Fine-tuning Large Language Models with Human-inspired Learning Strategies in Medical Question Answering

Anonymous ACL submission

Abstract

Training Large Language Models (LLMs) incurs substantial data-related costs, motivating the development of data-efficient training methods through optimised data ordering and selection. Human-inspired learning strategies, such as curriculum learning, offer possibilities for efficient training by ordering data according to common human learning practices. Despite evidence that curriculum learning improves performance of natural language understanding tasks in fine-tuning LLMs, its application to domain-specific question-answering remains underexplored. In this work, we comprehensively examine the effectiveness of human-inspired learning strategies for fine-tuning LLMs in medical question answering. Our work complements previous studies by extending the evaluation to non-curriculum-based learning across multiple language models, using both human-defined and automated data labels. Our results show moderate impact in using human-inspired learning strategies for fine-tuning LLMs, with maximum accuracy gains of 1.77% per model and 1.81% per dataset. However, the effectiveness of these learning strategies varies significantly across different model-dataset combinations, suggesting caution in generalising human-inspired strategies for fine-tuning language models. We also find that curriculum learning using LLM-defined question difficulty outperformed human-defined difficulty, highlighting the potential of using model-generated metrics in optimal curriculum design.

1 Introduction

Training Large Language Models (LLMs) incurs substantial data-related costs both in compute (Hoffmann et al., 2022; Jeon and Roy, 2022) and data-collection (Muennighoff et al., 2023; Xue et al., 2023). Recent efforts have been made to improve model performance through more efficient use of the same training data (Sachdeva et al., 2024; Hase et al., 2024). Building on the historic success of human-inspired machine learning methods (Sayal et al., 2023), human-inspired learning strategies also offer possibilities for organising data ordering according to human learning practices to achieve efficient training.

The most established technique for data ordering is curriculum learning, in which training samples are ordered from easiest to hardest (Hase et al., 2024; Xu et al., 2020). This method has led to some improvements in general knowledge acquisition (Lee et al., 2024), natural language reasoning (Maharana and Bansal, 2022) and information retrieval (Penha and Hauff, 2019) benchmarks. In addition, variations on curriculum learning, such as interleaving different subject areas have also been effective for increasing world-knowledge and commonsense reasoning (Lee et al., 2024).

Despite evidence that curriculum learning improves foundational natural language processing capabilities in fine-tuning LLMs, its application to domain-specific question-answering remains underexplored. Medical question-answering, in particular, is a high-stakes domain requiring accurate information retrieval, and several models fine-tuned on medical data have been recently released to address this need (Saab et al., 2024; Chen et al., 2023b; Singhal et al., 2023). Previous studies on curriculum learning have also considered only a single model and curriculum strategy at a time, which limits the generalisability of the results (Lee et al., 2024; Maharana and Bansal, 2022; Xu et al., 2020). Our study extends previous research by evaluating a range of human-inspired learning strategies, including non-curriculum-based ones, across multiple models and various data labelling scenarios for fine-tuning LLMs. Through this comprehensive evaluation, we aim to provide insights into the usefulness of human-inspired learning strategies for optimising the fine-tuning process of LLMs.

Specifically, our contributions are:

- **Broad-based evaluation of human-inspired learning strategies**: Unlike previous work that focused on individual language models and curriculum learning, we compared four LLMs of different sizes and architectures, and extended the analysis to non-curriculum-based learning strategies. Our findings indicate that effectiveness of human-inspired learning strategies varies significantly across different model-dataset combinations.

- **Compare machine-generated and human-generated data labels**: We introduced a novel automated method for annotating question difficulty and category, using ensemble LLM responses and text clustering to define the learning strategies. This approach leverages pre-trained LLMs to label data, offering a cost-effective alternative to human annotations. Our findings showed that using LLM-defined question difficulty yielded improved performance in curriculum learning compared to human-defined difficulty.

2 Related Work

Data-efficient fine-tuning on LLMs Data selection and ordering methods are essential for data-efficient fine-tuning of LLMs. Sachdeva et al. (2024) explored data-efficient fine-tuning by assessing training example quality using zero-shot reasoning and selecting diverse samples to represent the data distribution. Das and Khetan (2023) used unsupervised core-set selection to minimise data requirements while maintaining accuracy. Chen et al. (2023a) proposed a learning framework that uses an ordered data sampling algorithm to enable efficient learning of advanced language processing skills. In contrast to these approaches that select high-quality subsets of data, our research focuses on adopting human-inspired learning strategies for data ordering to enhance the efficiency of fine-tuning.

Human-inspired learning for fine-tuning Curriculum learning has been widely explored to fine-tuning language models for general-purpose natural language tasks. For example, Xu et al. (2020) demonstrated that defining question difficulty by cross-reviewing the training set with multiple teacher models and using that curriculum to fine-tune the BERT large model led to consistent performance improvements across various natural language understanding tasks by up to 1.3%. Similarly, Maharana and Bansal (2022) found that fine-tuning RoBERTa with fixed and adaptive curricula defined by a teacher model improved performance on five commonsense reasoning tasks by up to 2%. In addition, Lee et al. (2024) demonstrated that interleaving the curriculum by subjects outperformed other curriculum arrangements using Llama 2-13B, improving on the MMLU benchmark by up to 3% compared to randomly shuffled data. However, Campos (2021) found no statistically significant improvements when evaluating curriculum learning using a similar difficulty metrics to Xu et al. (2020) in language modelling. Our work builds on previous studies by extending the evaluation across multiple models, learning strategies and data labelling scenarios for the task of domain-specific question answering.

3 Methods

3.1 Experimental design

We conducted a comprehensive investigation into the optimal data-ordering strategy, inspired by human learning, for fine-tuning language models in medical question answering. Our study compared the effectiveness of five specific human-inspired learning strategies with a Random Shuffled baseline (Section 3.2), across four LLMs (Section 3.5) and three datasets (Section 3.3), resulting in a total of 24 fine-tuned models (6 strategies × 4 models). We then evaluated these fine-tuned models on three different datasets in the medical domain. Additionally, we implemented human-inspired learning strategies with model-generated data labels, resulting in three distinct data-labelling scenarios (Section 3.4). This brings the total to 72 fine-tuned models.

3.2 Human-inspired learning strategies

Figure 1 defines the five learning strategies using data orderings that mimic common learning practices adopted by humans. These strategies are defined based on two data labels: (i) a continuous measure of question difficulty and (ii) a discrete category to which each question belongs. In particular, *Blocked Learning* and *Interleaved Learning* are solely defined by category and are non-curriculum-based, while the rest use the difficulty measure to define the curriculum. The design of the learning strategies was inspired by Lee et al. (2024), who proposed incorporating blocking and interleaving practices into the curriculum arrange-

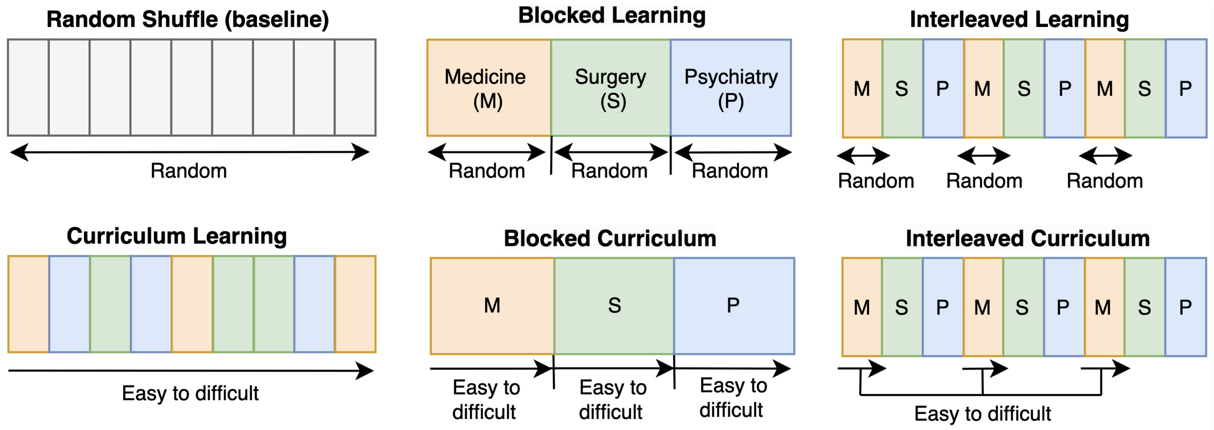


Figure 1: **Human-inspired learning strategies.** The five human-inspired learning strategies are demonstrated by ordering data based on a continuous measure of question difficulty (arranged by arrows) and category (indicated by block colours), alongside the *Random Shuffle* baseline. The first row presents non-curriculum-based strategies, and the second row presents curriculum-based strategies. (i) *Blocked Learning:* Questions are grouped by category, and randomised within each group. (ii) *Interleaved Learning:* Questions are grouped by category, then each category is randomly divided into three equal parts, and questions from each part are arranged in an interleaved manner. (iii) *Curriculum Learning:* Questions are sorted by difficulty in ascending order. (iv) *Blocked Curriculum:* Questions are grouped by category, and then arranged in ascending difficulty within each category. (v) *Interleaved Curriculum:* Following the Blocked Curriculum arrangement, questions in each category are further divided into three equal parts, and then interleaved.

ment. We modified their learning strategies by strictly sorting questions based on continuous values of question difficulty, instead of categorising questions into easy, medium, and hard classes, to avoid arbitrary distinctions between questions of similar difficulty.

The incorporation of human learning practices can potentialy improve the effectiveness of LLMs by structuring their learning process to promote better memory retention, generalisation, and prevent catastrophic forgetting (Luo et al., 2023). **Blocked Learning** groups questions by category, similar to blocked practice in education, where focusing on one subject at a time before moving to the next deepens understanding (North et al., 2017; Fazeli et al., 2017). **Interleaved Learning** mixes questions by different categories and revisits them periodically, similar to interleaved practice in education, which mitigates cognitive decay by bringing up old subjects and improves memory retention (Carvalho and Goldstone, 2014; Firth et al., 2019). **Curriculum Learning** sorts questions from easiest to hardest, similar to traditional educational curriculum where students build foundational knowledge before tackling more complex tasks (Wang et al., 2021). **Blocked Curriculum** combines the two by sorting questions within each category, allowing learners to build knowledge progressively in each category (Lee et al., 2024). **Interleaved Curriculum** (also called Spiral Curriculum), cycles through categories in rounds with increasing difficulty, mimicking the process of revisiting subjects with progressively challenging material to reinforce learning, and follows a global progression from simple to complex concepts across categories (Johnston, 2012).

3.3 Datasets

We fine-tuned on one medical question answering dataset, and evaluated on three to test generalisation. For fine-tuning, we used the Lekarski Egzamin Końcowy (LEK) dataset (Bean et al., 2024), which comprises of questions from the Polish medical licensing exams[1]. Unlike other medical multiple-choice datasets, LEK includes meta-information about human test takers' responses for each question, allowing us to assess question difficulty based on the actual performance of medical students. We used the English version of the questions from the last five exam sittings, between spring 2021 and spring 2023. The final dataset contains 874 unique questions divided into ten medical categories. For evaluation, we used the LEK dataset with cross-validation, as well as the official

[1]The LEK dataset is publically available at `https://cem.edu.pl/`

validation set of MedMCQA (Pal et al., 2022) [2], and the test set of MedQA (Jin et al., 2020), which are two popular medical question answering benchmarks.

3.4 Data labelling scenarios

We tested the effects of learning strategies defined by the following three data labelling scenarios on question difficulty and category:

- Difficulty defined by human responses and categories based on pre-existing labels (already exists in LEK);
- Difficulty defined by LLM responses and categories based on pre-existing labels;
- Difficulty defined by LLM responses and categories identified through clustering.

The automated data labels generated by LLM responses and clustering were tested to extend learning strategies to unlabelled data, where human annotations are expensive to obtain. The details of automated labelling are described below.

LLM-annotated question difficulty We prompted several general-purpose and medical LLMs to answer the questions in the training set, following the instruction prompt in Section 3.6. For each LLM, we computed an *expected accuracy* score for each question, defined as the probability that the LLM assigns to the correct choice index.

$$\mathbb{E}[\text{Acc}] = \sum_{c} P(c) \cdot \mathbb{1}(c = c^*), \tag{1}$$

where $P(c)$ is the probability assigned to choice $c \in \{A, B, C, D, E\}$, and $\mathbb{1}(c = c^*)$ is 1 if c is the correct answer c^*, otherwise 0. Essentially, this equates to the probability the model assigns to the correct answer.

The LLM-annotated difficulty for each question is defined as *(1 - expected accuracy)*, averaged across the LLMs. The LLMs used to compute difficulty on the LEK dataset are GPT-4 Turbo (OpenAI et al., 2024), GPT-3.5 (Brown et al., 2020), PaLM 2 (Anil et al., 2023), Mixtral 8x7B (Jiang et al., 2024), Meditron 70B (Chen et al., 2023b), and Llama 2 70B (Touvron et al., 2023). We present results using other ensemble models in Appendix A.4.

[2]Following Wu et al. (2023) and Chen et al. (2023b), we used the validation set as the MedMCQA test set does not publicly provide answer keys.

Clustering-based question categories To automate category assignment, we performed text clustering to group questions into semantically similar clusters, creating question categories based on the clustering. For clustering, we applied the BioMedBERT sentence embedding (Gu et al., 2020) to the question context and answer choices. We then used Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2020) for dimensionality reduction, followed by Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (McInnes and Healy, 2017). Although UMAP does not preserve pairwise distances, it retains global structure of data, making it suitable for clustering purposes. A density-based algorithm was chosen to handle noisy data and generate clusters with variable densities without specifying the number of clusters (Awasthi et al., 2013). Noise points identified by HDBSCAN were treated separately as an additional block in Blocked Learning. The hyperparameters of UMAP and HDBSCAN hyperparameters were optimised using Bayesian optimization to minimise the proportion of data points with a low probability (below 5%) of belonging to any cluster. The final hyperparameters for clustering are presented in Appendix A.3.

3.5 Language models

We used four open-source language models for fine-tuning: TinyLlama 1.1B (Zhang et al., 2024), Llama 2 7B (Touvron et al., 2023), Llama 2 13B (Touvron et al., 2023), and Mistral 7B (Jiang et al., 2023). Our selection ensures that we measure the effects of learning strategies across three varying sizes and two different model architectures. The TinyLlama 1.1B model follows the same architecture and tokenizer as Meta's Llama 2 models, with 1.1B parameters pre-trained on 3 trillion tokens (Zhang et al., 2024). For all Llama 2 models, we used the chat series optimised for dialogue, as they outperformed the base models in our experiments. All models were accessed via Hugging Face and fine-tuned on two NVIDIA RTX 6000 Ada cards. To optimise memory usage, the models loaded from Hugging Face were quantised to 4-bits with double quantization. We did not test larger models due to the computational costs of repeatedly fine-tuning large models.

3.6 Supervised fine-tuning

Instruction prompt We used zero-shot prompting for each question, starting with the following instruction:

Answer the following multiple-choice question by giving the most appropriate response. The answer should be one of [A, B, C, D, E].

This was followed by the question context and the multiple-choice answers. The response template began with *'Answer:'*, and the correct answer index was learned to be predicted as the next token during fine-tuning. Our zero-shot prompt structure is designed to reflect typical exam instructions and serves as a baseline for performance. The same prompt structure was used for inference, where the correct answer index was masked and predicted.

Fine-tuning method We employed the QLoRA (Dettmers et al., 2023) method for parameter-efficient fine-tuning of the linear layers in the LLMs. During fine-tuning, we disabled automatic data shuffling in PyTorch and trained the entire data sequence in a single epoch to maintain the specified learning order. Repeating the training across multiple epochs would violate the learning strategy design outlined in Figure 1; for example, looping through a Blocked practice multiple times would effectively turn it into an Interleaved practice. We also tried repeating the samples three times within each block to simulate multiple batches while maintaining the learning order, and obtained similar results to those with no data repetition. The hyperparameters used for fine-tuning each model are provided in Appendix A.3. To ensure fair comparisons, all learning strategies applied to a model used the same hyperparameters selected by grid search on the Random Shuffle baseline.

3.7 Model evaluation

Metrics for evaluation As we are dealing with multiple-choice questions with single-label answers, we relied on the LLMs to generate the next token as one of the option indexes [A, B, C, D, E] following the instruction prompt in Section 3.6. We used greedy decoding as the model's generated answer and compared it to the true answer to determine the *accuracy score*. To evaluate the effectiveness of learning strategies for fine-tuning, we calculated the *maximal accuracy gain* as the difference in accuracy score between the best-performing learning strategy and the Random Shuffle baseline.

Evaluation on learning strategies The accuracy score for each learning strategy and the Random Shuffle baseline was calculated by sampling each strategy five times and averaging the results. To ensure consistent comparisons, the category orders used in Blocked and Interleaved strategies remain consistent across all five samples.

4 Results and Discussion

4.1 Impact of fine-tuning with human-inspired learning strategies

Table 1 presents the accuracy scores for all learning strategies averaged across either datasets or models. Among the three data labelling scenarios, all learning strategies on average achieved a positive accuracy gain over Random Shuffle. The highest model-wise accuracy gain was 1.77%, and the highest dataset-wise accuracy gain was 1.81% (Table 2). Among the four models considered, TinyLlama-1.1B consistently demonstrated the highest accuracy gains (1.40%, 1.44%, and 1.77%) in all three sets of data labels. Following the definition of maximal accuracy gain (Section 3.7), the average maximal accuracy gain was 0.94% across models and 1.02% across datasets, both achieved with LLM-defined difficulty and pre-existing categories (Table 2).

	Data labelling scenarios		
Max accuracy gain	**(a)**	**(b)**	**(c)**
Top in models	1.40	1.44	**1.77**
Top in datasets	1.13	**1.81**	1.15
Average by models	**0.94**	**0.94**	0.80
Average by datasets	0.83	**1.02**	0.74

Table 2: **Maximal accuracy gains in models and datasets.** The *maximal accuracy gain* (in %) for a model or dataset is calculated as the difference between the best-performing learning strategy and the Random Shuffle baseline in Table 1. This table presents the maximal accuracy gains in models and datasets for three data labelling scenarios: (a) Human-defined difficulty and pre-existing categories, (b) LLM-defined difficulty and pre-existing categories, and (c) LLM-defined difficulty and clustered categories. *Top in models* and *Average by models* indicate the highest and average maximal accuracy gains across models, respectively. Similarly, *Top in datasets* and *Average by datasets* indicate the highest and average maximal accuracy gains across datasets.

Modest improvement of human-inspired learning strategies over Random Shuffle Overall, adopting a human-inspired learning strategy can yield an accuracy gain over Random Shuffle for any model or dataset when an appropriate learn-

Table 1: **Accuracy scores across models and datasets.** The accuracy scores (in %) for applying the human-inspired learning strategies in three data labelling scenarios shown in Tables (a)-(c). The scores are averaged across datasets and models. In the *Models* columns, accuracy scores are averaged across the three datasets for each model. In the *Datasets* columns, accuracy scores are averaged across the four models for each dataset. Learning strategies (in gray) in Tables (b) and (c) indicate unchanged results from Table (a) due to unchanged data labels. Abbreviations: *TinyLla.* = TinyLlama model, *Blocked Curri.* = Blocked Curriculum, *Interleaved Curri.* = Interleaved Curriculum, *AVG* = average.

Strategy	**Models**				**Datasets**			
	TinyLla. 1.1B	**Llama 2 7B**	**Llama 2 13B**	**Mistral 7B**	**LEK**	**Med MCQA**	**MedQA**	**AVG**
Random Shuffle	20.40	38.71	42.57	47.97	43.55	36.28	32.40	37.41
Curriculum	19.79	**39.05**	**43.68**	47.31	**44.68**	36.36	31.35	37.46
Blocked	20.47	38.46	42.83	48.10	43.99	36.45	31.97	37.47
Blocked Curri.	**21.80**	38.32	42.57	47.10	43.84	36.46	32.05	37.45
Interleaved	21.74	38.87	42.79	**48.88**	44.18	**37.04**	**32.99**	**38.07**
Interleaved Curri.	21.10	38.10	42.69	48.04	43.81	36.44	32.20	37.48

(a) Data labels: human-defined difficulty and pre-existing categories.

Strategy	**Models**				**Datasets**			
	TinyLla. 1.1B	**Llama 2 7B**	**Llama 2 13B**	**Mistral 7B**	**LEK**	**Med MCQA**	**MedQA**	**AVG**
Random Shuffle	20.40	38.71	42.57	47.97	43.55	36.28	32.40	37.41
Curriculum	20.88	**39.21**	42.82	48.39	**44.36**	36.86	32.26	37.83
Blocked	20.47	38.46	42.83	48.10	43.99	36.45	31.97	37.47
Blocked Curri.	**21.84**	37.89	42.67	48.71	43.64	37.20	32.51	37.78
Interleaved	21.74	38.87	42.79	48.88	44.18	37.04	**32.99**	38.07
Interleaved Curri.	21.67	<u>38.98</u>	**43.02**	**49.32**	44.22	**38.09**	32.43	**38.25**

(b) Data labels: LLM-defined difficulty and pre-existing categories.

Strategy	**Models**				**Datasets**			
	TinyLla. 1.1B	**Llama 2 7B**	**Llama 2 13B**	**Mistral 7B**	**LEK**	**Med MCQA**	**MedQA**	**AVG**
Random Shuffle	20.40	38.71	42.57	47.97	43.55	36.28	32.40	37.41
Curriculum	20.88	**39.21**	42.82	**48.39**	**44.36**	36.86	32.26	**37.83**
Blocked	20.95	38.23	**43.09**	47.94	43.22	36.77	**32.67**	37.55
Blocked Curri.	21.50	38.39	43.00	47.62	43.12	**37.43**	32.32	37.62
Interleaved	**22.17**	38.23	43.03	47.77	43.61	37.3	32.41	37.80
Interleaved Curri.	20.74	38.45	43.01	47.87	43.33	37.34	31.88	37.52

(c) Data labels: human-defined difficulty and clustered categories.

ing strategy is used. However, the optimal learning strategy is not consistent, which we will discuss in Section 4.2. The maximal accuracy gains are consistent in scale with the impact of curriculum learning found in some previous studies (up to 2%) (Maharana and Bansal, 2022; Xu et al., 2020), but are slightly lower than those reported by Lee et al. (Lee et al., 2024). Using similar Blocked and Interleaved Curriculum for fine-tuning Llama-13B on general knowledge tasks, their study showed Interleaved Curriculum consistently outperformed Blocked Curriculum, improving World Knowledge and Commonsense Reasoning benchmarks by 3.28% and 1.73%. We suspect two main reasons for the differences in results: a broader curriculum span and a clearer categorisation of difficulty levels. First, Lee et al. (Lee et al., 2024) used a synthetic dataset covering a wide range of subjects from secondary to graduate school levels, whereas our dataset focuses solely on graduate school medical exams, offering a narrower curriculum range. Additionally, they categorised questions into distinct difficulty levels of remembering, understanding, and applying knowledge based on Bloom's taxonomy (Bloom et al., 1956), while our medical questions are more semantically similar and lack such clear distinctions in difficulty. These factors likely contribute to the better performance of LLMs in curriculum-based learning in their study.

4.2 Generalisation of human-inspired learning strategies across contexts

As shown in Table 1, the accuracy gains over Random Shuffle varied significantly between models, and the best learning strategy was not consistent across models and datasets. Taking the case where we used human-defined difficulty and predefined categories as data labels (Table 1a), Curriculum Learning was the best learning strategy for Llama 2 7B and Llama 2 13B (+0.34 and +1.11), but failed to outperform the Random Shuffle for the other two models (-0.61 and -0.66). Among the four models, three different best learning strategies were identified, each achieving maximal accuracy gains for one or two models. However, only one strategy, Interleaved Learning, consistently outperformed Random Shuffle across all models. A similar pattern was observed in accuracy gains in datasets. Overall, Curriculum Learning scored the best strategy most often (8 out of 21 times), followed by Interleaved Learning (Table 1).

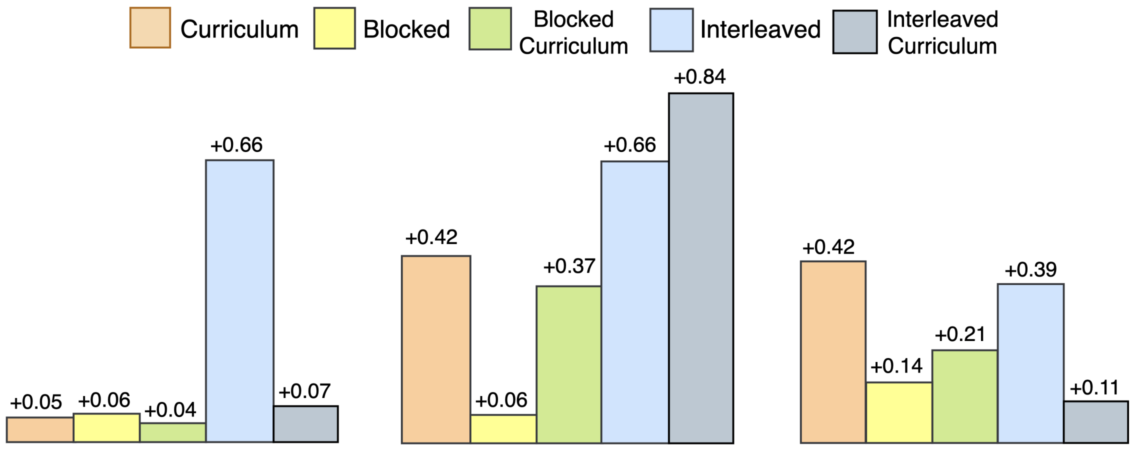
Variation of best learning strategy across models Most previous studies used a single model to examine the effectiveness of curriculum learning, consistently showing performance improvements on several data benchmarks (Xu et al., 2020; Maharana and Bansal, 2022; Lee et al., 2024). However, our study found that the best learning strategy for one model may not be optimal for another and may not even outperform the Random Shuffle baseline. Additionally, a strategy that consistently outperforms Random Shuffle across all models may not be the best for any specific model. Therefore, the effectiveness of a learning strategy for one model does not necessarily generalise to others.

Variation of best learning strategy across datasets We found that no single learning strategy was consistently the best across all datasets, even that strategy outperformed Random Shuffle on all datasets. This contrasts with the results of Lee et al. (Lee et al., 2024), where they found Interleaved Curriculum was consistently the best-performing strategy across multiple datasets compared to others. This discrepancy may be due to differences in experimental design, as discussed in Section 4.1. Although our results show that Interleaved Curriculum achieved the highest accuracy gain (+0.66) in Figure 2a, the margin of improvement compared to Lee et al. (Lee et al., 2024) was considerably smaller.

4.3 Performance of curriculum-based learning with LLM-defined difficulty

With pre-existing categories, we observed a modest accuracy increase in all curriculum-based learning strategies (Curriculum Learning, Blocked Curriculum, Interleaved Curriculum) when switching from human-defined to LLM-defined difficulty (Figure 2). With human-defined difficulty and pre-existing categories, only Interleaved Learning showed a noticeable accuracy improvement (+0.66) over Random Shuffle (Figure 2a). Upon switching to LLM-defined difficulty, there was an increase in accuracy across all three curriculum-based strategies: Curriculum Learning (+0.05 to +0.42), Blocked Curriculum (+0.04 to +0.37) and Interleaved Curriculum (+0.07 to +0.84) (Figure 2b). For each dataset, switching to LLM-defined difficulty resulted in the greatest increases for MedMCQA in Blocked Curriculum (+0.18 to +0.92) and Interleaved Curriculum (+0.16 to +1.81). For MedQA, the greatest increase was observed in Curriculum Learning (-1.05 to -0.14) (Appendix A.2). As a further evidence, we fine-tuned the MedQA training set (11.4k data) with the Mistral 7B model, the best-performing model among the four, using LLM-defined difficulty and clustered categories in an additional experiment (Appendix A.4). We again observed that Curriculum Learning (+0.70) consistently outperformed other learning strategies across all three datasets (Table 5). On the other hand, switching to clustered categories for fine-tuning had less noticeable effects on improving any specific learning strategy compared to using pre-existing categories (Figure 2c).

Potential of using LLM-defined difficulty for curriculum design These results indicate that using LLM responses to automatically generate a difficulty measure can enhance the effectiveness



(a) Human-defined difficulty + pre-existing categories

(b) LLM-defined difficulty + pre-existing categories

(c) LLM-defined difficulty + clustered categories

Figure 2: **Averaged accuracy gains for the learning strategies.** Each bar plot shows the accuracy gains (in %) over Random Shuffle, averaged over all model-data combinations for each learning strategy under the three data labelling scenarios.

of curriculum-based learning strategies, leading to more noticeable improvements. This aligns with the findings of previous studies that used language-model-ranked difficulty to define curriculum yielding consistent accuracy gains (Maharana and Bansal, 2022; Xu et al., 2020). These results suggest that model-generated difficulty may be a better indicator for training LLMs and highlight the potential of LLM-defined difficulty as a cost-effective alternative to human annotations for improved curriculum design.

4.4 Conclusions and future work

Our study conducted a comprehensive evaluation of fine-tuning LLMs with human-inspired learning strategies in medical question answering, focusing on four key dimensions: learning strategies, models, datasets, and data labelling scenarios. The main findings are as follows: First, human-inspired learning strategies showed moderate impacts, with the maximum accuracy gains of 1.77% per model and 1.81% per dataset. This indicates some transferability of human learning behaviours to LLMs in this task for data-efficient fine-tuning. Second, there was significant variability in the effectiveness of learning strategies across different models and datasets, with no single strategy universally outperforming the others. This suggests caution when generalising human-inspired learning strategies, as effectiveness for one model or dataset does not necessarily translate to others. Third, using LLM-defined difficulty metrics led to moderate accuracy improvements in the performance of curriculum-based learning strategies compared to human-defined difficulty. This highlights the potential of developing model-generated difficulty metrics to improve curriculum design over human-defined ones.

Future work could investigate the impacts of alternative clustering algorithms for fine-tuning. Given the broadness of clustering algorithms, a careful data sampling design could still lead to improved LLM performance. For example, Shao et al. (Shao et al., 2024) proposed ClusterClip Sampling, which balances common and rare samples during language model training based on clustered data distribution, outperforming random sampled data by 1%-2%. In addition, experiments could be extended to evaluate larger language models, such as those with 70B parameters, and specialised LLMs like medically fine-tuned models, to assess how model size and the amount of pre-trained knowledge affect the impact of learning strategies. Future experiments could also explore the temporal process of fine-tuning, investigating whether easy questions are answered correctly first and how the spectrum of correctly answered questions evolves throughout the fine-tuning process.

5 Limitations

We identify several limitations in our study design which may lead to result variations. First, we only ran the experiment five times for each learning strategy, and more repetitions would be needed for establishing more precise confidence intervals and statistical testing. Second, the LLM-defined difficulty measure relies on the choices of LLMs for response collection, and the results for clustered categories heavily depend on the clustering algorithm and its hyperparameters, both of which may introduce result variations. Third, the relatively small size of the LEK dataset for fine-tuning may limit the revelation of effects from learning strategies that may only emerge with more data points and longer training time. For example, the benefits of Interleaved Learning might become apparent over longer revision intervals and more frequent revision, which our dataset might not fully capture in the evaluation. Similarly, the span of question difficulties in the LEK dataset may be insufficient for effective Curriculum Learning. Future research could explore a curriculum that encompasses a broader spectrum of questions, spanning from fundamental medical concepts to advanced-level knowledge.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

R. Awasthi, A. Tiwari, and Seema Pathak. 2013. Analysis of mass based and density based clustering techniques on numerical datasets. *Journal of Information Engineering and Applications*, 3:29–34.

Andrew M. Bean, Karolina Korgul, Felix Krones, Robert McCraith, and Adam Mahdi. 2024. Exploring the landscape of large language models in medical question answering. *Preprint*, arXiv:2310.07225.

Benjamin S. Bloom, Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. McKay.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Daniel Campos. 2021. Curriculum learning for language modeling. *Preprint*, arXiv:2108.02170.

Paulo F. Carvalho and Robert L. Goldstone. 2014. Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology*, 5:936.

Mayee F. Chen, Nicholas Roberts, K. Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. 2023a. Skill-it! a data-driven skills framework for understanding and training language models. *ArXiv*, abs/2307.14430.

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*, abs/2310.16776.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

David Fazeli, Taheri Hamidreza, and Alireza Saberi Kakhki. 2017. Random versus blocked practice to enhance mental representation in golf putting. *Perceptual and Motor Skills*, 124:003151251770410.

Jonathan Firth, Ian Rivers, and James Boyle. 2019. A systematic review of interleaving as a concept learning strategy. *Social Science Protocols*, 2:1–7.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Peter Hase, Mohit Bansal, Peter Clark, and Sarah Wiegreffe. 2024. The unreasonable effectiveness of easy training data for hard tasks. *Preprint*, arXiv:2401.06751.

Jordan Hoffmann, Sebastian Borgeaud, A. Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, K. Simonyan, Erich Elsen, Jack W. Rae, O. Vinyals, and L. Sifre. 2022. Training compute-optimal large language models. *ArXiv*, abs/2203.15556.

Hong Jun Jeon and Benjamin Van Roy. 2022. An information-theoretic analysis of compute-optimal neural scaling laws. *ArXiv*, abs/2212.01365.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Howard Johnston. 2012. The spiral curriculum. Technical report, University of Florida. Accessed: June 9, 2024.

Bruce W. Lee, Hyunsoo Cho, and Kang Min Yoo. 2024. Instruction tuning with human curriculum. *Preprint*, arXiv:2310.09518.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *Preprint*, arXiv:2308.08747.

Adyasha Maharana and Mohit Bansal. 2022. On curriculum learning for commonsense reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 983–992, Seattle, United States. Association for Computational Linguistics.

Leland McInnes and John Healy. 2017. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 33–42. IEEE.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

Niklas Muennighoff, Alexander M. Rush, B. Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. *ArXiv*, abs/2305.16264.

Jamie North, Neil Bezodis, Colm Murphy, Oliver Runswick, Chris Pocock, and André Roca. 2017. The effect of consistent and varied follow-through practice schedules on learning a table tennis backhand.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun

Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

Gustavo Penha and Claudia Hauff. 2019. Curriculum learning strategies for ir: An empirical study on conversation response ranking. *Preprint*, arXiv:1912.08555.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, Si-Wai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of gemini models in medicine. *Preprint*, arXiv:2404.18416.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H. Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. 2024. How to train data-efficient llms. *Preprint*, arXiv:2402.09668.

Anu Sayal, Janhvi Jha, Chaithra N, Veethika Gupta, Ashulekha Gupta, Omdeep Gupta, and M. Memoria. 2023. Neural networks and machine learning. *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, pages 58–63.

Yunfan Shao, Linyang Li, Zhaoye Fei, Hang Yan, Dahua Lin, and Xipeng Qiu. 2024. Balanced data sampling for language model training with clustering. *Preprint*, arXiv:2402.14526.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:4555–4576.

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

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. Association for Computational Linguistics.

Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. 2023. To repeat or not to repeat: Insights from scaling llm under token-crisis. *ArXiv*, abs/2305.13230.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *Preprint*, arXiv:2401.02385.

A Appendix

A.1 Accuracy differences from Random Shuffle

We presented the accuracy differences (in %) of each learning strategy compared to the Random Shuffle baseline in Table 6, which follows a similar format to Table 1.

A.2 Accuracy gains by dataset

We presented a fine-grained analysis of the accuracy gains of learning strategies across each dataset in Figure 4, as an extension to Figure 2.

A.3 Hyperparameters for fine-tuning and clustering

We presented the hyperparameters for fine-tuning and clustering. For fine-tuning models, the fixed hyperparameters are as follows: For QLora, the parameters were set as $r = 16$, $\alpha = 64$ and dropout was set to 0.1. The optimizer used was AdamW. The learning rate decay followed a linear scheduler, the warmup steps were set to 0 and the maximum sequence length was set to 512. Table 3 shows the model-varying hyperparameters selected by grid search for each model. Table 4 shows the hyperparameters for clustering with UMAP and HDBSCAN, where the hyperparameters were selected using Bayesian Optimization within the specified ranges.

Table 3: **Model-varying hyperparameters for fine-tuning on LEK.** The hyperparameters were selected by grid search for each model on the Random Shuffle baseline. For fine-tuning Mistral 7B on the MedQA training set (Appendix A.4), we changed the learning rate to 1e-7 and kept the same batch size and gradient accumulation step. Abbreviations: *TinyLla.* = TinyLlama model, *Grad accum.* = gradient accumulation steps.

	TinyLla. 1.1B	**Llama 2 7B**	**Llama 2 13B**	**Mistral 7B**
Learning rate	5e-4	5e-5	1e-4	1e-4
Batch size	16	4	4	4
Grad accum.	1	2	2	2

Table 4: **Hyperparameters for clustering.** *Range* specifies the range of parameters for hyperparameter search, *Set* specifies the hyperparameter value chosen by Bayesian Optimization.

		LEK		**MedQA**	
		Range	Set	Range	Set
UMAP	**Number of Neighbours**	[8, 20]	15	[5, 30]	5
	Number of Components	[3, 15]	5	[3, 20]	17
HDBSCAN	**Minimum Cluster Size**	[25, 35]	25	[200, 250]	202

A.4 Results for fine-tuning on MedQA

As a further experiment, we presented the results for fine-tuning the MedQA training set (11.4k data) with the Mistral 7B model. We used LLM-defined difficulty and clustered categories, as the MedQA dataset does not contain pre-existing medical categories medqa. The LLMs used to compute the difficulty metrics are Mixtral 8x7B mixtral, Meditron 70B meditron, Llama 2 70B llama2 and Jamba jamba.

We observed that Curriculum Learning consistently outperformed other learning strategies across all three datasets (Table 5). Curriculum Learning also showed the highest accuracy gain over Random Shuffle (+0.70) compared to other learning strategies when averaged across all datasets (Figure 3).

Strategy	**LEK**	**Med MCQA**	**MedQA**	**AVG**
Random Shuffle	44.38	41.67	50.57	45.54
Curriculum	**45.40**	**42.19**	**51.14**	**46.24**
Blocked	44.76	41.70	50.71	45.72
Blocked Curri.	44.64	41.89	50.64	45.72
Interleaved	44.65	41.75	50.87	45.76
Interleaved Curri.	44.92	42.06	50.73	45.90

Table 5: **Accuracy scores of Mistral 7B fine-tuned on MedQA**. The accuracy scores (in %) were computed with LLM-defined difficulty and clustered categories as data labels.

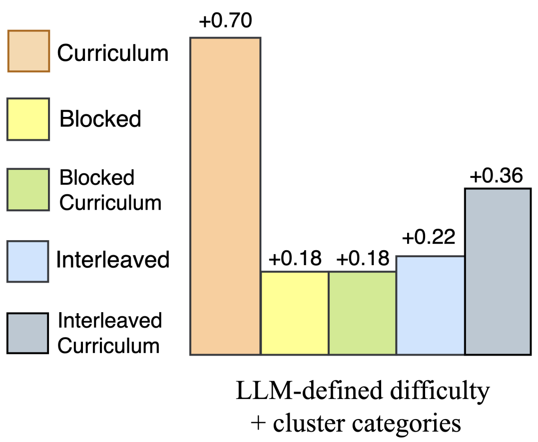


Figure 3: **Averaged accuracy gains of Mistral 7B fine-tuned on MedQA**. The bar plot shows the accuracy gains (in %) over Random Shuffle for each learning strategy, averaged across all datasets.

Table 6: **Accuracy differences compared to Random Shuffle baseline.** The accuracy differences (in %) of each learning strategy compared to Random Shuffle in three data-labelling scenarios shown in Tables (a)-(c). The accuracy difference for each model or dataset is calculated relative to the Random Shuffle baseline in Table 1. Learning strategies in gray in Tables (b) indicate unchanged results from Table (a) due to unchanged data labels. Abbreviations: *TinyLla.* = TinyLlama model, *Blocked Curri.* = Blocked Curriculum, *Interleaved Curri.* = Interleaved Curriculum, *AVG* = average.

Strategy	**Models**				**Datasets**			
	TinyLla. 1.1B	**Llama 2 7B**	**Llama 2 13B**	**Mistral 7B**	**LEK**	**Med MCQA**	**MedQA**	**AVG**
Curriculum	-0.61	**0.34**	**1.11**	-0.66	**1.13**	0.08	-1.05	0.05
Blocked	0.07	-0.25	0.26	0.13	0.44	0.17	-0.43	0.06
Blocked Curri.	**1.40**	-0.39	0.00	-0.87	0.29	0.18	-0.35	0.04
Interleaved	1.34	0.16	0.22	**0.91**	0.63	**0.76**	**0.59**	**0.66**
Interleaved Curri.	0.70	-0.61	0.12	0.07	0.26	0.16	-0.20	0.07

(a) Data labels: human-defined difficulty and pre-existing categories.

Strategy	**Models**				**Datasets**			
	TinyLla. 1.1B	**Llama 2 7B**	**Llama 2 13B**	**Mistral 7B**	**LEK**	**Med MCQA**	**MedQA**	**AVG**
Curriculum	0.48	**0.50**	0.25	0.42	**0.81**	0.58	-0.14	0.42
Blocked	0.07	-0.25	0.26	0.13	0.44	0.17	-0.43	0.06
Blocked Curri.	**1.44**	-0.82	0.10	0.74	0.09	0.92	0.11	0.37
Interleaved	1.34	0.16	0.22	0.91	0.63	0.76	**0.59**	0.66
Interleaved Curri.	1.27	0.27	**0.45**	**1.35**	0.67	**1.81**	0.03	**0.84**

(b) Data labels: LLM-defined difficulty and pre-existing categories.

Strategy	**Models**				**Datasets**			
	TinyLla. 1.1B	**Llama 2 7B**	**Llama 2 13B**	**Mistral 7B**	**LEK**	**Med MCQA**	**MedQA**	**AVG**
Curriculum	0.48	**0.50**	0.25	**0.42**	**0.81**	0.58	-0.14	**0.42**
Blocked	0.55	-0.48	**0.52**	-0.03	-0.33	0.49	**0.27**	0.14
Blocked Curri.	1.10	-0.32	0.43	-0.35	-0.43	**1.15**	-0.08	0.21
Interleaved	**1.77**	-0.48	0.46	-0.20	0.06	1.11	0.01	0.39
Interleaved Curri.	0.34	-0.26	0.44	-0.10	-0.22	1.06	-0.52	0.11

(c) Data labels: human-defined difficulty and clustered categories.

Figure 4: **Averaged accuracy gains for the learning strategies across datasets.** Each bar plot shows the accuracy gains (in %) for learning strategies over Random Shuffle across datasets. The results in each bar plot were averaged across models. Figures (a)-(c) represent three data labelling scenarios: (a) Human-defined difficulty with pre-existing categories; (b) LLM-defined difficulty with pre-existing categories; (c) LLM-defined difficulty with clustered categories.