Balancing Knowledge Breadth and Task Depth for Effective Domain Adaptation Fine-Tuning

Mu Zhang[1]*, Yuxiang Chu[1]*, Guangya Yu[1], Yongqi Fan[1], Weiyan Zhang[1], Hang Hu[1], Tong Ruan[1]†, Jingping Liu[2]†

[1]School of Information Science and Engineering
East China University of Science and Technology,
[2]School of Software Engineering, Sun Yat-sen University

Correspondence: zhang.mu@foxmail.com, ruantong@ecust.edu.cn

Abstract

Training large language models for domain adaptation poses a significant challenge in balancing the acquisition of domain knowledge with the retention of general abilities, often leading to catastrophic forgetting. While curriculum learning offers a promising direction, conventional methods typically rely on a single dimension of knowledge or task, which is insufficient to navigate the trade-off between knowledge breadth and task depth. In this paper, we propose a two-dimensional curriculum learning framework that coordinates model training along two orthogonal axes: the knowledge dimension and the task dimension. We first reconstruct the dataset by clustering instances according to their semantic similarity to general-domain data, and subsequently annotate them with a task hierarchy. Then, we design an integrated curriculum that develops from general to domain-specific knowledge clusters, and within each cluster, from lower- to higher-order cognitive tasks. Compared with the second-best method, our method improves accuracy on medical evaluations by 2.49% and on financial evaluations by 1.2%. Ablation and cross-domain experiments further demonstrate our method as a scalable and effective framework for structured domain adaptation in large language model fine-tuning. We have released the code in an anonymous repository at https://github.com/Melo-1017/Balancing-Knowledge-Breadth-and-Task-Depth.

1 Introduction

Recently, large language models (LLMs) have demonstrated strong performance across a broad spectrum of general tasks and emerged as the cornerstone of modern natural language processing (Ouyang et al., 2022; Achiam et al., 2023; Guo et al., 2025). However, practical applications in domains such as medicine (Chen et al., 2024b), finance (Xie et al., 2023), and law (Hou et al., 2025b) require specialized expertise beyond general language proficiency. Recent studies further show that medical deployment often involves multi-step clinical reasoning and diagnosis, which places higher demands on domain-adapted LLMs (Hou et al., 2025a). A common solution is to fine-tune LLMs on domain-specific corpora, which enhances in-domain performance but often degrades general abilities due to catastrophic forgetting (CF) (Luo et al., 2025; Song et al., 2025; Li et al., 2024). Therefore, a key challenge in domain adaptation for LLMs lies in balancing the acquisition of domain-specific knowledge with the preservation of general-purpose abilities.

Curriculum learning (CL) is a potential solution for addressing this challenge (Zhang et al., 2019). By assigning a difficulty score to each training instance and presenting data from "easy" to "hard", CL aims to guide models through a gradual, structured learning process (Chen et al., 2025). Existing curriculum-learning methods typically rely on a single difficulty dimension, such as task-based ordering (Yue et al., 2024) or domain-level sequencing (Dong et al., 2024). While effective in some settings, these one-dimensional designs make it difficult to simultaneously preserve general reasoning abilities and deepen domain expertise.

Insights from human learning indicate that knowledge acquisition often follows two complementary trajectories. Learners progress horizontally from general to specialized knowledge, while vertically advancing from simple to complex cognitive tasks (Bloom et al., 1956). This dual process broadens knowledge and deepens understanding, helping retain general skills while developing expertise. Guided by this principle, we explore how a two-dimensional curriculum can jointly organize training data along knowledge and task dimensions to achieve a balanced development of generality and specialization.

* Equal Contribution.
† Corresponding Authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 8287–8304

In this paper, we propose a two-dimensional curriculum learning framework for domain adaptation of LLMs. We leverage the bge-m3 (Chen et al., 2024a) model to embed both general-domain and domain-specific corpora into a shared semantic space and apply K-means clustering to form hierarchical knowledge levels. Within the domain-specific subset, we leverage the Qwen3-8B (Yang et al., 2025) model to annotate each instance with a task label derived from Bloom's taxonomy, grouped into three task complexity levels. Based on this reconstructed data, we design a curriculum that, at the macro level, orders clusters from general-like to domain-specific, and at the micro level, sequences samples within each cluster from lower- to higher-order tasks. Finally, an exponential mixing schedule gradually increases the proportion of domain data during supervised fine-tuning, enabling a smooth transition from general to domain learning.

Extensive experiments on medical, financial, and general benchmarks validate the effectiveness of this framework. Our method yields substantial improvements in domain-specific evaluations while maintaining strong performance on general benchmarks, and ablation and scaling studies highlight the complementary roles of knowledge- and task-level orderings in mitigating catastrophic forgetting and enhancing cross-domain robustness.

Contributions. The main contributions of this work are summarized as follows:

1. We introduce a two-dimensional curriculum that combines knowledge and task dimensions, jointly organizing training data by semantic similarity and task complexity to balance general reasoning and domain specialization.

2. We instantiate this framework with a unified semantic space for hierarchical knowledge clustering, Bloom's-taxonomy-based task labeling, and an exponential mixing schedule that gradually shifts from general to domain data.

3. Extensive experiments across medical, financial, and general benchmarks demonstrate consistent gains while preserving general abilities. Notably, our method improves Qwen3-8B accuracy by 2.49% on medical and 1.2% on financial evaluations over the second-best baseline. Ablation and scaling studies further confirm the framework's robustness.

2 Related Works

Catastrophic Forgetting Catastrophic Forgetting refers to the phenomenon where neural networks forget previously learned knowledge when trained on new tasks, posing a major challenge in continual learning (French, 1999; De Lange et al., 2021). Existing approaches can be roughly divided into four categories. Regularization-based methods constrain parameter updates to preserve important past knowledge (Kirkpatrick et al., 2017; Zenke et al., 2017; Li and Hoiem, 2017). Rehearsal-based methods replay stored or generated samples to maintain previous performance (Chaudhry et al., 2019; Buzzega et al., 2020). Architectural and gradient-based methods isolate parameters or project gradients to reduce task interference (Rusu et al., 2016; Serra et al., 2018; Farajtabar et al., 2020). Recently, prompt- and adapter-based methods leverage pretrained models with lightweight modules to mitigate forgetting efficiently (Wang et al., 2022). Despite these various efforts, achieving stable and scalable knowledge learning across multiple domains remains a challenge.

Curriculum Learning Curriculum learning refers to a training paradigm inspired by the human learning process, where a model is trained on easier samples or subtasks before gradually moving to harder ones (Zhang et al., 2019). It typically aims to improve optimization and generalization by controlling the order of training data, and has been studied through predefined curricula based on heuristic difficulty measures such as length, noise, or complexity (Wei et al., 2016), self-paced schemes that adaptively select or weight samples according to model competence (Kumar et al., 2010), and teacher-guided or transfer-based approaches that estimate difficulty via auxiliary models (Weinshall et al., 2018). For large language models, CL-inspired techniques include instruction data selection (Cao et al., 2023), progressive learning from increasingly complex explanation traces (Mukherjee et al., 2023), and token- or signal-level reweighting to emphasize informative tokens (Lin et al., 2024). However, these methods typically organize learning along a single axis (e.g., instruction quality, explanation complexity, or token importance), whereas we construct an explicit two-dimensional

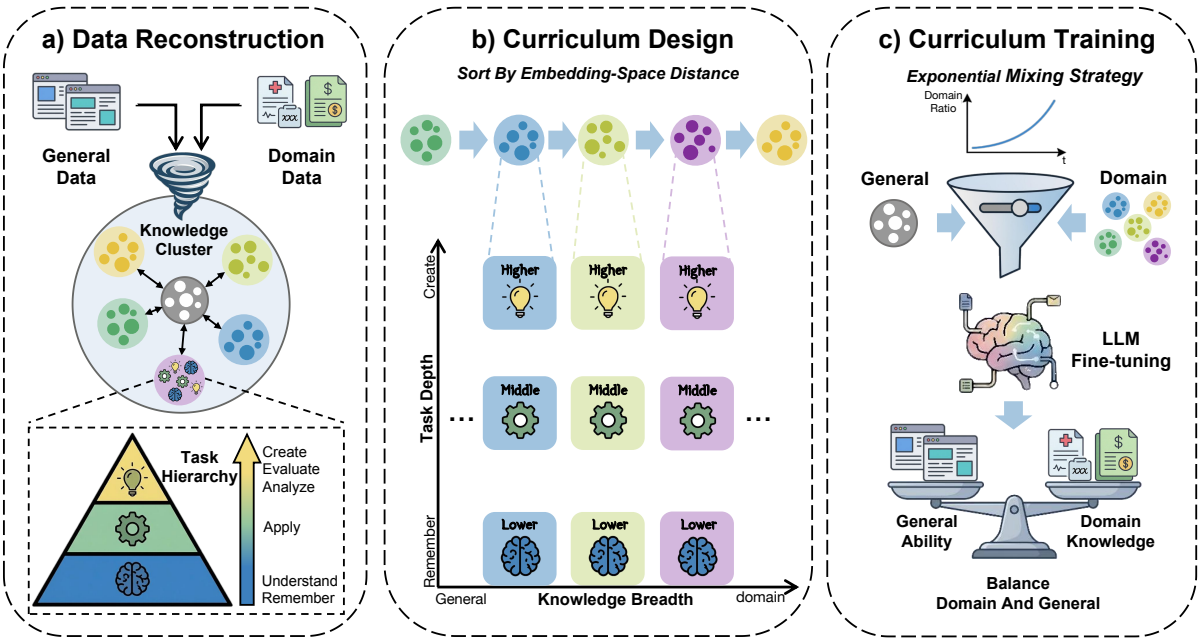

Figure 1: Overview of the proposed two-dimensional curriculum learning framework. The pipeline comprises (a) Data Reconstruction with knowledge clustering and task annotation, (b) Curriculum Design along knowledge dimension and task dimension, and (c) Curriculum Training with an exponential mixing schedule.

curriculum over knowledge clusters and task hierarchies to jointly control knowledge breadth and task complexity for domain adaptation.

3 Problem Statement

When training large language models on both general and domain data, a key challenge lies in mitigating catastrophic forgetting of general abilities while enhancing domain knowledge. Formally, let the dataset be $\mathcal{D} = \mathcal{D}_g \cup \mathcal{D}_s$, where $\mathcal{D}_g$ and $\mathcal{D}_s$ denote the general and domain datasets, respectively. Each instance $x \in \mathcal{D}$ corresponds to an input–output pair reflecting either general abilities or domain knowledge.

Curriculum learning typically addresses this issue by assigning each sample x a difficulty score $d(x)$, and then using a scheduling function π to order the dataset: $\pi(\mathcal{D}) = (x_1, x_2, \ldots, x_{|\mathcal{D}|})$, such that easier examples precede harder ones. This gradual progression allows the model to acquire domain knowledge while retaining general abilities.

4 Methods

We propose a two-dimensional curriculum learning framework that jointly incorporates knowledge-level learning and task-level learning. Together, these stages transform heterogeneous corpora into a structured learning trajectory that balances general-purpose knowledge retention with domain-specific specialization. As illustrated in Figure 1, the framework consists of three stages: (1) Data Reconstruction, (2) Curriculum Design, and (3) Curriculum Training.

4.1 Data Reconstruction

To exploit intrinsic data relationships, we restructure the training corpus along two dimensions: semantic similarity and task complexity. This stage transforms the data by mapping heterogeneous instances into a unified space for knowledge clustering and annotating them with hierarchical task labels.

Vector Embedding. To facilitate a unified semantic representation, we embed both the general corpus $\mathcal{D}_g$ and the domain corpus $\mathcal{D}_s$ into a shared latent vector space using a pretrained model. This unified representation bridges the distributional gap between heterogeneous datasets, ensuring that semantically similar instances are positioned in close proximity regardless of their data source.

Knowledge Clustering. Building upon the previous vector space embedding, vectors are grouped into clusters via K-means (Lloyd, 1982) optimization:

$$\min_{\{C_k\}_{k=1}^{K}} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|\mathbf{v}_i - \boldsymbol{\mu}_k\|^2 \tag{1}$$

where $x_i \in \mathcal{D}$, $\mathbf{v}_i$ is its embedding, and $\boldsymbol{\mu}_k$ denotes the centroid of cluster C_k. To measure inter-cluster relationships, cosine distance is employed:

$$\text{dist}(C_a, C_b) = 1 - \frac{\boldsymbol{\mu}_a \cdot \boldsymbol{\mu}_b}{\|\boldsymbol{\mu}_a\| \|\boldsymbol{\mu}_b\|} \tag{2}$$

Clusters closer to the general centroid $\boldsymbol{\mu}_g$ are semantically aligned with $\mathcal{D}_g$, while more distant clusters encode increasingly domain-specific content from $\mathcal{D}_s$. This spatial organization provides the basis for the knowledge dimension in curriculum design.To verify that this distance metric genuinely reflects domain specificity rather than random distributional shifts, we conducted a quantitative correlation analysis using Inverse Document Frequency (IDF) to measure term rarity. Our results demonstrate a highly significant positive correlation between a cluster's distance from μ_g and its term rarity (detailed in Appendix A), supporting the use of this metric as a proxy for knowledge depth.

Task Hierarchy. To capture cognitive variation, we adopt Bloom's taxonomy (Bloom et al., 1956), a well-established framework for categorizing educational objectives into six hierarchical levels: remembering, understanding, applying, analyzing, evaluating, and creating. This hierarchy mirrors the model's learning trajectory, progressing from basic pattern recall to complex reasoning and synthesis. This hierarchical structuring is supported by recent research. Specifically, curriculum ordering based on cognitive levels has been shown to significantly outperform random shuffling in instruction tuning (Lee et al., 2024). Furthermore, the validity of Bloom's Taxonomy as a robust framework for characterizing the cognitive depth of LLMs has been confirmed by recent evaluations (Huber and Niklaus, 2025), which provides a theoretical basis for our approach. Prior work has explored domain-oriented tools for structured text annotation in specialized corpora (Lin et al., 2022). Different from such tool-oriented approaches, we assign task labels automatically with Qwen3-8B based on Bloom's taxonomy. However, the boundaries between adjacent levels in the original taxonomy (e.g., Analyzing vs. Evaluating) can be subtle and prone to ambiguity, leading to lower classification consistency. To mitigate annotation noise and enhance robustness, we consolidate these six levels into three strata: a lower level (L), corresponding to remembering and understanding; a middle level (M), reflecting the application of knowledge; and a higher level (H), encompassing analysis, evaluation, and creation. Human validation (detailed in Appendix B) confirms that this coarse-grained mapping significantly improves alignment with human judgment compared to the direct 6-way classification, ensuring a more reliable curriculum signal.

Formally, each instance is encoded as a triplet

$$x_i = \{\mathbf{v}_i, c_i, t_i\}, \quad x_i \in \mathcal{D} \tag{3}$$

where $\mathbf{v}_i$ denotes the feature representation, c_i is the cluster identifier, and $t_i \in \{L, M, H\}$ specifies the task complexity level.

4.2 Curriculum Design

Building upon the structured data, the second stage specifies the sequencing of training instances along two dimensions: knowledge dimension and task dimension. The overarching goal of this stage is to transform the raw partitions into a pedagogically informed learning trajectory. By organizing clusters and tasks in a systematic order, the curriculum mitigates cognitive overload, preserves general abilities, and facilitates the incremental acquisition of domain knowledge.

Knowledge Curriculum. The knowledge curriculum is constructed by ordering clusters according to their semantic similarity to the general centroid $\boldsymbol{\mu}_g$:

$$D_i = \text{dist}(c_i, \boldsymbol{\mu}_g) = 1 - \frac{\boldsymbol{\mu}_{c_i} \cdot \boldsymbol{\mu}_g}{\|\boldsymbol{\mu}_{c_i}\| \|\boldsymbol{\mu}_g\|} \tag{4}$$

Training begins with clusters most closely aligned with $\mathcal{D}_g$ and gradually transitions toward those representing highly domain-specific content from $\mathcal{D}_s$.

Task Curriculum. Within each knowledge cluster, training instances are further sequenced by task complexity. This ensures that the model initially acquires competence in low-order tasks ($t = L$) before progressing to higher-order reasoning tasks ($t = H$). Such intra-cluster progression not only aligns with Bloom's taxonomy but also reduces the likelihood of cognitive overload when complex tasks are introduced prematurely.

Two-Dimensional Curriculum. The final curriculum integrates the two dimensions of knowledge and task into a unified sequencing principle.

Datasets	Domain	#Ori	#Sel
Infinity Instruct	General	1.5M	10K
MedInstruct	Medical	52K	5K
MedThoughts	Medical	8K	5K
Finance Alpaca	Finance	52K	10K

Table 1: Statistics of the datasets used for supervised fine-tuning. #Ori denotes the number of original samples, while #Sel denotes the number of selected samples.

Formally, the curriculum's ordering is determined by a lexicographic rule. An instance x_i precedes x_j (denoted $x_i \prec x_j$) if and only if

$$x_i \prec x_j \quad \Leftrightarrow \quad \begin{cases} D_i < D_j, \\ D_i = D_j \text{ and } t_i \prec t_j \end{cases} \tag{5}$$

At a macro level, this curriculum enforces a progression across clusters that expands the *breadth* of knowledge: the model first consolidates general abilities from $\mathcal{D}_g$ (low D_i) before gradually traversing toward increasingly domain knowledge from $\mathcal{D}_s$ (high D_i). At a micro level, within each cluster, the curriculum deepens the *depth* of learning by guiding the model through tasks of escalating complexity ($L \to M \to H$).

4.3 Curriculum Training

In the final stage, we adopt an exponential mixing strategy, where samples from $\mathcal{D}_s$ are gradually interleaved with those from $\mathcal{D}_g$ according to a dynamic schedule.

Let the size of general datasets be $|\mathcal{D}_g|$ and domain datasets be $|\mathcal{D}_s|$. At iteration index t (relative to $|\mathcal{D}_g|$), we define the progression variable

$$p_t = \frac{t}{|\mathcal{D}_g|}, \quad p_t \in (0, 1] \tag{6}$$

which indicates the relative position in the training process.

Formally, the expected number of domain-specific samples at iteration t is

$$\mathbb{E}[n_t] = r_{\max} \, p_t^k, \tag{7}$$

where $k > 0$ controls the growth steepness, and $r_{\max} = \frac{2|\mathcal{D}_s|}{|\mathcal{D}_g|}$ is the maximum insertion ratio determined by dataset sizes.

The exponential rule thus guides the model from stable general training ($\mathcal{D}_g$) to focused domain specialization ($\mathcal{D}_s$).

5 Experiments and Results

In this section, we present the details of our experimental setup (Section 5.1), the baselines used for comparison (Section 5.2), and the experimental results (Section 5.3).

5.1 Experimental Setup

Datasets. For supervised fine-tuning, we use data from three domains. In the **general domain**, we adopt Infinity Instruct (Li et al., 2025), a large-scale instruction-following corpus. For the **medical domain**, we incorporated two resources: MedInstruct (Zhang et al., 2023), a synthetic dataset containing approximately 52k medical instruction–response pairs generated by GPT-4 (Achiam et al., 2023), and MedThoughts, an extension of MedQA (Jin et al., 2021) that augments answers with reasoning traces distilled from DeepSeek-R1 (Guo et al., 2025). In the **financial domain**, we utilized Finance Alpaca, an Alpaca-style dataset built from Alpaca, FiQA, and GPT-generated financial examples. To disentangle domain-specific learning from general capabilities, we remove all medical- and financial-related data from the general corpus, and the specific removal methodology is detailed in Appendix C. Detailed training dataset statistics and preprocessing are provided in Appendix D.

For evaluation, we use benchmarks from three domains. In the **medical domain**, we adopt MedQA, MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and the medical subsets of MMLU-Pro (Wang et al., 2024) and GPQA (Rein et al., 2024). For the **financial domain**, we use FinQA (Chen et al., 2021), FPB (Malo et al., 2014), and Headlines (Sinha and Khandait, 2021). For **general domain** performance, we evaluate on MMLU (Hendrycks et al., 2021), MMLU-Pro, and ARC (Clark et al., 2018). To avoid contamination, all medical and financial subsets are excluded when reporting general-purpose results.

Models and Implementation. To demonstrate the generalization ability of our method across different model families, we adopt two backbones: Qwen3 (Yang et al., 2025), including Qwen3-8B, Qwen3-4B, and Qwen3-1.7B, and LLaMA3 (Grattafiori et al., 2024), where we employ LLaMA3.1-8B. For the data reconstruction process, we utilize `bge-m3` as the embedding backbone to capture dense semantic representations for

Method	Medical						General			
	MedQA	MMC.QA	PM.QA	MMLU-P.	GPQA	Avg.	MMLU	MMLU-P.	ARC	Avg.
Qwen3-8B										
Base	0.5734	0.5316	0.6680	0.5452	0.4384	0.5513	0.6313	0.4325	0.7500	0.6046
Shuffle	0.6095	0.5419	0.7460	0.6566	0.4923	0.6093	0.6932	0.5501	0.8191	0.6875
PDPC	0.5981	**0.6101**	0.7480	0.6697	0.4948	0.6241	**0.7449**	0.5836	0.8579	0.7288
DMT	0.5545	0.5522	0.6140	0.5706	0.4564	0.5495	0.6415	0.4325	0.7969	0.6236
Ours	**0.6692**	0.6014	**0.7510**	**0.6951**	**0.5282**	**0.6490**	0.7291	**0.5878**	**0.8703**	**0.7291**
LLaMA3.1-8B										
Base	0.4700	0.3450	0.5220	0.2751	0.2976	0.3819	0.3646	0.1725	0.4516	0.3296
Shuffle	0.6205	**0.4929**	0.6980	0.5166	0.3471	0.5350	0.6163	0.3658	0.7457	0.5759
PDPC	0.6201	0.4873	0.7120	0.5115	0.3413	0.5344	0.6168	**0.3727**	**0.7783**	**0.5893**
DMT	0.6102	0.4852	0.6930	0.5100	0.3348	0.5266	0.6131	0.3651	0.7704	0.5829
Ours	**0.6311**	0.4917	**0.7190**	**0.5224**	**0.3792**	**0.5487**	**0.6202**	0.3686	0.7704	0.5864

Table 2: The effectiveness of the proposed two-dimensional curriculum is assessed on both medical and general-domain benchmarks. We compare its performance with that of Base, Shuffle, PDPC, and DMT. Specifically, MMC.QA, PM.QA, and MMLU.P correspond to MedMCQA, PubMedQA, and MMLU-Pro, respectively. The best-performing results are indicated in **bold**, and the second-best results are underlined.

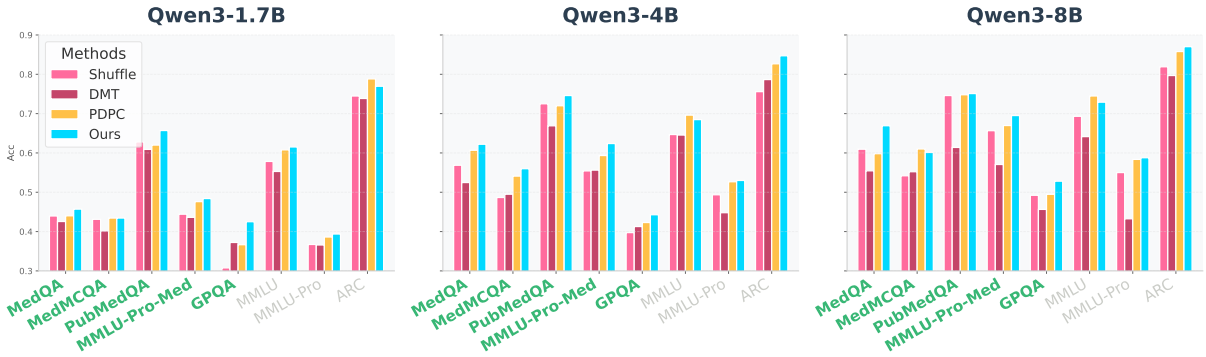

Figure 2: Scaling behavior of the two-dimensional curriculum on the Qwen3 family (1.7B/4B/8B). Across model scales and benchmarks, the method consistently outperforms most baselines, indicating robustness to parameter size.

knowledge clustering. Fine-tuning is conducted via supervised fine-tuning with LoRA, using a rank of 8. All models are trained using AdamW with a learning rate of 5×10^{-5} for 3 epochs and a batch size of 64. Comprehensive implementation details and a thorough analysis of computational resource consumption are provided in Appendix F.

5.2 Baselines

We compare our method with several baseline strategies: (1) **Shuffle** mixes general and domain data uniformly without structured ordering. (2) **PDPC** (Zhang et al., 2025) partitions data by perplexity difference (PD) to progressively train from low- to high-PD samples. (3) **DMT** (Dong et al., 2024) adopts a two-stage strategy: initial training on domain data, followed by fine-tuning on a mixture of general and domain data (with ratio K=1/256). (4) **Reverse Curriculum** inverts our proposed framework by transitioning from domain-specific to general clusters and arranging task hierarchy from higher- to lower-order.

5.3 Results

Overall Performance. Table 2 and 4 summarize the experimental results of our method and several baselines, including *Base*, *Shuffle*, *PDPC*, and *DMT*, evaluated on two backbone models—Qwen3-8B and LLaMA3.1-8B—across multiple benchmarks spanning the medical, financial, and general

Method	Medical						General			
	MedQA	MMC.QA	PM.QA	MMLU-P.	GPQA	Avg.	MMLU	MMLU-P.	ARC	Avg.
Qwen3-8B										
Base	0.5734	0.5316	0.6680	0.5452	0.4384	0.5513	0.6313	0.4325	0.7500	0.6046
Ours-Reverse	0.6097	0.5806	0.7230	0.6097	0.4871	0.6020	0.6717	0.5270	0.7645	0.6544
w/o Knowledge	0.5679	0.5689	0.7440	0.6456	0.4871	0.6027	0.7081	0.5551	0.8694	0.7109
w/o Task	0.6504	0.5656	**0.7540**	0.6247	0.5153	0.6220	0.6935	0.5442	0.8199	0.6859
Ours	**0.6692**	**0.6014**	0.7510	**0.6951**	**0.5282**	**0.6490**	**0.7291**	**0.5878**	**0.8703**	**0.7291**
LLaMA3.1-8B										
Base	0.4700	0.3450	0.5220	0.2751	0.2976	0.3819	0.3646	0.1725	0.4516	0.3296
Ours-Reverse	0.6009	0.4754	0.6790	0.5016	0.3230	0.5160	0.6010	0.3531	0.7525	0.5689
w/o Knowledge	0.6177	0.4841	0.7040	0.5206	0.3811	0.5415	0.6151	0.3652	0.7320	0.5708
w/o Task	0.6025	0.4819	0.7130	0.5179	**0.3820**	0.5395	0.6107	0.3649	**0.7721**	0.5826
Ours	**0.6311**	**0.4917**	**0.7190**	**0.5224**	0.3792	**0.5487**	**0.6202**	**0.3686**	0.7704	**0.5864**

Table 3: Ablation studies on our curriculum. The experiments include Ours-Reverse (reversing the curriculum order), w/o Knowledge (removing the knowledge-level curriculum), and w/o Task (removing the task-level curriculum).

Method	Finance				General			
	FinQA	FPB	Headlines	Avg.	MMLU	MMLU-P.	ARC	Avg.
Base	0.0174	0.2969	0.2120	0.1754	0.3646	0.1725	0.4516	0.3296
Shuffle	0.0575	0.6856	0.6353	0.4595	0.5669	0.3193	0.7030	0.5297
PDPC	0.0480	0.7237	0.6590	0.4769	**0.5755**	0.3233	0.7047	0.5345
DMT	0.0593	0.4753	0.5530	0.3625	0.5052	0.2543	0.6126	0.4574
Ours-Reverse	0.0366	0.6258	0.6410	0.4345	0.5639	0.3185	0.7081	0.5302
w/o Knowledge	0.0619	0.7000	0.6160	0.4593	0.5441	0.3135	0.6928	0.5168
w/o Task	0.0584	0.6959	0.6010	0.4518	0.5648	0.3212	**0.7286**	0.5382
Ours	**0.0645**	**0.7311**	**0.6710**	**0.4889**	0.5609	**0.3389**	0.7209	**0.5402**

Table 4: Cross-domain evaluation on finance benchmark. The curriculum transfers effectively to finance, delivering the best overall balance between domain and general performance.

domains.

Across both models, our method achieves the best overall average performance (SOTA). Specifically, on Qwen3-8B, our method attains an average accuracy of 0.6490 on medical tasks and 0.7291 on general tasks, outperforming the second-best method (PDPC) by +2.49% and +0.03%, respectively. On LLaMA3.1-8B, our method reaches an average of 0.5487 in the medical domain, exceeding the next-best result by +1.37%, while achieving comparable performance in the general domain.

At the task level, our method consistently ranks first or second across nearly all datasets and model backbones, demonstrating strong cross-domain robustness. Notably, while other approaches exhibit larger performance fluctuations across backbones—for instance, PDPC performs as the second-best method on Qwen3-8B but is surpassed by Shuffle on LLaMA3.1-8B—our method maintains stable superiority on both, indicating greater consistency and generalization capability.

Overall, these results confirm the effectiveness and robustness of our proposed two-dimensional curriculum strategy across diverse domains and model architectures.

Scaling Analysis. To further investigate the scalability of our method, we conduct experiments on Qwen models of different parameter sizes, including Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, with results illustrated in Figure 2. Across nearly all benchmarks, our method consistently outperforms baseline approaches under different parameter regimes, demonstrating strong robustness to model scale. In particular, while PDPC occasion-

ally surpasses our method on a small subset of general benchmarks, it lags significantly behind on domain-specific tasks. By contrast, our two-dimensional curriculum simultaneously maintains superior performance on both general and specialized benchmarks, striking a balance that neither naive shuffling nor preference-based curricula achieve. This indicates that the advantages of our method are not limited to a specific scale, but extend across lightweight and larger backbones, making it a scalable solution for practical deployment.

In addition to model scaling, we also evaluated the framework's robustness when scaling the training data. Expanding the SFT dataset to 60k instances yields consistent performance improvements and sustained mitigation of catastrophic forgetting, with full experimental details provided in Appendix I.

Domain Generalization. We further evaluate on the financial domain to test generality beyond medical tasks. The results are presented in Table 4. The results demonstrate that our method achieved the highest performance across both the financial and general-domain test sets. Specifically, it attained a score of 0.4889 in the financial domain, surpassing the second-best method, PDPC, by 1.2%, and reached an average accuracy of 0.5402 on the general-domain test set, outperforming the next-best ablated variant by 0.2%.

Interestingly, *w/o Task* performs relatively well on ARC but lags behind on domain-specific tasks, which also suggests that task-level learning is particularly critical for specialized reasoning. Overall, the consistent improvements across both medical and financial settings demonstrate that our curriculum design is not confined to a single domain, but serves as a general strategy for domain adaptation with large language models.

Hyperparameter Analysis. To validate the robustness of our framework configurations, we conducted sensitivity analyses on the Qwen3-8B model, focusing on the number of knowledge clusters (K) and the data mixing strategy. As illustrated in Figure 3(a), the algorithm demonstrates strong adaptability regarding the choice of K, indicating that performance is relatively insensitive to this hyperparameter provided extreme settings are avoided. While an inverted-U trend is observable—where a very small K is too coarse to capture distinct knowledge boundaries and an excessively large K overly fragments the semantic space—the performance variance remains minimal across a broad effective range (e.g., from 50 to 200). We finally utilized $K = 100$ for our main experiments as it yields the optimal balance. Furthermore, we verified the effectiveness of our exponential mixing strategy by comparing it with Uniform (fixed 50% ratio) and Linear (k=1) schedules. As shown in Figure 3(b), the proposed exponential strategy consistently outperforms the baselines. Specifically, it surpasses the Uniform strategy by +2.88% on medical tasks and +1.03% on general tasks. This confirms that a non-linear, gradual increase in domain data density is crucial for stabilizing training dynamics and mitigating catastrophic forgetting during the domain adaptation process.

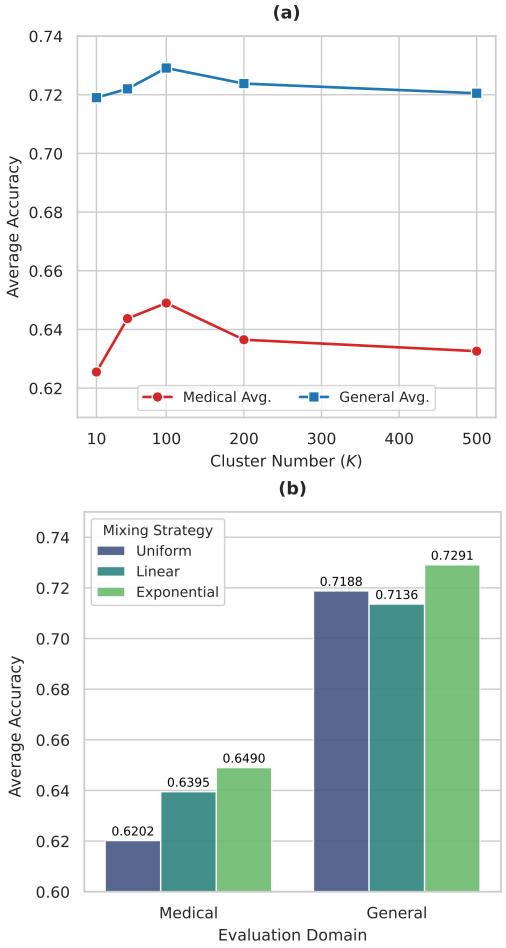

Figure 3: Ablation studies on hyperparameter sensitivity and mixing strategies using Qwen3-8B. (a) The plot shows how performance changes with different numbers of knowledge clusters (K). (b) The plot compares model performance under three data mixing strategies: Uniform, Linear, and Exponential.

Ablation Study. To disentangle the contributions of our two curriculum dimensions, we conduct a series of ablation studies, with results presented in Table 3. Our analysis reveals that each

component plays a distinct and complementary role.

Removing the knowledge curriculum (*w/o Knowledge*), which provides a structured progression from general to domain-specific content, causes a more significant performance drop on general-purpose benchmarks. This is particularly evident on MMLU (from 0.7291 to 0.7081 on Qwen3-8B) and ARC (from 0.7704 to 0.7320 on LLaMA3.1-8B), underscoring its importance in preserving foundational knowledge and preventing catastrophic forgetting. Conversely, eliminating the task curriculum (*w/o Task*), which organizes learning by task complexity, leads to a more pronounced performance decline on specialized domain benchmarks. For example, performance degrades on MedQA (from 0.6311 to 0.6025 on LLaMA3.1-8B) and MedMCQA (from 0.6014 to 0.5656 on Qwen3-8B), which indicates that the task hierarchy is crucial for developing deep, nuanced understanding within a specific field.

Furthermore, reversing the knowledge trajectory (*Ours-Reverse*) consistently underperforms our proposed method, confirming that beginning with highly specialized data harms generalization. These findings validate that our two-dimensional curriculum is indispensable, with the knowledge component fostering broad generalizability and the task component cultivating deep domain expertise.

6 Conclusion

We propose a two-dimensional curriculum learning framework for supervised fine-tuning of large language models. Our method organizes training data along knowledge similarity and task complexity, progressing from general to domain-specific clusters and from lower- to higher-order tasks, which helps preserve general reasoning while acquiring domain expertise and mitigating catastrophic forgetting. Experiments on medical, financial, and general benchmarks show consistent improvements over strong baselines, and ablation, scaling, and transfer studies confirm the complementary roles of the two curriculum dimensions and the robustness of the framework across model sizes and domains.

Limitations

Our study has two main limitations: 1) Due to limited computational resources, we only fine-tune models up to 8B parameters, which may limit the assessment of scalability to larger LLMs. 2) Our evaluation focuses on medical and financial domains; broader validation on additional specialized areas (e.g., law or engineering) is needed to further test the generality of the proposed framework.

Acknowledgments

This work is supported by the Shanghai Natural Science Foundation Project under Grant 25ZR1402116.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. 2020. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. 2023. Instruction mining: Instruction data selection for tuning large language models. *arXiv preprint arXiv:2307.06290*.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. 2019. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024b. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. 2025. Self-evolving curriculum for llm reasoning. *arXiv preprint arXiv:2505.14970*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and

William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. How abilities in large language models are affected by supervised fine-tuning data composition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–198, Bangkok, Thailand. Association for Computational Linguistics.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. 2020. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pages 3762–3773. PMLR.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ruihui Hou, Shencheng Chen, Yongqi Fan, Guangya Yu, Lifeng Zhu, Jing Sun, Jingping Liu, and Tong Ruan. 2025a. Msdiagnosis: A benchmark and framework for evaluating large language models in multi-step clinical diagnosis. *Knowledge-Based Systems*, 330:114524.

Zhitian Hou, Zihan Ye, Nanli Zeng, Tianyong Hao, and Kun Zeng. 2025b. Large language models meet legal artificial intelligence: A survey. *arXiv preprint arXiv:2509.09969*.

Thomas Huber and Christina Niklaus. 2025. LLMs meet bloom's taxonomy: A cognitive view on large language model evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246, Abu Dhabi, UAE. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

M Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bruce W Lee, Hyunsoo Cho, and Kang Min Yoo. 2024. Instruction tuning with human curriculum. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1281–1309, Mexico City, Mexico. Association for Computational Linguistics.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024. Revisiting catastrophic forgetting in large language model tuning. *arXiv preprint arXiv:2406.04836*.

Jijie Li, Li Du, Hanyu Zhao, Bo-wen Zhang, Liangdong Wang, Boyan Gao, Guang Liu, and Yonghua Lin. 2025. Infinity Instruct: Scaling Instruction Selection and Synthesis to Enhance Language Models. *arXiv preprint*. ArXiv:2506.11116 [cs].

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

Yupian Lin, Tong Ruan, Ming Liang, Tingting Cai, Wen Du, and Yi Wang. 2022. Dotat: A domain-oriented text annotation tool. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–8.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and 1 others. 2024. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pages 4548–4557. PMLR.

Ankur Sinha and Tanmay Khandait. 2021. Impact of news on the commodity market: Dataset and results. In *Future of Information and Communication Conference*, pages 589–601. Springer.

Shezheng Song, Hao Xu, Jun Ma, Shasha Li, Long Peng, Qian Wan, Xiaodong Liu, and Jie Yu. 2025. How to alleviate catastrophic forgetting in llms finetuning? hierarchical layer-wise and element-wise regularization. *arXiv preprint arXiv:2501.13669*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149.

Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. 2016. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2314–2320.

Daphna Weinshall, Gad Cohen, and Dan Amir. 2018. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International conference on machine learning*, pages 5238–5246. PMLR.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yuanhao Yue, Chengyu Wang, Jun Huang, and Peng Wang. 2024. Distilling instruction-following abilities of large language models with task-aware curriculum planning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6030–6054, Miami, Florida, USA. Association for Computational Linguistics.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. *arXiv preprint arXiv:1905.05816*.

Xuemiao Zhang, Xu Liangyu, Feiyu Duan, Yongwei Zhou, Sirui Wang, Rongxiang Weng, Jingang Wang, and Xunliang Cai. 2025. Preference curriculum: LLMs should always be pretrained on their preferred data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21181–21198, Vienna, Austria. Association for Computational Linguistics.

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

A Verification of Distance Signal

To assess whether the embedding-distance signal serves as a useful proxy for domain specificity, rather than merely reflecting random distributional shift, we conducted a quantitative analysis centered on terminology rarity.

Specifically, we used inverse document frequency (IDF) as a measure of term rarity. The IDF model was trained on our general instruction corpus, allowing us to quantify, from a general-domain perspective, how rare the vocabulary in each domain-specific cluster is.

For embedding computation and clustering, we followed the methodology described in the main paper. We used the `bge-m3` embedding model and set $K = 100$ for clustering. We then computed the distance between each cluster centroid and the general centroid, and analyzed its correlation with the average term rarity within that cluster.

Experimental Results

- Pearson correlation coefficient: $r = 0.38$ ($p < 0.001$)
- Spearman correlation coefficient: $\rho = 0.38$ ($p < 0.001$)

The extremely low p-values indicate that the relationship is highly statistically significant. The moderate positive correlation suggests that clusters located farther from the general centroid tend to contain more specialized and rarer terminology.

These results support our core hypothesis: embedding distance can serve as an effective proxy for measuring domain specificity and knowledge difficulty relative to the general domain.

B Task Hierarchy Details

To assess the reliability of our automated task complexity annotation, we conducted a human validation study. We randomly sampled 120 instances from the annotated corpus and invited human experts to classify them according to Bloom's taxonomy definitions. We compared the agreement between the Qwen3-8B predictions and human annotations under two settings: the original 6-level taxonomy and our proposed 3-level strata (L, M, H). The results are summarized below:

6-Level Classification: The direct classification into six levels resulted in an accuracy of 70% with a Cohen's kappa of 0.64. Discrepancies primarily occurred between adjacent cognitive levels (e.g., Understanding vs. Applying), reflecting the inherent ambiguity of fine-grained distinctions in automated tagging.

3-Level Strata (Ours): By grouping the labels into the L, M, and H strata, the accuracy significantly improved to 90.83%, and the Cohen's kappa increased to 0.8510, indicating substantial agreement.

These findings validate our design choice: while Qwen3-8B provides strong semantic understanding, the consolidation into three levels filters out fine-grained noise, providing a more stable and accurate progression for the curriculum learning framework.

Taxonomy Setting	Accuracy	Cohen's kappa
6-Level Classification	0.7000	0.6400
3-Level Strata (Ours)	**0.9083**	**0.8510**

Table 5: Results of the human validation study on sampled instances.

C Data Filtering

Label-based Filtering

Since the general corpus we used, Infinity Instruct, comes with built-in metadata labels, we first leveraged its original labeling system to directly remove all samples tagged under the "Medical" or "Financial" domains. This step filtered out approximately 93% of the domain-related data.

Keyword-based Filtering

To further clean potential domain-specific samples not covered by the metadata labels, we performed a second-stage filtering using a high-frequency domain-specific keyword list for matching and removal. This step filtered out the remaining approximately 7% of domain-related data.

The specific keywords used are as follows:

Medical domain keywords: diagnosis, treatment, clinical, symptom, patient, hospital, disease, surgery, drug.

Financial domain keywords: stock, revenue, investment, market, bank, profit, asset, currency, shareholder.

D Training Datasets

To enable both general instruction following and domain-specific reasoning, we curated supervised fine-tuning (SFT) data from three domains: general, medical, and financial. This section details the datasets and sampling strategy used in SFT.

General Domain

We adopt the **Infinity Instruct** (Li et al., 2025) dataset, a large-scale, instruction-following corpus containing millions of high-quality, general-purpose and dialogue-style examples. Constructed via a two-stage pipeline of instruction selection and evolution, the dataset emphasizes comprehension, reasoning, and multi-turn conversation.

During preprocessing, all medical- and financial-related samples were filtered to ensure domain separation. The resulting corpus spans diverse topics, including commonsense reasoning, mathematics, programming, and dialogue generation.

To maintain training efficiency and domain balance, we randomly selected 10,000 general-domain samples for SFT.

Medical Domain

To enhance the model's ability in medical reasoning and communication, we combine two complementary datasets:

- **MedInstruct** (Zhang et al., 2023) contains approximately 52,000 instruction–response pairs generated by GPT-4 (Achiam et al., 2023) from expert-designed seed instructions. Each sample includes an `instruction`, optional `input`, and `response`, covering clinical and consultation scenarios. Quality was controlled through a curated evaluation set.
- **MedThoughts**, an extension of MedQA (Jin et al., 2021), augments each question with step-by-step reasoning paths distilled from DeepSeek-R1 (Guo et al., 2025). Each entry provides both a final answer and a verified reasoning trace, ensuring logical and factual soundness.

Together, these datasets support both factual coverage and interpretable reasoning. We randomly sampled 10,000 medical-domain examples for SFT.

Financial Domain

For financial-domain adaptation, we use the **Finance Alpaca** dataset, an Alpaca-style instruction–response corpus derived from Alpaca, FiQA, and GPT-generated financial examples. It spans topics such as investment, banking, market analysis, and corporate finance.

This dataset supports domain-specific reasoning while retaining instruction-following capabilities. We randomly selected 10,000 financial-domain samples for training.

E Evaluation Datasets

We evaluate our models on a diverse set of benchmarks across medical, financial, and general-purpose domains. The following datasets are used for evaluation.

Medical Domain

- **MedQA** (Jin et al., 2021): A USMLE-style multiple-choice dataset evaluating medical knowledge, clinical reasoning, and factual accuracy.
- **MedMCQA** (Pal et al., 2022): A large-scale MCQ benchmark with over 190,000 questions from medical entrance exams, focusing on factual recall and domain reasoning.
- **PubMedQA** (Jin et al., 2019): A biomedical QA dataset with research questions, PubMed abstracts, and categorical answers (*yes/no/maybe*), testing scientific comprehension and evidence-based reasoning.
- **MMLU-Pro (medical subsets)** (Wang et al., 2024): A challenging subset of MMLU-Pro targeting advanced biomedical knowledge and multi-step reasoning.
- **GPQA (medical subsets)** (Rein et al., 2024): Contains graduate-level medical questions with multi-hop reasoning requirements.

Financial Domain

- **FinQA** (Chen et al., 2021): A numerical reasoning dataset based on financial reports, assessing integration of textual and quantitative understanding.
- **Financial PhraseBank (FPB)** (Malo et al., 2014): A sentiment classification dataset of financial news sentences labeled by polarity.
- **Headlines** (Sinha and Khandait, 2021): A benchmark of short market-related headlines annotated with sentiment or price movement, measuring semantic understanding in finance.

General-Purpose Benchmarks

- **MMLU** (Hendrycks et al., 2021): A broad benchmark spanning 57 subjects, including STEM, humanities, and social sciences.
- **MMLU-Pro** (Wang et al., 2024): An enhanced version of MMLU with balanced distractors and higher difficulty; medical and financial subsets are excluded in general evaluations.
- **ARC** (Clark et al., 2018): Focuses on commonsense and scientific reasoning using grade-school science questions.

Evaluation Protocol

To ensure a fair comparison, we exclude medical and financial subsets from general-purpose evaluations to prevent domain leakage, and we carefully eliminate overlap between the evaluation datasets and the SFT training corpora. This evaluation suite provides a comprehensive assessment of factual knowledge, reasoning ability, and domain-specific expertise.

F Implementation Details

We fine-tune all SFT datasets for 3 epochs with a batch size of 64 on NVIDIA A800 GPUs. The fine-tuning process is conducted using the LLaMA-Factory (Zheng et al., 2024) framework, with a learning rate of 5e-5. We adopt the LoRA (Low-Rank Adaptation) technique to efficiently adapt large language models, setting the LoRA rank to 8 and the LoRA alpha to 16. After fine-tuning, we perform batch evaluations using the vLLM (Kwon et al., 2023) engine to ensure efficient and consistent inference performance.

In the task annotation stage, we employed the Qwen3-8B model to generate task labels for the domain-specific corpus. The annotation prompts were adapted and refined based on those proposed in (Lee et al., 2024), as shown below:

Prompt

Your task is to classify tasks into Bloom's taxonomy. The classes and their description are provided below:
Remember: Recall facts and basic concepts.
[Examples]: define, duplicate, list, memorize, repeat, state
Understand: Explain ideas or concepts.
[Examples]: classify, describe, discuss, explain, identify, locate, recognize, report, select, translate
Apply: Use information in new situations.
[Examples]: execute, implement, solve, use, demonstrate, interpret, operate, schedule, sketch
Analyze: Draw connections among ideas.
[Examples]: differentiate, organize, relate, compare, contrast, distinguish, examine, experiment, question, test
Evaluate: Justify a stand or decision.
[Examples]: appraise, argue, defend, judge, select, support, value, critique, weigh
Create: Produce new or original work.
[Examples]: design, assemble, construct, conjecture, develop, formulate, author, investigate
Now, classify the following problem into ONE category only. Do not explain. Do not execute the problem. Only output the category.
Problem (do not execute, only classify):
problem
Output format (exactly one word):
Remember
Understand
Apply
Analyze
Evaluate
Create

We also conducted a detailed analysis of the computational overhead introduced by our framework using NVIDIA A800 GPUs. The preprocessing stage consists of two components: embedding-based clustering and LLM-based task labeling. While knowledge clustering is computationally negligible (approximately $\sim$ 4 minutes for 20k samples), the task complexity labeling requires approximately $\sim$ 3 hours for 10k domain samples.

To highlight the efficiency of our approach, we compared this overhead with the second-best baseline, PDPC. PDPC necessitates a dual-model scoring process involving both a weak and a strong proxy model. In our reproduction using 10k samples, the weak proxy (Llama-3.1-8B on $1\times$A800 GPU) required 2 hours, while the strong proxy (Llama-3.3-70B on $4\times$A800 GPUs) required 7 hours. This results in a total preprocessing time of 9 hours for PDPC, which is approximately $3\times$ slower than our method.

Furthermore, PDPC imposes a significant hardware burden by requiring multi-GPU setups to host the 70B-parameter model, whereas our method operates efficiently on a single GPU. Crucially, this preprocessing is a one-time initialization cost. In contrast to the recurrent computational demand of the supervised fine-tuning phase, which requires approximately $\sim$ 13 hours per run in our setup, the constructed curriculum structure can be reused indefinitely across multiple training iterations. Consequently, the amortized time cost becomes marginal, making our framework a highly cost-effective solution for achieving significant per-

Method	Stage 1		Stage 2	
	General	Medical	General	Medical
Sequential	0.7460	0.5320	0.6774	0.6442
Ours	0.7344	0.5925	0.7291	0.6490

Table 6: Performance evolution across two stages for Sequential and Ours (Dual-Dimensional Curriculum) settings.

formance gains in domain adaptation.

G Continual Learning Metrics

We conducted a two-stage evaluation ($\text{stage}_1 \rightarrow \text{stage}_2$) based on Qwen3-8B. The "Sequential (General $\rightarrow$ Medical)" setting follows standard CL with explicit task boundaries. In contrast, "Ours (Dual-Dimensional Curriculum)" evaluates the midpoint and end of a single continuous mixed curriculum. While "Ours" lacks explicit boundaries for strict CL metrics, we compute standard FWT/BWT for the Sequential setting and provide a complementary process-level analysis for Ours. The results are shown in Table 6.

Forward Transfer (FWT)

Under the Sequential setting, FWT $= 0.5320 - 0.5513 = -0.0193$ (the base model achieves 0.5513 in the medical domain), indicating that training solely on the general domain yields no positive transfer to the subsequent medical task. For Ours, although strict FWT computation does not apply, the stage 1 medical performance already reaches 0.5925 (significantly higher than 0.5320), demonstrating that our continuous curriculum establishes domain-specific capabilities more effectively early on.

Backward Transfer (BWT)

Under the Sequential setting, BWT $= 0.6774 - 0.7460 = -0.0686$, indicating a noticeable degradation in general-domain capabilities due to the explicit stage transition. For Ours, the general-domain performance drops by only 0.0053 ($0.7344 \rightarrow 0.7291$). This shows that continuous curriculum training suffers minimal performance drift while maintaining superior final medical performance (0.6490 vs. 0.6442).

H Data-Source Ablation

To determine whether the performance improvements in the medical domain arise from our proposed curriculum ordering strategy or simply from the inclusion of reasoning-heavy data (i.e., MedThoughts), we conducted a data-source ablation study.

We evaluated the model under three different training data configurations:

- **Only MedInstruct**: Training exclusively on instruction-based medical data.
- **Only MedThoughts**: Training exclusively on reasoning-augmented medical data.
- **All (Ours)**: Training on the full mixture of both data sources using our proposed curriculum.

To ensure a fair comparison, all other training settings—including the total number of training samples and the curriculum scheduling strategy—were kept identical across the three settings. The results are shown in Table 7.

I Scaling Experiment Details

Due to space limitations, we present the extended experiments in this section. Specifically, we first report the results of scaling the supervised fine-tuning dataset to 60k instances in Table 8. We then provide the detailed results for models of different scales, including Qwen3-8B, 4B, and 1.7B, in Table 9, 10, and 11, respectively.

Method	Medical						General			
	MedQA	MMC.QA	PM.QA	MMLU-P.	GPQA	Avg.	MMLU	MMLU-P.	ARC	Avg.
Only MedInstruct	0.6602	0.6172	0.7540	0.6777	0.5244	0.6467	0.7387	0.5472	0.8931	0.7263
Only MedThoughts	0.6783	0.5867	0.7477	0.6923	0.5461	0.6502	0.7191	0.5610	0.8595	0.7132
All	0.6692	0.6014	0.7510	0.6951	0.5282	0.6490	0.7291	0.5878	0.8703	0.7291

Table 7: Performance comparison across different medical data sources under the same curriculum strategy.

Method	Medical						General			
	MedQA	MMC.QA	PM.QA	MMLU-P.	GPQA	Avg.	MMLU	MMLU-P.	ARC	Avg.
Base	0.5734	0.5316	0.6680	0.5452	0.4384	0.5513	0.6313	0.4325	0.7500	0.6046
DMT	0.5632	0.5503	0.6550	0.5663	0.4745	0.5618	0.6603	0.4835	0.8115	0.6517
PDPC	0.6294	0.6238	0.7360	0.6856	0.5218	0.6393	0.7413	0.5806	0.8732	0.7317
Ours	0.6853	0.5989	0.7510	0.7165	0.5421	0.6587	0.7465	0.5782	0.8979	0.7408

Table 8: Scaling experiment results on 60k SFT dataset using Qwen3-8B.

Method	Medical						General			
	MedQA	MMC.QA	PM.QA	MMLU-P.	GPQA	Avg.	MMLU	MMLU-P.	ARC	Avg.
Qwen3-8B										
Base	0.5734	0.5316	0.6680	0.5452	0.4384	0.5513	0.6313	0.4325	0.7500	0.6046
Shuffle	0.6095	0.5419	0.7460	0.6566	0.4923	0.6093	0.6932	0.5501	0.8191	0.6875
PDPC	0.5981	**0.6101**	0.7480	0.6697	0.4948	0.6241	**0.7449**	0.5836	0.8579	0.7288
DMT	0.5545	0.5522	0.6140	0.5706	0.4564	0.5495	0.6415	0.4325	0.7969	0.6236
Ours-Reverse	0.6097	0.5806	0.7230	0.6097	0.4871	0.6020	0.6717	0.5270	0.7645	0.6544
w/o Knowledge	0.5679	0.5689	0.7440	0.6456	0.4871	0.6027	0.7081	0.5551	0.8694	0.7109
w/o Task	0.6504	0.5656	**0.7540**	0.6247	0.5153	0.6220	0.6935	0.5442	0.8199	0.6859
Ours	**0.6692**	0.6014	0.7510	**0.6951**	**0.5282**	**0.6490**	0.7291	**0.5878**	**0.8703**	**0.7291**

Table 9: Scaling experiment results on Qwen3-8B.

Method	Medical						General			
	MedQA	MMC.QA	PM.QA	MMLU-P.	GPQA	Avg.	MMLU	MMLU-P.	ARC	Avg.
Qwen3-4B										
Base	0.3034	0.3428	0.3610	0.2620	0.2956	0.3130	0.4360	0.1945	0.4308	0.3538
Shuffle	0.5687	0.4867	0.7250	0.5543	0.3974	0.5464	0.6468	0.4936	0.7559	0.6321
PDPC	0.6068	0.5411	0.7200	0.5934	0.4230	0.5769	**0.6962**	0.5267	0.8267	0.6832
DMT	0.5247	0.4950	0.6690	0.5563	0.4128	0.5316	0.6455	0.4480	0.7866	0.6267
Ours-Reverse	0.6025	0.5259	0.7230	0.5602	0.3717	0.5567	0.6814	0.5144	0.8037	0.6665
w/o Knowledge	0.6056	0.5303	0.7390	**0.6354**	0.4364	0.5893	0.6586	0.4609	0.8260	0.6485
w/o Task	0.6056	0.5474	0.7290	0.6280	0.4211	0.5862	0.6883	0.4595	0.8319	0.6599
Ours	**0.6221**	**0.5599**	**0.7460**	0.6239	**0.4428**	**0.5989**	0.6848	**0.5300**	**0.8471**	**0.6873**

Table 10: Scaling experiment results on Qwen3-4B.

Method	Medical						General			
	MedQA	**MMC.QA**	**PM.QA**	**MMLU-P.**	**GPQA**	**Avg.**	**MMLU**	**MMLU-P.**	**ARC**	**Avg.**
Qwen3-1.7B										
Base	0.3362	0.3526	0.3440	0.2465	0.2948	0.3148	0.4299	0.1965	0.4146	0.3470
Shuffle	0.4397	0.4312	0.6280	0.4442	0.3076	0.4501	0.5784	0.3672	0.7448	0.5635
PDPC	0.4401	0.4344	0.6200	0.4762	0.3666	0.4675	0.6078	0.3862	**0.7883**	**0.5941**
DMT	0.4256	0.4019	0.6090	0.4362	0.3723	0.4490	0.5529	0.3659	0.7385	0.5524
Ours-Reverse	0.4359	0.4085	0.6190	0.4475	0.3256	0.4473	0.5882	0.3783	0.7613	0.5759
w/o Knowledge	0.4485	**0.4509**	0.6230	0.4736	0.4128	0.4818	0.5786	0.3690	0.7369	0.5615
w/o Task	0.4438	0.4212	0.6370	0.4338	0.3076	0.4487	0.5745	0.3857	0.7558	0.5720
Ours	**0.4571**	0.4344	**0.6570**	**0.4840**	**0.4251**	**0.4915**	**0.6152**	**0.3938**	0.7696	0.5929

Table 11: Scaling experiment results on Qwen3-1.7B.