Evaluating Fine-Tuning Efficiency of Human-Inspired Learning Strategies in Medical Question Answering

Yushi Yang[*], **Andrew M. Bean**, **Robert McCraith**, **Adam Mahdi**
Oxford Internet Institute
University of Oxford

Abstract

Fine-tuning Large Language Models (LLMs) incurs considerable training costs, driving the need for data-efficient training with optimised data ordering. Human-inspired strategies offer a solution by organising data based on human learning practices. This study evaluates the fine-tuning efficiency of five human-inspired strategies across four language models, three datasets, and both human- and LLM-labelled data in the context of medical question answering. These strategies achieve the best accuracy gain of 1.81% and an average gain of 1.02% across datasets, with interleaved strategies delivering the best average results. However, the best strategy varies across model-dataset combinations, limiting the generalisability of the effects of any single strategy. Additionally, LLM-defined question difficulty outperforms human-defined labels in curriculum-based learning, showing the potential of model-generated data as a cost-effective alternative for optimising fine-tuning. [2]

1 Introduction

Training large language models (LLMs) is costly, both in terms of data collection efforts [28, 38] and computational resources [15, 16]. To mitigate these costs, recent research has focused on data-efficient training methods that optimise data ordering [33]. One promising approach is human-inspired learning strategies, which organise data based on human learning principles [40, 37]. Among these strategies, curriculum learning, where samples are arranged from easiest to hardest, is particularly popular [13, 37]. This method has proven effective in language tasks such as knowledge acquisition [21], natural language reasoning [24], and information retrieval [32].

While curriculum learning has shown promise, the broader potential of human-inspired data ordering for fine-tuning remains unclear. Specifically, previous studies have focused on (i) a narrower set of models [24, 37], with fewer comparisons across multiple models; (ii) limited evaluation in domain-specific applications, especially in high-stakes areas like medical question answering; and (iii) primarily using one set of data labels, with less emphasis on comparing human- versus machine-defined labels to reflect different perspectives on data ordering. To address these gaps, we conduct a comprehensive evaluation of human-inspired learning strategies for fine-tuning in medical question answering, testing models of various sizes using both human- and LLM-defined question difficulty labels. We focus on the medical question-answering domain because of the scarcity of high-quality medical questions needed to train effective medical LLMs, highlighting the importance of efficient fine-tuning strategies [22, 25]. Our findings offer empirical insights into the efficiency and generalisability of these strategies for fine-tuning LLMs. Specifically, we find that:

[*]Correspondence to: Yushi Yang, <yushi.yang@oii.ox.ac.uk>

[2]The code is available at: https://github.com/Oxford-AI-for-Society/human-learning-strategies/.

38th Conference on Neural Information Processing Systems (NeurIPS 2024)
Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability (FITML)

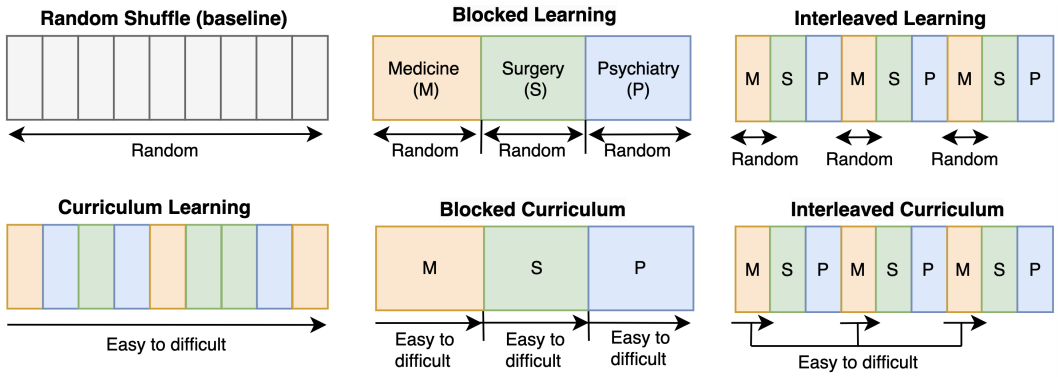


Figure 1: **Human-inspired learning strategies.** Along the *Random Shuffle* baseline, five human-inspired learning strategies order data based on question difficulty (shown with arrows) and question category (represented by block colors). The first row shows non-curriculum-based strategies, while the second row shows curriculum-based strategies.

- *Human-inspired learning yields moderate improvements over random shuffling.* These strategies result in the best accuracy gain of 1.81% and an average gain of 1.02% across datasets, with interleaved strategies providing the best average results.
- *Human-inspired learning lacks generalisation across model-data combinations.* The best strategy varies across model-dataset combinations, suggesting caution when generalising the effects of any one strategy to other models based on single-model results.
- *LLM-defined difficulty outperforms human labels in curriculum-based learning.* We automatically labelled question difficulty using an ensemble of LLM responses. The results show that switching to LLM-defined difficulty modestly improves the performance of curriculum-based strategies, offering a cost-effective alternative to human annotations for optimising fine-tuning.

2 Experimental design

We evaluated the efficiency of human-inspired data ordering strategies for fine-tuning LLMs in medical question answering. Our study compared 5 human-inspired learning strategies against a Random Shuffle baseline (Section 2.1) across 4 LLMs (Section 2.2) and 3 medical datasets (Section 2.3). We also applied these strategies with both human- and model-generated labels on question difficulty, creating 2 data-labelling scenarios (Section 2.5). This results in 144 fine-tuned models (144 = 6 strategies × 4 models × 3 datasets × 2 data labels).

2.1 Human-inspired learning strategies

Figure 1 presents five human-inspired learning strategies. These strategies are defined based on two data labels: (i) a continuous measure of *question difficulty* and (ii) a discrete measure of *question category*. Among them, Blocked Learning and Interleaved Learning rely solely on question category and do not follow a curriculum. The remaining strategies use the difficulty measure to create a curriculum, ranking questions from easiest to hardest.

The five learning strategies are based on common educational practices for human learning. *Blocked Learning* groups questions by category, like blocked practice in education [29, 9], where focusing on one topic deepens understanding. *Interleaved Learning* alternates questions from different categories, revisiting them periodically to reduce cognitive decay and improve retention [6, 10]. *Curriculum Learning* orders questions from easiest to hardest by question difficulty, akin to traditional curricula where basic knowledge precedes complex tasks [35]. *Blocked Curriculum* combines blocked and curriculum learning, arranging questions by difficulty within each category to progressively build knowledge [21]. *Interleaved Curriculum* cycles through categories with increasing question difficulty, reinforcing learning by revisiting topics with harder material [20].

2.2 Language models

We fine-tuned four open-source language models of varying sizes and architectures: TinyLlama 1.1B [39], Llama 2 7B [34], Llama 2 13B [34], and Mistral 7B [17]. TinyLlama shares the architecture of Llama 2 with only 1.1B parameters [39]. For the two Llama 2 models, we used the chat versions optimised for dialogue, focusing on fine-tuning medical knowledge rather than instruction styles. All models were accessed via Hugging Face and fine-tuned on two NVIDIA RTX 6000 Ada cards. To optimise memory, the models were quantised to 4-bits with double quantisation.

2.3 Datasets

We fine-tuned on a single medical question answering dataset with human-defined labels on question difficulty and categories, then evaluated on three datasets to test the models' generalisation. For fine-tuning, we used the Lekarski Egzamin Końcowy (LEK) dataset [4], consisting of multiple-choice questions from Polish medical licensing exams[3]. Unlike other medical exam datasets, LEK includes test-takers' responses for each question, allowing us to assess question difficulty based on actual human performance. We used the English version of questions from the last five exams (spring 2021 to spring 2023), resulting in a final dataset of 874 unique questions across ten medical categories. For evaluation, we used (i) the LEK test set with cross-validation, (ii) the official MedMCQA validation set [31][4], and (iii) the official MedQA test set [19].

2.4 Supervised fine-tuning

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

We used the QLoRA [8] method for parameter-efficient fine-tuning on the linear layers of the language models. To preserve the specified learning order in Figure 1, we disabled automatic data shuffling when loading data and trained the entire sequence in a single epoch. Training over multiple epochs would disrupt the learning order specified; for example, repeating Blocked practice multiple times would effectively turn it into an Interleaved practice, and re-running a curriculum would cause jumps from difficult to easy questions. We also experimented with repeating samples three times within each category to increase learning data, which yielded similar performance to no repetition. The fine-tuning hyperparameters, selected via grid search, are provided in Appendix B.

2.5 Data labelling scenarios

We evaluated the effectiveness of learning strategies in two data labelling scenarios: with (a) human-defined question difficulty and (b) LLM-defined question difficulty. Human-defined difficulty is calculated as the proportion of incorrect answers from human test takers for each question in the LEK dataset, with a higher error rate indicating more difficult questions. We tested LLM-defined difficulty to extend learning strategies to unlabelled data, where human annotations are costly, which is especially true in high-stakes fields such as medical question answering.

We prompted six LLMs to answer the questions in the LEK dataset, using the instruction prompt in Section 2.4. For each LLM, we calculated an *expected accuracy* for each question as the probability the LLM assigns to the correct answer during next-token prediction:

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

where $P(c)$ is the probability assigned to answer index $c \in \{A, B, C, D, E\}$, and $\mathbb{1}(c = c^*)$ is 1 if c is the correct answer c^*, otherwise 0. The LLM-defined difficulty for each question is defined as

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

[4]We used the validation set for evaluation following Wu et al. [36] and Chen et al. [7], as the test set does not publicly provide answer keys.

Table 1: **The accuracy scores (in %) for learning strategies averaged across datasets or models.** This table shows that the best strategy varies across models and datasets. In the *Models* columns, scores are averaged over three datasets; in the *Datasets* columns, scores are averaged over four models. The *Mean* column shows the accuracy scores for a learning strategy averaged across all model-dataset combinations. Note that here *Tiny* stands for TinyLlama.

Strategy	Models				Datasets			Mean
	Tiny 1.1B	Llama 2 7B	Llama 2 13B	Mistral 7B	LEK	Med MCQA	Med QA	
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 label: human-defined difficulty

Strategy	Models				Datasets			Mean
	Tiny 1.1B	Llama 2 7B	Llama 2 13B	Mistral 7B	LEK	Med MCQA	Med QA	
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	38.98	**43.02**	**49.32**	44.22	**38.09**	32.43	**38.25**

(b) Data label: LLM-defined difficulty

(1 - expected accuracy), averaged across the LLMs. The models used to compute difficulty include GPT-4 Turbo [30], GPT-3.5 [5], PaLM 2 [2], Mixtral 8x7B [18], Llama 2 70B [34], and Meditron 70B [7], covering a wide range of architectures and pre-training medical knowledge. These models differ from those used for fine-tuning to ensure an unbiased assessment of question difficulty. They showed reasonable per-question agreement with an average correlation of 0.54.

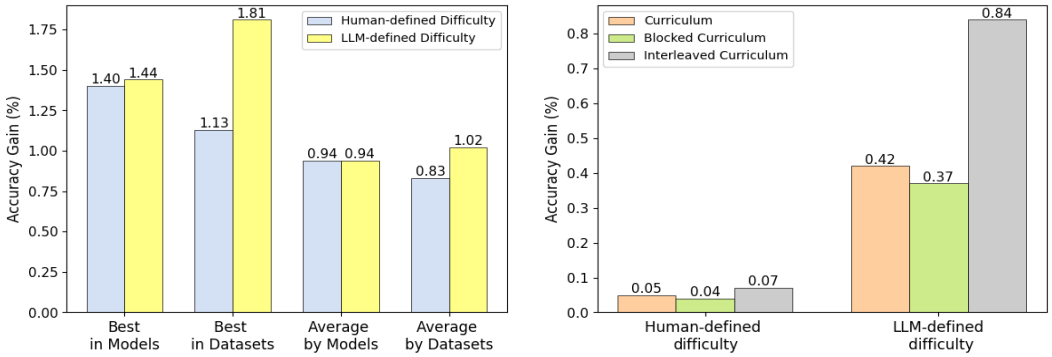
2.6 Evaluation metrics

We used greedy decoding to generate the model's answer index and compared it to the correct answer to calculate the *accuracy score*. For each model, accuracy scores for a learning strategy were averaged across all datasets, and for each dataset, scores were averaged across all models (Table 1). For each model or dataset, the *accuracy gain* is measured as the difference between the best-performing strategy and the Random Shuffle baseline. The *mean accuracy gain* of a strategy is measured as its accuracy gains averaged across all model-dataset combinations (the 'Mean' column in Table 1). Accuracy scores for each strategy across each model-dataset combination were averaged over five runs, with question category orders in Blocked and Interleaved Learning kept consistent across runs.

3 Results

Impact of human-inspired learning strategies Overall, adopting a human-inspired learning strategy yields the best accuracy gain of 1.81% and an average gain of 1.02% across datasets (1.44% and 0.94% across models) over Random Shuffle (Figure 2a), both significant at the 5% level in a paired t-test. Switching to LLM-defined difficulty leads to higher accuracy gains both in models and datasets (Figure 2a). Model-wise, TinyLlama-1.1B gave the highest accuracy gains in models (1.40% and 1.44%) in both data labelling scenarios.

Generalisation of human-inspired learning strategies Table 1 presents the accuracy scores for all learning strategies, averaged across either datasets or models, with a total of 14 models and datasets



(a) Effects of human-inspired learning strategies.

(b) Accuracy gains for curriculum-based learning.

Figure 2: **Comparison of accuracy gains from learning strategies with human- and LLM-defined difficulty.** (a) shows the highest and average accuracy gains (in %) of the best strategy across models and datasets, compared to the Random Shuffle baseline. (b) shows that using LLM-defined difficulty improves accuracy scores for all curriculum-based strategies, with each bar showing the mean accuracy gains (in %) across all model-data combinations.

shown in the table columns. Overall, Curriculum Learning was the top strategy in 5 out of 14 cases, followed by Interleaved Learning (4 cases) and Interleaved Curriculum (3 cases). Notably, Interleaved Learning consistently outperformed Random Shuffle across all models and datasets in both labelling scenarios (Table 1a and 1b). However, when switching to LLM-defined difficulty, the best strategy shifts from Interleaved Learning to Interleaved Curriculum (Table 1).

Moreover, there is no single strategy performing the best across all models in either data labelling scenario. For example, with human-defined difficulty (Table 1a), while Interleaved Learning consistently outperformed Random Shuffle, it wasn't always the best for specific models. Curriculum Learning was most effective for Llama 2 7B and Llama 2 13B (up to +1.11%) but under-performed compared to Random Shuffle on the other two models (up to -0.66%). This suggests that strategies effective for one model may not generalise well to others, a trend also seen across datasets.

We also experimented with automating question category labelling using text clustering to group semantically similar questions. Switching to clustering-based question categories yielded accuracy gains at a similar scale (best accuracy gain of 1.77% across models and 1.15% across datasets), with a similar pattern observed, where no single strategy consistently outperformed the others across models or datasets (Appendix C).

Performance of LLM-defined difficulty Switching from human-defined to LLM-defined difficulty increased accuracy scores across all three curriculum-based strategies (Figure 2b), with all gains significant at the 10% level in a paired t-test. The largest improvements were seen in MedMCQA for Blocked Curriculum (+0.74%) and Interleaved Curriculum (+1.65%), and in MedQA for Curriculum Learning (+0.91%). Additionally, fine-tuning Mistral 7B, our best-performing model, on the MedQA training set (11.4k data) using LLM-defined difficulty showed that curriculum learning outperformed all other strategies across all datasets (Appendix D). These findings suggest that using LLM-defined difficulty can improve the performance of curriculum-based learning strategies over human labels.

4 Discussion

Level of improvements The highest accuracy gains in our study (1.81% and 1.44%) align with previous findings on curriculum learning. Xu et al. [37] reported a 1.3% improvement in natural language understanding tasks by using multiple teacher models to define question difficulty for BERT fine-tuning, while Maharana and Bansal [24] showed a 2% gain in common sense reasoning with RoBERTa fine-tuned on fixed and adaptive curricula. Our results are lower than Lee et al. [21], who reported gains of 3.28% and 1.73% on World Knowledge and Commonsense Reasoning benchmarks for Llama-13B using Interleaved Curriculum.

Two key factors may explain the difference. First, Lee et al. [21] used a broader curriculum, covering subjects from secondary to graduate level, while we focused on a specialised graduate-level medical curriculum, with a narrower scope of content. Second, their approach used Bloom's taxonomy to

classify questions by distinct difficulty levels, whereas our dataset, with more semantically similar medical questions, exhibits subtler difficulty variations. These differences in curriculum scope and difficulty categorisation may account for the variation in performance gains.

Limited generalisation Most previous studies demonstrated the effectiveness of curriculum learning using a single model, showing consistent improvements across several benchmarks [37, 24, 21]. However, these results are insufficient to extrapolate that curriculum learning generalises well to other models, even though it is often assumed in these studies. Our findings suggest that a strategy effective for one model does not generalise to others, emphasising the need for caution when generalising results and drawing conclusions from limited model-specific evidence. This limited generalisation across models and datasets may be due to variations in how different LLMs perceive question difficulty and inconsistencies in question categories across medical datasets.

Benefits of LLM-defined labels Our results show that using LLM-generated difficulty modestly improves the performance of curriculum-based learning strategies. This aligns with previous studies, which found that language-model-ranked difficulty led to consistent accuracy gains in curriculum learning across benchmarks [24, 37]. Unlike those studies, which first fine-tuned models with randomly shuffled data before ranking difficulty, we measured the difficulty directly, leveraging the pre-training knowledge of LLMs to define the labels. This demonstrates the potential of LLM-defined difficulty labels as a cost-effective alternative to human labels for optimising fine-tuning order in curriculum-based learning.

5 Limitations

Fine-tuning data size may affect performance The relatively small size of the LEK dataset, due to the limited human labels on medical questions, may restrict the visibility of certain effects of learning strategies, especially those that only emerge with larger data. For example, the benefits of Interleaved Learning might become only apparent over longer revision intervals, which our dataset might not fully capture. Similarly, the relative narrow range of question difficulties in the LEK dataset may limit the effectiveness of curriculum-based learning.

Alternative notions for defining LLM-based difficulty The LLM-defined question difficulty was labelled based on the average accuracy on questions from ensemble LLMs, which may not fully capture all aspects of difficulty. Alternative measures, such as topic-specific perplexity to assess topic familiarity or the average language modeling loss on sequences as an indicator of LLMs' pre-training medical knowledge, could be explored further.

Limited exploration of fine-tuning methods Our study focused exclusively on supervised fine-tuning using the QLoRA approach. Future work could explore the effects of other fine-tuning methods, such as domain-adaptive pretraining (DAPT) [12], continual learning [1], or adapter-based fine-tuning [14].

6 Conclusions

We evaluated human-inspired learning strategies for fine-tuning LLMs in medical question answering, with the evaluation conducted across four dimensions: learning strategies, models, datasets, and data labelling scenarios. Our findings show moderate improvements (up to 1.81%) in fine-tuning with those strategies, suggesting some transferability of human learning behaviours to LLMs for more data-efficient tuning. However, the best learning strategy varied significantly across model-dataset combinations, with no single strategy consistently outperforming others, indicating limited generalisation of their effects across contexts. Additionally, we find that using LLM-defined difficulty measures provide moderate accuracy improvements over human-defined difficulty, highlighting their potential as a cost-effective alternative to human labels for optimising fine-tuning.

References

[1] Rahaf Aljundi. Continual learning in neural networks, 2019. *arXiv:1910.02718*.

[2] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, et al. Palm 2 technical report, 2023. *arXiv:2305.10403*.

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

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

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. Language models are few-shot learners, 2020. *arXiv:2005.14165*.

[6] Paulo F. Carvalho and Robert L. Goldstone. Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology*, 5:936, 2014. doi: 10.3389/fpsyg.2014.00936.

[7] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, et al. Meditron-70b: Scaling medical pretraining for large language models, 2023. *arXiv:2311.16079*.

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

[9] David Fazeli, Taheri Hamidreza, and Alireza Saberi Kakhki. Random versus blocked practice to enhance mental representation in golf putting. *Perceptual and Motor Skills*, 124: 003151251770410, 04 2017. doi: 10.1177/0031512517704106.

[10] Jonathan Firth, Ian Rivers, and James Boyle. A systematic review of interleaving as a concept learning strategy. *Social Science Protocols*, 2:1–7, 2019. doi: 10.7565/ssp.2019.2650.

[11] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, October 2021. ISSN 2637-8051. doi: 10.1145/3458754.

[12] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, et al. Don't stop pretraining: Adapt language models to domains and tasks, 2020. *arXiv:2004.10964*.

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

[14] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Boshen Ding, et al. On the effectiveness of adapter-based tuning for pretrained language model adaptation, 2021. *arXiv:2106.03164*.

[15] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, et al. Training compute-optimal large language models, 2022. *arXiv:2203.15556*.

[16] Hong Jun Jeon and Benjamin Van Roy. An information-theoretic analysis of compute-optimal neural scaling laws, 2023. *arXiv:2212.01365*.

[17] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, et al. Mistral 7b, 2023. *arXiv:2310.06825*.

[18] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, et al. Mixtral of experts, 2024. *arXiv:2401.04088*.

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

[20] Howard Johnston. The spiral curriculum, 2012. URL `https://files.eric.ed.gov/fulltext/ED538282.pdf`. Accessed: 2-Nov-2024.

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

[22] Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, et al. Do we still need clinical language models?, 2023. *arXiv:2302.08091*.

[23] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, et al. Jamba: A hybrid transformer-mamba language model, 2024. *arXiv:2403.19887*.

[24] Adyasha Maharana and Mohit Bansal. 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, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.72.

[25] Supun Manathunga and Isuru Hettigoda. Aligning large language models for clinical tasks, 2023. *arXiv:2309.02884*.

[26] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 33–42. IEEE, 2017. doi: 10.1109/ICDMW.2017.12.

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

[28] Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, et al. Scaling data-constrained language models, 2023. *arXiv:2305.16264*.

[29] J.S. North, N.E. Bezodis, C.P. Murphy, O.R. Runswick, C. Pocock, et al. The effect of consistent and varied follow-through practice schedules on learning a table tennis backhand. *Journal of Sports Sciences*, 37(6):613–620, 2018. doi: 10.1080/02640414.2018.1522683.

[30] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. Gpt-4 technical report, 2023. *arXiv:2303.08774*.

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

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

[33] Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, et al. How to train data-efficient llms, 2024. *arXiv:2402.09668*.

[34] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. *arXiv:2307.09288*.

[35] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:4555–4576, 2021. doi: 10.1109/TPAMI.2021.3069908.

[36] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, et al. Pmc-llama: Towards building open-source language models for medicine, 2023. *arXiv:2304.14454*.

[37] Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, et al. Curriculum learning for natural language understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.542.

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

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

[40] Tianhui Zhang, Danushka Bollegala, and Bei Peng. Learning to predict concept ordering for common sense generation, 2023. *arXiv:2309.06363*.

A Mean accuracy gains across data labelling scenarios

In this section, we present detailed results on the mean accuracy gains (in %) of learning strategies across three data labelling scenarios: (a) Human-defined difficulty with human-defined categories, (b) LLM-defined difficulty with human-defined categories, and (c) LLM-defined difficulty with clustered categories, as shown in Figure 3.

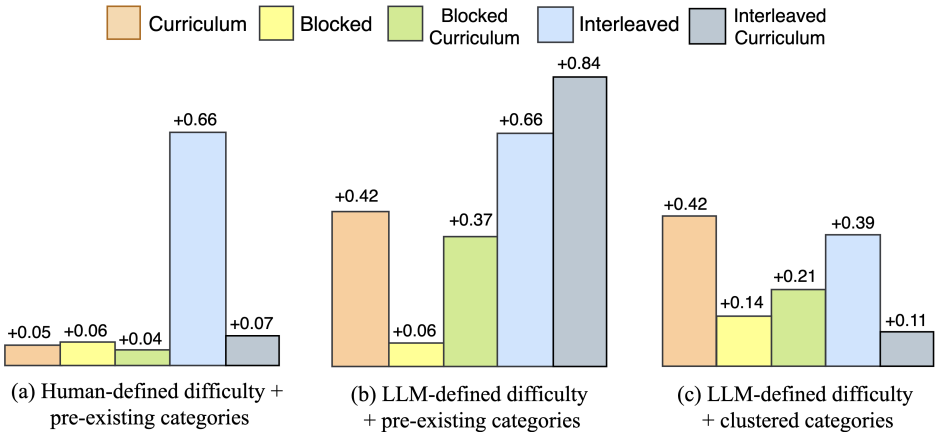


Figure 3: **Mean accuracy gains for the learning strategies.** Each bar plot shows the mean accuracy gains (in %) of the learning strategies when averaged across model-dataset combinations. Figures (a)-(c) represent three data labelling scenarios.

B Hyperparameters for fine-tuning

In this section, we present the hyperparameters for fine-tuning LLMs. The fixed hyperparameters for QLoRA were set as follows: $r = 16$, $\alpha = 64$, with a dropout of 0.1. AdamW was used as the optimizer, and the learning rate decay followed a linear schedule with no warmup steps. The maximum sequence length was set to 512. Table 2 details the model-specific hyperparameters selected via grid search for each model on the Random Shuffle baseline. The grid search ranges were: learning rate in [5e-4, 1e-4, 5e-5, 1e-5, 5e-6, 1e-6, 5e-7, 1e-7], batch size in [4, 8, 16], and gradient accumulation steps in [1, 2, 4].

Table 2: **Model-specific hyperparameters for fine-tuning**. For each model, the hyperparameters were selected by grid search on the Random Shuffle baseline. Abbreviations: *Grad accum.* = gradient accumulation steps. *TinyLla.* = TinyLlama.

	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

C Automatic labelling on question categories

In this section, we detail the process of automating question category assignment using text clustering to group semantically similar questions. First, we applied BioMedBERT sentence embeddings [11] to the question context and answers. We then used Uniform Manifold Approximation and Projection (UMAP) [27] for dimensionality reduction, followed by Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [26] for clustering. HDBSCAN was chosen for its ability to handle noisy data and form clusters with variable densities without requiring a pre-defined number of clusters [3]. Table 3 shows the accuracy scores of learning strategies with

LLM-defined difficulty and clustered categories. Table 4 shows the hyperparameters for clustering with UMAP and HDBSCAN for automated category labelling. The hyperparameters of UMAP and HDBSCAN were optimised using Bayesian optimisation to minimise the number of data points with low cluster probability (below 5%).

Table 3: **The accuracy scores (in %) for learning strategies averaged across datasets or models, using LLM-defined difficulty and clustered categories as labels.** The best strategy still varies across models and datasets. In the *Models* columns, scores are averaged over three datasets; in the *Datasets* columns, scores are averaged over four models.

Strategy	**Models**				**Datasets**		
	TinyLla. 1.1B	**Llama2 7B**	**Llama2 13B**	**Mistral 7B**	**LEK**	**Med MCQA**	**Med QA**
Random Shuffle	20.40	38.71	42.57	47.97	43.55	36.28	32.40
Curriculum	20.88	**39.21**	42.82	**48.39**	**44.36**	36.86	32.26
Blocked	20.95	38.23	**43.09**	47.94	43.22	36.77	**32.67**
Blocked Curri.	21.50	38.39	43.00	47.62	43.12	**37.43**	32.32
Interleaved	**22.17**	38.23	43.03	47.77	43.61	37.39	32.41
Interleaved Curri.	20.74	38.45	43.01	47.87	43.33	37.34	31.88

Table 4: **Hyperparameters for clustering with UMAP and HDBSCAN.** The *Range* row specifies the range for hyperparameter search in Bayesian Optimisation, the *Set* row specifies the hyperparameter value selected.

		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

D Fine-tuning on MedQA training set

In this section, we present the results of fine-tuning the MedQA training set (11.4k data points) using the Mistral 7B model, our best-performing model. We employed LLM-defined difficulty and clustered categories as data labels, as the MedQA dataset lacks pre-existing human-defined medical categories [19]. The LLMs used to compute difficulty measures were Mixtral 8x7B [18], Meditron 70B [7], Llama 2 70B [34], and Jamba [23]. Table 5 shows that Curriculum Learning consistently outperformed other strategies across all three datasets, with the highest average accuracy gain over Random Shuffle (+0.70%).

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

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