasets/gunman02/health-care-magic

the instruction part is encoded into an embedding using the LLM. We extract the last hidden states of the LLM and compute the average across all tokens in the sequence to form a fixed-size embedding vector. These embeddings represent the semantic content of the instruction.

2. K-Center Sampling: Using these embeddings, the K-Center sampling algorithm selects k data points in a greedy manner. The goal is to maximize the minimum distance between any pair of selected data points, ensuring that the sampled data points are as distinct as possible. This promotes diversity in the selected dataset and minimizes the risk of overfitting to similar data points.

The pseudo codes of this greedy K-Center sampling process are shown in Algorithm 2:

Algorithm 2: Greedy K-Center Sampling

```
Input: Intermediate set
    $S_{mid} = \{s_1, s_2, \ldots, s_n\}$, model $M$, data
    budget $k$
Output: Final selected set $\mathcal{S}$
Step 1: Encode data in $S_{mid}$ using model $M$;
foreach $s_i \in S_{mid}$ do
    Encode $s$ using $M$ to obtain the embedding
      $e_s$ ;
end
Step 2: Run K-Center greedy algorithm;
Initialize $\mathcal{S} \leftarrow \emptyset$ ;
Initialize min_distances $\leftarrow \infty$ ;
for $i = 1$ to $k$ do
    if $\mathcal{S} = \emptyset$ then
        Select $s_j \in S_{mid}$ randomly and add it to
          $\mathcal{S}$ ;
    else
        $min\_distances_j = \min_{s_i \in \mathcal{S}} \|e_{s_j} - e_{s_i}\|_2, \quad \forall s_j \in S_{mid} \setminus \mathcal{S}$;
        Select $s^* = \arg\max_{s_j \in S_{mid} \setminus \mathcal{S}} min\_distances_j$;
        Add $s^*$ to $\mathcal{S}$;
    end
end
return $\mathcal{S}$
```

D Baseline Implementations

Due to differences in task settings and datasets, we re-implement the baselines using their publicly available codes. We adapt their data selection strategies to our domain adaptation fine-tuning task on the medical instruction fine-tuning dataset and models. The re-implementation details are as follows and our use of the code repositories complies with their respective licenses:

(1) Alpagasus: (Chen et al., 2023a) We adopt the open-sourced implementation[2], officially verified by the original authors. Given the scale of our full training dataset, applying GPT-4 annotation to the entire set would incur substantial financial cost due to API usage. Constrained by our budget, we randomly sample 70K training samples and assess their quality using the provided prompt with GPT-4o. From data scoring above the default threshold of 4.5, we randomly select 5K samples.

(2) DEITA: (Liu et al., 2023) We utilize the official implementation from the public GitHub repository[3] and directly download their trained data quality and complexity scorer models from HuggingFace[45] without modification. The scorers are applied to randomly sampled 70K training data. We then select the top 5K samples with the highest scores in both quality and complexity.

(3) IFD: (Li et al., 2024a,b) The Instruction Following Difficulty (IFD) method begins by calculating the instruction-following difficulty scores for each data point through model forward propagation. Given that our full domain dataset consists of over 1.9 million samples, performing this step on the entire dataset would be computationally prohibitive. Therefore, we randomly sample 60K samples from the training set, an amount comparable to the dataset size used in our 3DS after Stage#1. We compute IFD scores for this subset, and, following the recommendations in the original paper, select the samples with highest scores. The data budget is constrained to 5k samples, ensuring consistent with our main experimental setup.

(4) MoDS: (Du et al., 2023) For the MoDS baseline, We follow the original paper's implementations, using the reward model `reward-model-deberta-v3-large-v2`[6] to score the full dataset. We then obtain samples with scores above 0.5, yielding a subset of 120k high-quality data samples. From this subset, we apply K-Center sampling to select 2k seed samples for model warm-up training. Subsequently, the trained model perform inference on the 120k high-quality subset, and these predictions are rescored using the same reward model. Data samples where model's generated answers score below 0 are deemed necessary and are combined with the seed samples. From this merged set, we randomly select 5k samples as the final training data, and train models from scratch on this final data.

(5) LESS: (Xia et al., 2024) The LESS method involves constructing a gradient library based on the original data, which incurs significant computational costs, particularly for the large dataset like ours. Similarly,

[2] https://github.com/gpt4life/alpagasus

[3] https://github.com/hkust-nlp/deita

[4] https://huggingface.co/hkust-nlp/deita-quality-scorer

[5] https://huggingface.co/hkust-nlp/deita-complexity-scorer

[6] https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2

we sample 60k data points to compute the gradients. Unlike the original LESS method that targets specific downstream tasks and uses samples from the targeting dataset to construct a validation set, our domain adaptation fine-tuning scenario does not involve fixed downstream tasks. Therefore, we randomly selected an additional 100 samples from the training set as the validation set. Then we run the provided codes and select 5k training samples.

E Implementation Details

The difficulty thresholds in our experiments are determined based on model performance on a hold-out CMB-validation set composed of 280 samples provided in the CMB benchmark (Wang et al., 2023c). All experiments are conducted using the PyTorch 2.4.0 in Python 3.9, on 8 NVIDIA H100 GPUs and an Intel(R) Xeon(R) CPU, with both training and inference performed using half-precision FP16 for efficiency. We employ the LoRA fine-tuning method, targeting all linear modules within the model, with a learning rate of 5×10^{-5}, a batch size of 64, and a single epoch of training. The learning rate is scheduled using a cosine decay scheduler with a warmup ratio of 0.1. The LoRA rank is set to 8, and the input sequence length is cut off at 1024 tokens. DeepSpeed Zero-3 is used to optimize distributed training. For instruction scoring, response generation, and training, we use templates corresponding to each model, implemented through the llamafactory project (Zheng et al., 2024).

Due to the high computational cost of training and testing LLMs, most existing instruction data selection studies conduct experiments with a single run for efficiency (Li et al., 2024b; Du et al., 2023). We adopt this approach as well. However, to assess the reliability of our results, we perform the random selection experiment three times. The results show consistent performance with low variance (MMCU: 0.07; CMB 0.01 for `Qwen1.5-7B-Chat`) and narrow error bars (±0.26 and ±0.08 for `Qwen1.5-7B-Chat`), demonstrating that our findings are statistically stable and reliable.

F Evaluation Metrics

To evaluate the performance of LLMs on multi-task medical choice questions, we instruct the models to provide only the correct answer and adopt the widely-used metric, **Exact Match (EM)**, as recommended by prior work (Zhu et al., 2021; Karpukhin et al., 2020). An answer is deemed correct under the EM metric if its form exactly matches all the correct answers listed in the ground truth. The EM score is computed as follows:

$$EM = \frac{\text{Number of Correctly Matched Answers}}{\text{Total Number of Answers}} \times 100\%.$$

For open-domain medical Q&A tasks, we employ **ROUGE-R** (Xu, 2023; Jiang et al., 2024) and **Bilingual Evaluation Understudy (BLEU)** to assess the quality of the LLMs' responses.

BLEU-N Specifically, **BLEU-1** is used to measure answer precision, and **BLEU-4** evaluates answer fluency by considering higher-order n-gram consistency. **BLEU** evaluates the similarity of generated responses to the ground truth using the following formula:

$$\text{BLEU-N} = BP \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log p_n\right),$$

where p_n is the precision of n-grams, BP is the Brevity Penalty, calculated as:

$$BP = \begin{cases} 1, & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \leq r \end{cases}.$$

Here c is the length of the generated response, and r is the length of the reference response.

ROUGE-R quantifies the recall of retrieved knowledge in the LLMs' responses, emphasizing their ability to comprehensively cover the information relevant to the query. For a generated response R and a reference G, ROUGE-R is computed as:

$$\text{ROUGE-R} = \frac{|R \cap G|}{|G|},$$

where $|R \cap G|$ denotes the number of overlapping n-grams between the generated response and the reference, and $|G|$ is the total number of n-grams in the reference.

During implementation, We use the 'rouge' package to calculate ROUGE scores and the 'nltk' module to compute BLEU scores (from BLEU-1 to BLEU-4), utilizing the smoothing function for BLEU and the default settings for ROUGE.

G Supplementary Experiments

G.1 Win Rates Evaluation

When evaluating model performance on the open Q&A dataset CMB-Clin, in addition to traditional metrics such as BLEU1, BLEU4 and Rouge scores, we conduct a pair-wise comparison to more thoroughly compare the fine-tuned models' medical analysis ability. In this experiment, we randomly sample 100 answers from each model and employ GPT-4o, a highly capable LLM, as the judge to determine which model generates a better answer. Below, we present the prompt used to instruct GPT-4o to compare answers from two models in this qualitative evaluation. To ensure a fair comparison and eliminate any possible positional bias in GPT-o4, we randomly assign the answers from each model as "Student 1" or "Student 2" throughout the experiment.

Results shown in Figure 3 demonstrate that 3DS exhibits substantially higher win rates compared to all other baselines. Notably, the larger and stronger models `Baichuan2-13B-Chat`, `Qwen1.5-7B-Instruct` and `Qwen2.5-7B-Instruct` generally show higher win rates compared to `LLaMA3-8B-Instruct`, which indicates that 3DS also exhibits scalability. This evaluation provides qualitative evidence that 3DS effectively enhances the model to deliver more clinically accurate outputs.

CMB-Clin Evaluation Prompt

You are now a medical expert guiding students in analyzing medical cases. You have two students, Student 1 and Student 2. You assess them through real medical case questions and choose the one with the best answer to become your assistant.

[High-Quality Answer Criteria]
1. The answer should address the question directly and solve the problem posed.

2. The description of symptoms should be comprehensive and accurate, and the diagnosis should be the most reasonable inference based on all relevant factors and possibilities.

3. The treatment recommendation should be effective and reliable, considering the severity or stage of the condition.

4. The prescription should consider indications, contraindications, and dosages, being both effective and reliable.

[Judgment Instructions]
Please compare the answers of Student 1 and Student 2. You need to tell me whether Student 1 is [better], [worse], or [equal] to Student 2. Compare their answers, refer to the question and the correct answer, and determine which one meets the given requirements more closely. Please only output one of the following: [Student 1 is better than Student 2], [Student 1 is worse than Student 2], or [Student 1 and Student 2 are equal]. Do not output any other words.

[Case Example]
Here is the [Question]:
`<Insert medical question here>`

Here is the [Standard Answer]:
`<Insert standard answer here>`

Here is [Student 1]'s answer:
`<Insert Student 1's answer here>`

Here is [Student 2]'s answer:
`<Insert Student 2's answer here>`
Please compare the two answers and give your judgment.

G.2 Comparison with Existing Medical LLMs

Model	CMB-Exam	MMCU-Med
Baichuan2-13B-3DS	47.37	51.08
Qwen1.5-7B-3DS	61.96	66.09
Qwen2.5-7B-3DS	**79.06**	**85.70**
Meditron-7B	11.20	12.16
Huatuo-II-7B	27.69	47.18
Huatuo-II-34B	59.54	66.10

Table 10: Performance comparisons with existing medical LLMs.

To further validate the practical utility of 3DS, we conduct comparisons with existing medical LLMs. We compare 3DS fine-tuned models to established medical LLMs, including open-source models MediTron (Chen et al., 2023b) (7B version due to its similar size to Qwen models), and state-of-the-art Chinese medical LLMs HuatuoGPT-II-7B, and HuatuoGPT-II-34B (Chen et al., 2024), to see whether our framework can benefit the construction of medical LLMs. The results presented in Table 10 show that, MediTron-7B, as an English-based LLM, demonstrates limited performance on Chinese medical benchmarks. Huatuo-II-7B also falls short on to similar-sized Qwen models. Huatuo-II-34B, with nearly five times the size of Qwen1.5-7B and Qwen2.5-7B, achieves only comparable performance.

It is worth noting that the performance of fine-tuned models is closely tied to the capability of the base model, so relative improvements achieved through domain-specific fine-tuning are more important than absolute performance. Still, the strong performance of models fine-tuned with 3DS validates its practical utility and efficiency for developing medical domain LLMs, paving ways for more building more powerful and advanced models in the future.

G.3 More Results for Ablation on 3DS

In the ablation studies in 4.3, for CMB-Clin benchmark, we only report the ROUGE score. We provide BLEU-1, BLEU-4 scores and win-rates of the experiments in Table 11 and Table 12. Results are consistent with previous observations that the original 3DS significantly outperforms ablation variants, supporting the validity of our designed two-stage framework and three data difficulty metrics.

LLM Turbo	Baichuan2-13B-Chat		Qwen-1.5-7B-Instruct	
Metric	BLEU-1	BLEU-4	BLEU-1	BLEU-4
w/o Stage#1	14.13	29.60	15.50	31.94
w/o Stage#2	20.56	46.86	21.55	47.39
Stage#2 into Stage#1	21.48	50.16	21.73	52.27
w/o D1	22.55	51.75	24.14	55.12
w/o D2	22.22	52.06	20.48	49.59
w/o D3	20.86	49.40	22.27	50.18
3DS	**24.15**	**63.51**	**24.40**	**60.32**

Table 11: Performance (BLEU-1, BLEU-4) on *CMB-Clin* for ablation experiments. The best performance is highlighted in **bold**.

LLM Turbo	Baichuan2-13B-Chat			Qwen-1.5-7B-Instruct		
Metric	Win	Tie	Lose	Win	Tie	Lose
vs w/o Stage#1	66.5	9.0	24.5	70.5	3.0	26.5
vs w/o Stage#2	66.0	15.5	28.5	66.0	5.5	28.5
vs Stage#2 into Stage#1	63.5	18.0	18.5	54.5	2.5	43.0

Table 12: Win-rates (%) of GPT-4o judgment on *CMB-Clin*, comparing 3DS with stage ablation variants.

G.4 Ablation on Data Budgets

We conduct ablation experiments varying the selection data budgets. Results in Table 13 show that increasing the training data size initially boosts performance as the model learns to align with domain-specific knowledge. However, beyond a certain point (5K), performance degradation arise due to potential data redundancy and reduced diversity.

H Parameter Selection Guidelines

In 3DS's Stage#2 Decomposed Difficulty-based Data Selection, data within a moderate difficulty range are selected. How to determine the optimal difficulty range is thus essential. We provide selection guidelines based on our experiments. We identify that the 25%-75% difficulty range is a robust choice. For model-specific optimization, we recommend this implementation procedure:

- Model Capability Profiling: Conduct pre-fine-tuning validation to benchmark the model's baseline performance. Strong domain task performance suggests higher difficulty thresholds, while weaker models benefit from more conservative ranges.

- Hyperparameter Search: Implement search over potential ranges and select the values that yield the best performance on the validation set. This allows for adapting the difficulty range to the model's specific strengths and weaknesses.

I Law Domain Experiment Details

To assess the generalization ability of our model-centric data selection framework beyond the medical domain, we conduct experiments on the law domain, utilizing **CaseHOLD** dataset (Zheng et al., 2021). This dataset consists of over 53,000 multiple-choice questions derived from U.S. court decisions. Each instance presents a case citation context along with five candidate legal holdings, of which only one is correct. The task simulates legal reasoning by requiring models to identify the option that best matches the cited precedent.

We follow a standard instruction-tuning setup by converting CaseHOLD into an Alpaca-style format. The `instruction` is fixed to a law domain-specific prompt:

CaseHOLD Instruction

As a law expert, please select the option that best matches the legal holding cited in the case. Answer with the option letter only (A/B/C/D/E).

The `input` contains the case citation context and five formatted candidate holdings:

CaseHOLD Input

Case Citation Context: [citing_context]
Options: A. [holding_0] B. [holding_1] . . . E. [holding_4]

We fine-tune `LLaMA3-8B-Instruct` on 5K training samples selected from the CaseHOLD training set using three different strategies: **(1)** Random Selection, **(2)** IFD (Li et al., 2024b), a strong instruction filtering baseline, and **(3)** our proposed model-centric selection framework 3**DS**. All models are trained under the same hyperparameters, and each experiment is repeated three times with different random seeds. We report the mean accuracy and standard deviation on a selected 1K samples from the CaseHOLD test set.

J Data Evaluation Prompts

J.1 Data Quality Evaluation Prompt

In the pilot study and the first stage of 3DS, we utilize a prompt to instruct models to evaluate data quality on its internal knowledge. Inspired by existing works (Chen et al., 2024; Wang et al., 2023c; Liu et al., 2023), the model is asked to assess data quality across five dimensions: Instruction Complexity, Response Relevance, Response Thoroughness, Response Logic and Knowledge Richness. We provide the model with detailed scoring guidelines. The specific prompt used in this process is shown below.

Quality Evaluation Prompt

You are an AI assistant with medical expertise. Your task is to objectively assess the quality of the medical dialogue between the user and assistant based on your knowledge, and provide a score. The data may consist of single or multi-turn dialogues. You should evaluate based on the complexity of the question, relevance of the response, thoroughness, logical coherence, and knowledge richness, and provide an overall score. Focus on medical-specific characteristics to ensure accuracy.

[Evaluation Criteria]

1. *Question Complexity*: Evaluate the complexity of the user's question. If the question requires deep understanding, reasoning, or medical knowledge, score above 80.

2. *Response Relevance*: Assess if the assistant's response is directly aligned with the question. Score above 80 for responses tightly related to the question.

3. *Response Thoroughness*: Check if the response thoroughly addresses the question with sufficient detail. A score above 80 reflects comprehensive answers.

4. *Response Logic*: Ensure the response follows clear reasoning and logic. A score above 80 reflects well-structured reasoning.

5. *Knowledge Richness*: Determine whether the response demonstrates rich, specialized medical knowledge. A score above 80 indicates depth and accuracy.

[Scoring Guidelines]

[80-100]: Excellent. High complexity, thoroughness, relevance, logic, and knowledge richness, meeting medical standards.

[60-79]: Good. Strong performance but with minor deficiencies in logic or knowledge.

[40-59]: Fair. Noticeable issues such as unclear logic or insufficient depth.

[20-39]: Poor. Fails to properly address the medical issue or lacks substance.

[0-19]: Very Poor. Lacks relevance, logic, or medical knowledge.

[Start Conversation]
Refer to the guidelines and score the following dialogue data based on the criteria. Follow the output format strictly:
{score:}

Dialogue:
`<qa_pairs>`
Output:

J.2 Data Difficulty Evaluation Prompt

In the second empirical study, we prompt models to rate overall data difficulty according to its knowledge. The specific prompt used in this process is shown below.

Overall Difficulty Evaluation Prompt

You are a medical expert. I will provide you with an instruction related to the medical field. Based on your knowledge, please evaluate the difficulty of this instruction.

1. *Medical Knowledge Complexity*: Does this instruction involve complex medical knowledge?

2. *Reasoning Complexity*: Does answering this instruction require multi-step reasoning, integration of multiple sources of information, or handling clinical uncertainty?

3. *Overall Challenge*: Considering the above factors, what is the overall difficulty of this instruction?

Based on these considerations, please provide a **comprehensive difficulty rating** from 1 (very easy) to 5 (very difficult). Only output the score; do not provide any explanation.
Instruction to evaluate:
`{instruction}`

Please return an integer between 1 and 5, representing the overall difficulty of the instruction for you. Only output the score and nothing else.

J.3 Decomposed Difficulty Prompts

In the ablation study where we collapse Stage#2 in 3DS into Stage#1, using prompts to instruct model to score the three decomposed data difficulties. The prompts utilized are listed below.

Instruction Following Difficulty Prompt

Based on your existing knowledge, evaluate the difficulty of understanding the following instruction. The higher the complexity and ambiguity of the instruction, the more difficult it is for the model to understand. Please provide a score between 0 and 1, where a higher score indicates that the instruction is more difficult for you to understand.

Instruction to be evaluated: `{instruction}`

Please return a real number between 0 and 1, representing the difficulty of understanding the instruction. Only output the score, and do not output anything else.

Response Confidence Difficulty Prompt

Based on your existing knowledge, evaluate the difficulty of confidently and definitively providing the following evaluated response to the instruction. The more difficult it is to confidently provide this response, the higher the difficulty. Please provide a score between 0 and 1, where a higher score indicates greater difficulty in answering confidently.

Instruction: `{instruction}`
Response to be evaluated: `{generated output}`

Please return a real number between 0 and 1, representing the difficulty of confidently providing the response to the instruction. Only output the score, and do not output anything else.

Response Correctness Difficulty Prompt

Based on the following instruction and the standard answer, evaluate the difficulty of providing the correct standard answer. If the instruction is complex or the answer requires high expertise, making it difficult to provide the correct answer, the difficulty will be higher. Please provide a score between 0 and 1, where a higher score indicates greater difficulty in providing the correct answer.

Instruction: `{instruction}`
Standard Answer: `{output}`

Please return a real number between 0 and 1, representing the difficulty of providing the correct answer. Only output the score, and do not output anything else.

K Bias and Fairness Considerations

Fairness and bias are critical considerations, particularly in sensitive domains like healthcare. While our approach demonstrates promising results in fine-tuning LLMs for medical tasks, it is essential to acknowledge its limitations and potential implications concerning fairness and bias. Our method employs the LLM to evaluate data quality and calculate data difficulty. Although the evaluation prompts and difficulty calculation metrics are designed to be neutral, the inherent biases in the base model may still influence the selection results. And the LoRA fine-tuning's impact on LLM fairness also needs further investigations (Bui and Von Der Wense, 2024). Another source of potential bias arises from the composition of our training data, which predominantly consists of Chinese medical texts. While this dataset effectively reflects the health conditions and medical practices of East Asian populations, it may limit the generalizability to other regions or demographics. Current LLM data selection methods generally prioritize factors such as difficulty, quality, or diversity, without addressing fairness or examine what data is included or excluded. They focus on improving model performance on standard benchmarks, while the impact of these methods on fairness, safety, and truthfulness benchmarks, such as SafetyBench (Zhang et al., 2024) and TruthfulQA (Lin et al., 2022), remains underexplored. Therefore, we recognize that these issues are valuable directions for future research. Investigating how data selection and fine-tuning methods impact LLM fairness and safety will be essential for developing more equitable and reliable LLMs.

	Medtalk_singleround	**English translation**
Question	医生请问怀孕时母亲得了甲亢会遗传个孩子么？之前得过甲亢怀孕时又复发了但是没有吃药想知道宝宝会不会被遗传？	Q: Doctor, can hyperthyroidism during pregnancy be inherited by the baby? Mine recurred during pregnancy, but I didn't take medication. Will my baby be affected?
Answer	甲亢与遗传有一定的关系，但不是都会遗传的，不要担心。建议平时合理安排饮食，需要高热量高蛋白质高维生素和低碘的饮食。精神要放松。适当休息，避免重体力活。	A: Hyperthyroidism has some genetic links but is not always inherited. Don't worry. Maintain a balanced diet rich in calories, protein, and vitamins, but low in iodine. Stay relaxed, rest adequately, and avoid strenuous activities.
	Medknowledge_KG	**English translation**
Question	颜面部凹陷的手术治疗有些什么？	Q: What surgical treatments are available for facial depression?
Answer	自体颗粒脂肪移植；自体脂肪移植；自体脂肪干细胞移植；自体脂肪颗粒移植。	A: Autologous micro-fat grafting, autologous fat transplantation, autologous stem cell-enriched fat grafting, and autologous fat particle transplantation.
	Medknowledge_webqa	**English translation**
Question	请描述寄生虫病的治疗方法。	Q: Please describe the treatment methods for parasitic diseases.
Answer	以消灭寄生虫为主，根据虫种采用最有效的驱虫药物。在感染较重而寄主较衰弱时，可给予支持疗法，有外科并发症时应及时进行外科处理。	A: The primary goal is to eliminate the parasites by using the most effective antiparasitic drugs based on the specific type of parasite. In cases of severe infection where the host is weakened, supportive therapy may be provided. Surgical intervention should be performed promptly if complications arise.
	Medtask_promptcblue	**English translation**
Question	问诊对话的实体抽取：两个礼拜片感冒了，现在还有点咳嗽，会不会和这个有关？选项：具体的药物名称，医疗操作，医学检查检验，症状。	Q: Entity extraction in diagnostic dialogues: "Caught a cold two weeks ago, still have a bit of a cough now—could it be related?" Options: specific medication names, medical procedures, medical tests and examinations, symptoms.
Answer	上述句子中的实体包含：症状实体：感冒，咳嗽。	A: The entities in the above sentence include: Symptom entities: cold, cough.
	QA_website	**English translation**
Question	每天下午低烧三十六七°C，有时胸闷，没有咳嗦，盗汗，乏力的，有没有得肺结核的可能?	Q: Low-grade fever of 36-37°C every afternoon, occasional chest tightness, no cough, night sweats, or fatigue—could this indicate a possibility of tuberculosis?
Answer	你这个体温其实从临床上来讲，不算是低烧，一般来讲，37度二以上才算是低热，所以说你这个跟集合的关系不是特别大的，你倒是可以看一下有没有病毒感染的可能，再一个，有没有新冠的问题?	A: From a clinical perspective, this temperature doesn't qualify as a low-grade fever—typically, temperatures above 37.2°C are considered low-grade. Therefore, its connection to tuberculosis is unlikely. However, you might want to check for the possibility of a viral infection or consider whether it could be related to COVID-19.

Table 9: Examples For various type dataset

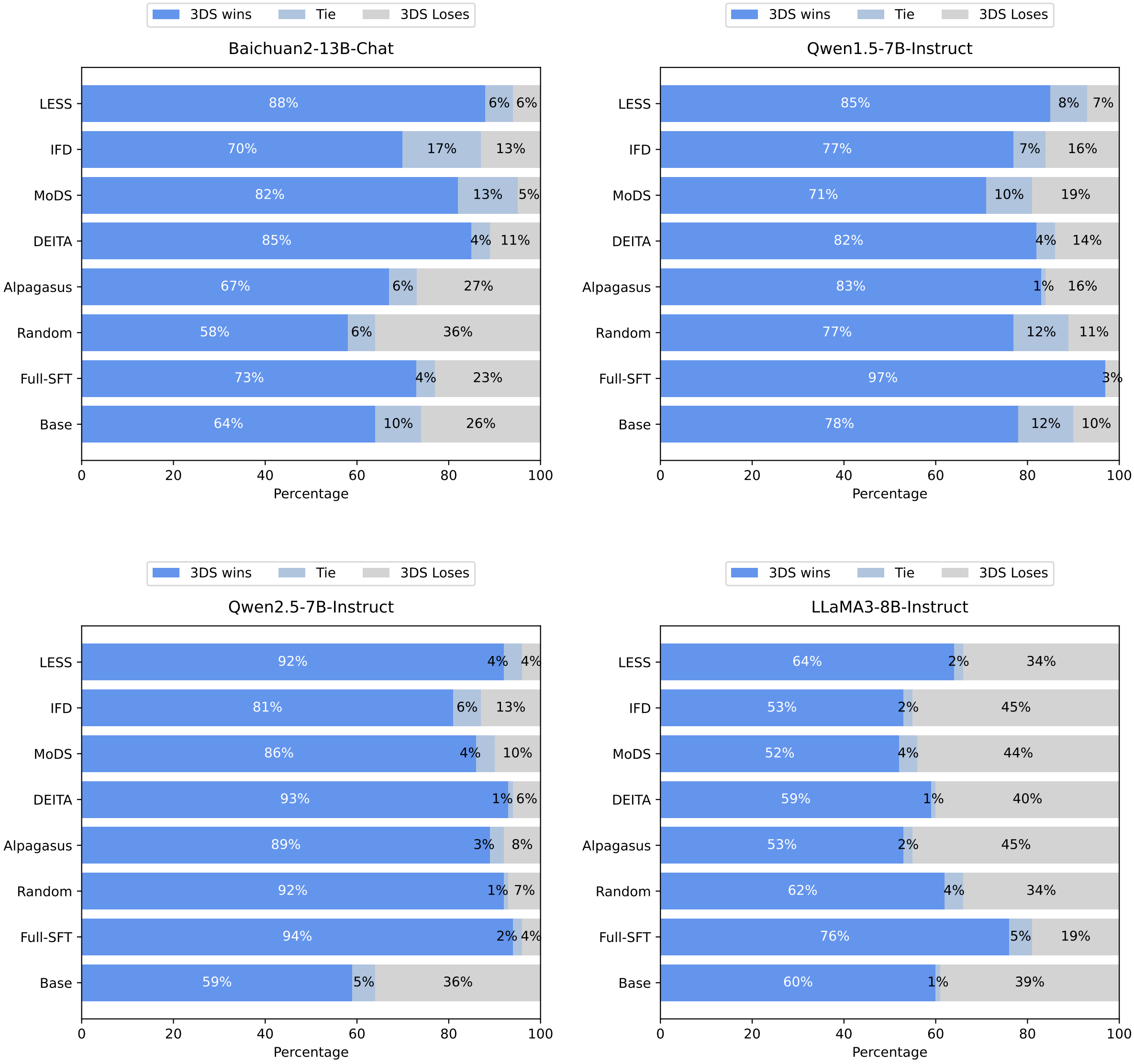


Figure 3: GPT-4o judgement of CMB-Clin.

Model	Dataset	3k	4k	5k	6k	7k
Baichuan2-13B-Chat	CMB-Exam	46.87	47.30	**47.37**	46.95	46.98
	MMCU-Medical	48.67	49.91	**51.08**	50.16	50.27
Qwen1.5-7B-Instruct	CMB-Exam	60.47	60.45	**61.96**	60.78	60.53
	MMCU-Medical	63.64	63.92	**66.09**	64.49	64.10

Table 13: Performance comparison of models trained on different data budgets.