MoL for LLMs: Dual-Loss Optimization to Enhance Domain Expertise While Preserving General Capabilities

Anonymous ACL submission

Abstract

Although Large Language Models (LLMs) perform well in general tasks, domain-specific applications suffer from hallucinations and accuracy limitations. Continual Pre-Training (CPT) approaches encounter two key issues: (1) domain-biased data degrade general language skills, and (2) improper corpus-mixture ratios limit effective adaptation. To address these, we propose a novel framework, Mixture of Losses (MoL), which decouples optimization objectives for domain-specific and general corpora. Specifically, cross-entropy (CE) loss is applied to domain-corpus to ensure knowledge acquisition, while Kullback-Leibler (KL) divergence aligns general-corpus training with the base model's foundational capabilities. This dual-loss architecture preserves universal skills while enhancing domain expertise, avoiding catastrophic forgetting. Empirically, we validate that a 1:1 domain-to-general corpus ratio optimally balances training and overfitting without the need for extensive tuning or resource-intensive experiments. Furthermore, our experiments demonstrate significant performance gains compared to traditional CPT approaches, which often suffer from degradation in general language capabilities; our model achieves 27.9% higher accuracy on the Math-500 benchmark in the non-think reasoning mode, and an impressive 83.3% improvement on the challenging AIME25 subset in the think mode, underscoring the effectiveness of our approach.

1 Introduction

Despite the remarkable success of Large Language Models (LLMs) in general text and code generation tasks (Grattafiori et al., 2024; Yang et al., 2024, 2025; Guo et al., 2025), challenges persist in domain-specific applications, notably in the form of hallucinations and inadequate accuracy. Continual Pre-Training (CPT) strategies have been proposed to address these issues (Sun et al., 2020; Jin et al., 2021b; Mendieta et al., 2023). However, two major problems arise with such approaches. Firstly, there is the challenge of maintaining general capabilities in CPT. Due to the limited quantity and quality of domain-specific data, along with its divergence from general data distributions, certain general competencies of LLMs may experience unpredictable degradation, even catastrophic forgetting (Cossu et al., 2024). Secondly, the determination of the optimal mixture ratio between the general-corpus and downstream domain-corpus remains a persistent challenge. While a sufficient proportion of general-corpus data is indispensable to preserve the model's foundational capabilities, identifying the ideal balance between the two corpora remains elusive, resulting in suboptimal performance of the fine-tuned model (Mehta et al., 2023; Wu et al., 2022). Recent work introduces the domain-specific Scaling Law to determine the optimal mixture ratio in CPT (Que et al., 2024). However, this Scaling Law primarily focuses on achieving an optimal compromise between domain-specific capabilities and general capabilities.

This paper introduces a novel training framework based on Mixture of Losses (MoL) computation to elegantly address the above two primary problems in CPT. During training, domain-corpus and general-corpus are randomly shuffled, but distinct loss functions are applied to each dataset type. Specifically, traditional cross-entropy (CE) loss is employed for domain-corpus to ensure effective learning of domain knowledge, while the loss for general-corpus is calculated using the Kullback-Leibler (KL) divergence relative to the base LLM (Hinton et al., 2015). This dual-strategy approach ensures that LLMs effectively incorporate specialized domain knowledge through CE optimization while maintaining the stability of their general capabilities via KL divergence. Furthermore, the inherent dichotomy of corpora into general and domain-specific categories naturally suggests an optimal 1:1 ratio between the two datasets (Abdelhamid and Desai, 2024; Carriero et al., 2025). This balanced training configuration mitigates potential model biases that could arise from dataset imbalance, ensuring more equitable learning across both knowledge domains. Our main contributions are summarized as follows.

(1) The MoL framework ensures the simultaneous preservation of general capabilities and enhanced domain-specific performance through its dual-loss architecture. By decoupling the optimization objectives for domain-corpus (via CE) and general-corpus (via KL divergence), the model avoids the degradation of foundational skills while systematically absorbing specialized knowledge. This is empirically validated through controlled experiments that demonstrate consistent performance gains in domain tasks without sacrificing general capabilities.

(2) We empirically establish the rationale behind the 1:1 corpus ratio as an optimal balance for hybrid training. This not only provides a principled guideline for dataset composition but also generalizes across diverse domains, eliminating the need for costly hyperparameter tuning for ratio optimization.

2 Related Work

Domain-specific CPT The domain-specific CPT paradigm is primarily designed to enhance the performance of LLMs on downstream tasks within specialized domains, such as medical consultation and legal Q&A systems (Qiu et al., 2024; Singhal et al., 2023; Yue et al., 2024). Typically, researchers need to curate high-quality domain-corpus alongside a certain volume of general-corpus for CPT. However, determining the optimal proportion of these two data components remains a challenging and computationally intensive task, often requiring extensive GPU resources for iterative optimization to achieve satisfactory results (Cossu et al., 2024; Mehta et al., 2023; Wu et al., 2022). Recent advancements in domain-specific Scaling Laws have attempted to provide systematic guidelines for corpus composition in CPT (Que et al., 2024), yet practical implementation still proves cumbersome and heavily dependent on numerous fitting experiments for calibration.

LLMs Distillation To transfer the capabilities of LLMs to a smaller one, knowledge distillation is commonly used (Hinton et al., 2015; Gou et al., 2021). When only the teacher model's API is accessible or there are vocabulary mismatches between the models, the black-box distillation approach is typically employed (Taori et al., 2023; Chiang et al., 2023; Peng et al., 2023). However, for open source LLMs with a shared vocabulary, white-box distillation is generally preferable (Sanh et al., 2019; Wang et al., 2020; Song et al., 2020). This method leverages the per token KL divergence between the teacher- and student-model distributions to compute the training loss. To mitigate the tendency of student models to overemphasize low-probability regions in the teacher distribution, recent studies have proposed substituting the conventional forward KL divergence with reverse KL divergence (Gu et al., 2023).

Learning without Forgetting In traditional neural network frameworks, incrementally introducing new capabilities into multitask architectures typically requires access to all task datasets, which is often impractical due to the inaccessibility of historical data and the prohibitive computational costs associated with retraining (Caruana, 1997). In the context of convolutional neural network (CNN) classification tasks, a regularization strategy combining KL divergence with CE loss in a weighted formulation has been proposed to address catastrophic forgetting when the model capabilities are incrementally expanded(Li and Hoiem, 2017). Empirical evaluations demonstrate that this approach achieves performance comparable to the upper bound established by joint training of all tasks simultaneously, offering a computationally efficient alternative to full retraining while mitigating the degradation of previously learned skills.

3 Methods

The roles of domain-corpus and general-corpus in CPT fundamentally differ. Domain-corpus are primarily designed to enhance a model's domain-specific capabilities by fine-tuning its understanding and generation within specialized contexts. In contrast, general-corpus serve to preserve and refine the model's general capabilities, which are critical for both ensuring broad applicability in diverse tasks and maintaining foundational competencies such as chain-of-thought (CoT) reasoning (Jaech et al., 2024; Xie et al., 2024).

A notable method to preserve the capabilities of LLM during training is the use of KL divergence as an objective function (Adler et al., 2021).

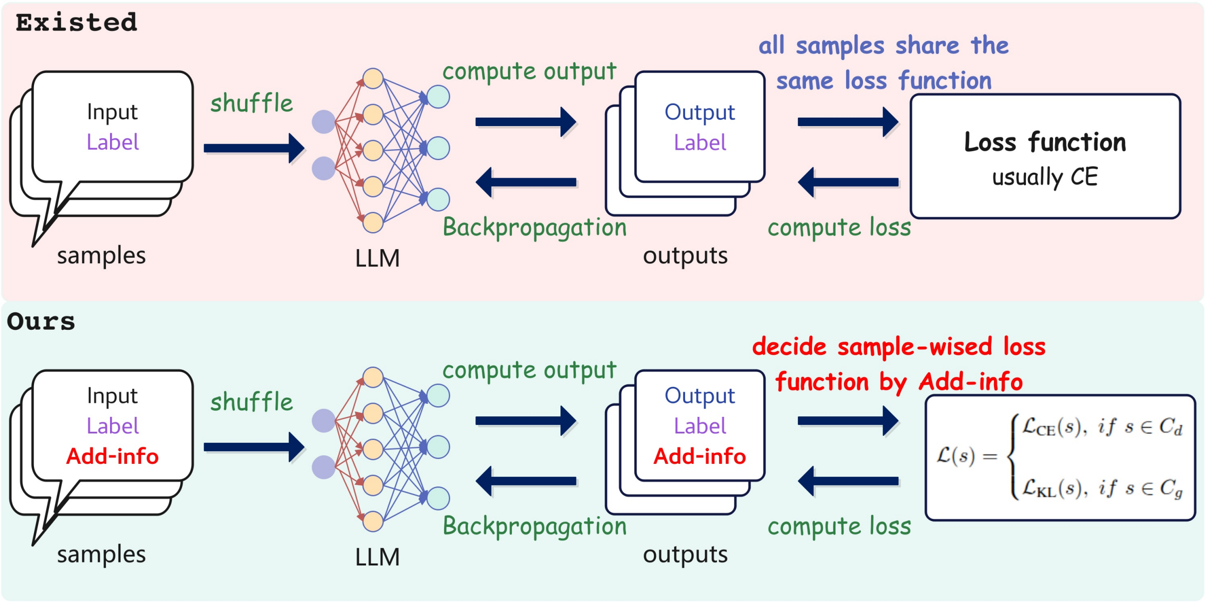

Figure 1: Schematic illustration of the MoL framework architecture. Unlike existed single-objective pre-training approaches, our MoL framework introduces an additional metadata input ("add-info") to distinguish between domain-specific and general corpora during training. This information determines the loss function selection: CE loss for domain corpora and KL divergence loss for general corpora (highlighted in red). The model's forward computation and backpropagation mechanisms retain the standard implementation pipeline of traditional LLMs.

Unlike traditional CE loss, which enforces deterministic "hard labels" by treating each token as an absolute target, KL divergence treats the output probability distribution of a base model as "soft labels" (Gu et al., 2023). This approach allows the target model to learn context-dependent generation patterns rather than memorizing fixed token sequences.

Thus, our MoL framework integrates domain-specific and general corpora under a dual-perspective optimization strategy in CPT. For domain-corpus, we employ the CE loss with hard labels to enforce precise domain-specific knowledge acquisition. While for general-corpus, the KL divergence loss with soft labels is adopted to preserve the model's pre-existing generalization capabilities. The loss function $\mathcal{L}$ for each sequence s in our MoL framework is formulated as follows:

$$\mathcal{L}(s) = \begin{cases} \mathcal{L}_{\text{CE}}(s),\ \ if\ \ s \in C_d, \\ \\ \mathcal{L}_{\text{KL}}(s),\ \ if\ \ s \in C_g, \end{cases} \qquad (1)$$

$$\mathcal{L}_{\text{CE}}(s) = -\frac{1}{n_s}\sum_i log\, p_\theta(s_i), \qquad (2)$$

$$\mathcal{L}_{\text{KL}}(s) = \frac{1}{n_s}\sum_i \text{KL}[p_\theta || p_0](s_i), \qquad (3)$$

where the average is performed over the total number of effective tokens (n_s). p_0 represents the probability distribution of the base LLM, and p_θ denotes the probability distribution of the CPT model parameterized by θ. The sets C_d and C_g correspond to domain-specific and general corpora, respectively. The specific operational workflow of our MoL framework is illustrated in Figure 1. This dual-loss architecture parallels human cognitive development: the KL divergence loss maintains alignment with foundational knowledge (like retaining language fundamentals during domain expertise acquisition), while the CE loss drives intentional knowledge expansion (similar to targeted skill development). The combination ensures that the model both preserves its general capabilities and systematically builds domain-specific expertise through complementary learning modes.

In practice, we introduce a small coefficient α to slightly adjust the final loss function to ensure training stability (Müller et al., 2019). Specifically, for the domain-corpus, the loss function is defined as

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{KL}}, \qquad (4)$$

while for the general-corpus, it is formulated as

$$\mathcal{L} = \alpha\mathcal{L}_{\text{CE}} + (1-\alpha)\mathcal{L}_{\text{KL}}. \qquad (5)$$

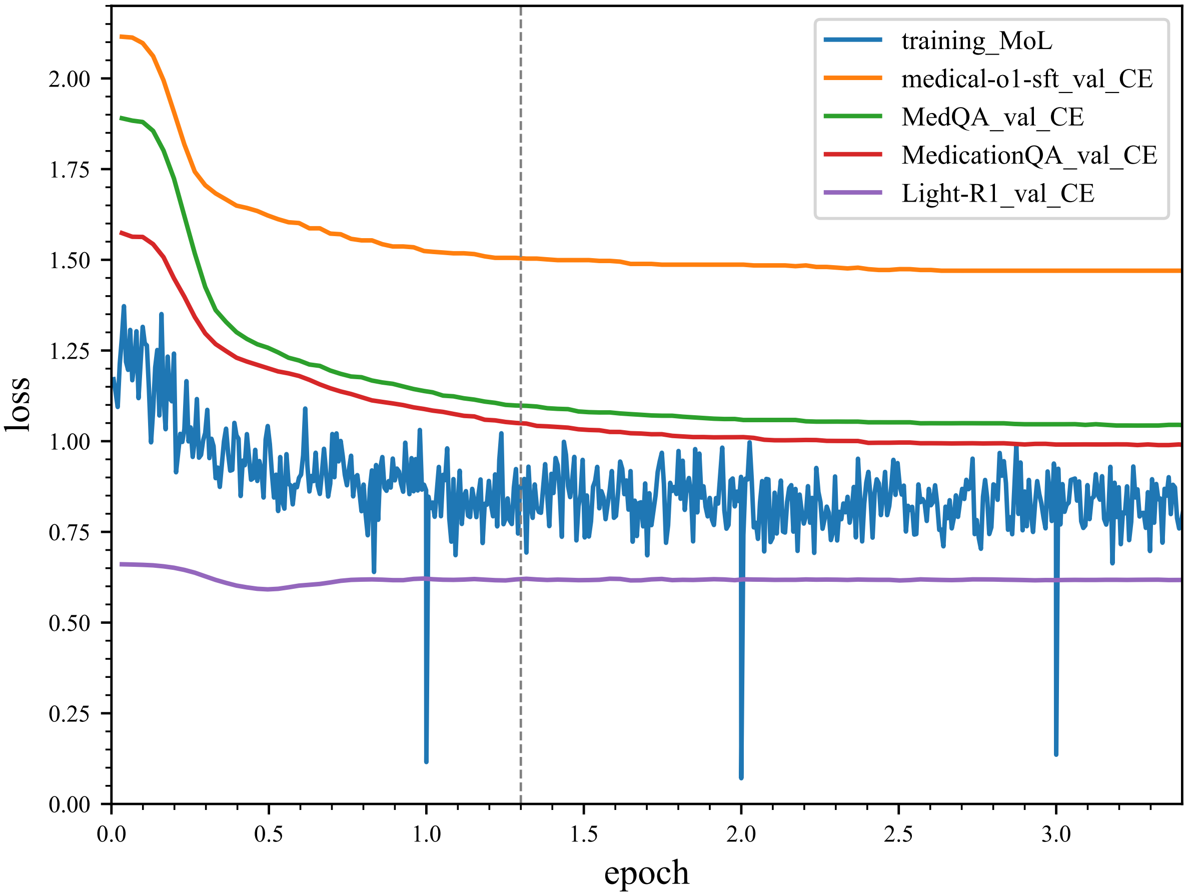

Figure 2: Training loss evolution across aggregated datasets and individual subsets, depicting CE loss dynamics for both domain-specific and general corpora. "train_MoL" represents the loss on the training set under the MOL framework. The domain-corpus include medical-o1-sft, MedicationQA, and MedQA, while Light-R1 is the general-corpus. The "_val" indicates the validation set, and "_CE" denotes CE loss. The validation set's CE loss is calculated every 10 steps, resulting in a smoother curve compared to the training set loss. The CE loss for general-corpus remains nearly constant throughout training, while the domain-corpus exhibits a steady decline in loss until reaching convergence at nearly 1.3 epochs (marked by the dashed Line).

In our experiments, α is set to 0.01 unless otherwise specified.

In Equation 3, we adopt the proposal of reverse KL divergence to mitigate overestimation of low-probability regions in the base model's output distribution (Gu et al., 2023). To further enhance regularization effectiveness, we introduce a cross-model probability aggregation scheme that jointly considers the probability distributions from both the base LLM and the CPT model for low-probability tokens, as formulated in Equation A. This optimization framework significantly reduces GPU memory consumption during KL divergence computation.

4 Experiments

Our objective is to validate the efficacy of the MoL training framework through empirical evaluation in the medical domain. Specifically, we conduct CPT on a hybrid dataset comprising medical-domain corpora and open source corpus using an open source model architecture. Subsequently, we evaluate the trained model's performance in both the medical domain and general domain to assess the validity and robustness of the proposed framework under real-world application scenarios.

Base Model The open source Qwen3-8B model (Yang et al., 2025) serves as the base for CPT. Additionally, this model is utilized to compute KL divergence during training within the MoL framework for consistency in optimization.

Training Our training data are derived from two primary sources:
(1) Domain-corpus: Training sections from the medical-o1-sft (Chen et al., 2024), MedicationQA (Abacha et al., 2019), and MedQA (Jin et al.,

		Qwen3-8B	+ D&G 1:1
Domain	MedQA	74.87	**77.25**
	MMLU-cli	78.86	**80.52**
General	C-Eval	67.45	**77.65**
	MMLU	72.56	**75.79**
Coding	MBPP	**68.40**	68.00
	HumanEval	**86.59**	82.32
Math (Non-thinking)	MATH-500	**85.40**	84.40
Math (Thinking)	MATH-500	96.60	**97.80**
	AIME 24	76.67	**80.00**
	AIME 25	66.67	**73.33**

Table 1: Performance comparison of various models across different task categories, including Domain, General, Coding, and Math tasks. The metrics represent the accuracy or performance scores achieved by each model on the respective tasks. The **D&G 1:1** refers to the training of base model using an nearly equal mix of domain-specific and general corpora.

2021a).

(2) General-corpus: High-quality chain-of-thought data from the open source Light-r1 corpus (Wen et al., 2025), serving as a supplementary training resource for broader reasoning capabilities.

With applying chat-template and concatenating, we can adopt a mixed CPT strategy, enabling training both textual and QA samples within a single pipeline. (Yang et al., 2024) The total templated and concatenated domain-corpus comprises approximately 10,000 samples, with 100 samples randomly selected as the validation set. For the general-corpus, we use the stage1 part of the Light-r1 dataset with 76 K training samples, allowing for flexible adjustment of the ratio between domain-specific and general corpora during experimental design.

Experiments are conducted using the Low-rank adaptation (LoRA) training approach (with a rank of 64) (Hu et al., 2022). All training is performed with the model's context length fixed at 8,192 tokens, ensuring compatibility with long input sequences. For LoRA training, we use a learning rate of 1e-4. All other hyperparameters remain consistent across experiments, including a cosine decay learning schedule with a warm-up ratio of 0.1 and a global batch size of 128.

Evaluation We perform a comprehensive evaluation of the trained models. The evaluation focuses on its performance in terms of domain, general knowledge, mathematics, and coding capabilities. The evaluation dataset of the trained model contains these benchmarks:

- **Domain Tasks:** We use benchmarks including predefined test set of MedQA (Jin et al., 2021a), including 3426 Chinese questions and 1273 US questions. And MMLU-cli, the medicine-related test data in MMLU (Hendrycks et al., 2020), including 134 Anatomy questions, 264 Clinical questions, 143 College biology questions, 172 college medicine questions, 99 Medical Genetic questions and 271 professional medicine questions.
- **General Tasks:** MMLU (Hendrycks et al., 2020) and C-Eval (Huang et al., 2023) (5 shots)
- **Coding Tasks:** MBPP (Austin et al., 2021) and HumanEval (Chen et al., 2021)
- **Math Tasks:** MATH-500 (Lightman et al., 2023), AIME 24 and AIME 25 (AIME, 2025).

The evaluation of Domain, General, and Coding tasks occurs under a non-thinking mode. In contrast, MATH-500 is assessed under both thinking and non-thinking modes. AIME 24 and AIME 25 are exclusively evaluated in thinking mode. For all models operating in thinking mode, we employ a sampling temperature of 0.6, a top-p value of 0.95, and a top-k value of 20. In the non-thinking mode for General, Coding, and Math Tasks, the sampling hyperparameters are configured as follows: temperature = 0.7, top-p = 0.8, top-k = 20, and presence penalty = 1.5. The settings of evaluation parameter above are fully consistent with the official Qwen3 . For domain tasks evaluated in non-thinking mode, the sampling hyperparameters are set with a temperature of 0.01. For both thinking and non-thinking modes, the maximum output length is capped at 30,720 tokens. Non-thinking mode is achieved by setting "enable_thinking=False" (Yang et al., 2025).

		D&G 1:1	D&G 1:0.5	D&G 1:1.5	D&G 1:2
Domain	MedQA	**77.25**	75.68	77.17	77.19
	MMLU-cli	**80.52**	79.22	79.87	79.59
General	C-Eval	**77.65**	77.04	77.20	77.59
	MMLU	75.79	**76.48**	74.73	69.36
Coding	MBPP	68.00	66.00	67.00	**69.00**
	HumanEval	82.32	81.10	80.49	**83.54**
Math (Non-thinking)	MATH-500	**84.40**	84.20	81.20	80.60
Math (Thinking)	MATH-500	**97.80**	96.40	97.20	96.40
	AIME 24	**80.00**	**80.00**	76.67	70.00
	AIME 25	**73.33**	70.00	70.00	66.67

Table 2: Performance evaluation of model variants trained on Qwen3-8B across diverse task categories, including Domain, General, Coding, and Math tasks. The **D&G** ratios indicate the adjusted proportions of domain-specific to general corpora used for training, showing the influence of these ratios on model performance across different tasks.

5 Results

5.1 Main Results

Determination of Optimal Training Steps We first conduct experiments on Qwen3-8B using a nearly 1:1 ratio of medical and general corpora to investigate the training dynamics of the MoL framework. As shown in Figure 2, the CE loss for general corpora remains nearly constant throughout training, due to the use of KL divergence as the loss function for these samples. In contrast, the CE loss for domain-specific corpora exhibits a consistent downward trend until convergence. This observation aligns precisely with our hypothesis that MoL can effectively enhance domain knowledge while preserving general language capabilities. Notably, we observe that all datasets' CE losses approach convergence at approximately 1.3 training epochs. This finding establishes a critical reference point for subsequent model comparisons, and therefore we standardize all evaluations at this epoch for fair performance assessment across different training paradigms.

Performance Evaluation at Convergence The results of 1.3 training epochs are presented in Table 1. Our model using the MoL training framework demonstrates superior performance over the base Qwen3-8B model across three critical dimensions: domain-specific capabilities, general abilities, and math reasoning. We also observe a significantly larger discrepancy in C-Eval performance scores before and after training, which was primarily attributed to insufficient instruction-following (IF) capability in the base model. This limitation leads to systematic misinterpretation of multiple-choice answers during evaluation. However, the implementation of the MoL training approach effectively resolves this issue, resulting in complete elimination of parsing errors in the CPT model. Detailed comparisons are presented in Appendix B.

Given that the open source medical corpus has probably already been exposed to Qwen3, further training on these domain-specific data may yield limited performance improvements in specialized medical tasks. To address this limitation and further validate the robustness of our MoL approach, we conducted additional experiments using a different foundational architecture and an internal corpus. The domain-to-general data ratio was carefully balanced by augmenting the general component through threefold repetition of another internal general-corpus. This setup enabled us to maintain sufficient training scale while ensuring domain relevance. The results, as shown in Appendix C, demonstrate that the proposed method achieves notable performance gains on the internal domain evaluation set compared to the baseline. Importantly, the model's general linguistic capabilities remain consistent with the base model's performance on standard benchmarks, confirming that domain adaptation does not come at the cost of foundational language proficiency.

Optimal Domain-to-General Corpus Ratio Regarding the optimization problem of domain-to-general corpus ratio, we conduct extensive experiments using the medical domain-specific corpus as the fixed unit and vary the proportion of general-corpus (0.5, 1, 1.5, and 2) accordingly, as illustrated in Table 2. The experimental results demonstrate that, for our MoL framework, a ratio near 1:1

		D&G 1:1	**D&G 1:1 CE**	**D&G 1:1** ($\alpha = 0.5$)
Domain	MedQA	77.25	**77.57**	73.19
	MMLU-cli	**80.52**	80.24	78.95
General	C-Eval	**77.65**	76.23	73.86
	MMLU	75.79	57.71	**76.99**
Coding	MBPP	**68.00**	62.20	64.80
	HumanEval	82.32	78.66	**84.15**
Math (Non-thinking)	MATH-500	**84.40**	66.00	82.40
Math (Thinking)	MATH-500	**97.80**	94.20	96.60
	AIME 24	**80.00**	63.33	70.00
	AIME 25	**73.33**	40.00	56.67

Table 3: Performance evaluation of model variants trained on Qwen3-8B across diverse task categories, including Domain, General, Coding, and Math tasks. The **D&G 1:1** corresponds to the definition provided in Table 1. The **D&G 1:1 CE** configuration utilizes CE as the loss function across all data. The final column represents results obtained with a α parameter of 0.5 in Equation 4 and 5.

achieves the most balanced performance between domain-specific and general language capabilities. This configuration consistently outperforms other tested ratios across most of the evaluation metrics, indicating that the optimal trade-off point for this specialized loss mechanism lies in maintaining approximately equal proportions of domain-specific and general corpora.

5.2 Ablation Studies and Critical Analysis

5.2.1 Influence of KL Divergence

To evaluate the necessity of KL divergence in our framework, we conduct one control experiment under the optimal mixture ratio of 1:1 (Table 2). This experiment replaces all KL divergence calculations with CE counterparts. Our training results demonstrate significantly superior performance compared to this alternative, as detailed in Table 3. The significant performance gains, particularly a 27.9% ($\frac{84.40-66.00}{66.00}$) increase in accuracy on the Math-500 benchmark in the non-think reasoning mode and an impressive 83.3% ($\frac{73.33-40.00}{40.00}$) improvement on the challenging AIME 25 subset in the think mode. These results not only reflect the superiority of our method, but also suggest that the alternative approach using CE may suffer from a degradation in generalization capability under the same experimental setup. This decline in general performance across different reasoning modes and benchmarks further supports the necessity of KL divergence in enhancing the robustness and adaptability of our model.

Regarding domain-specific capabilities, we observe that the CE losses of various domain-specific corpora under the MOL framework are closely aligned with those of the control experiment that employs CE exclusively, as shown in Figure 3 **(A)**. This suggests that our method is on par with the traditional CE-based training approach in terms of domain adaptation and specialization. However, as previously discussed, our framework significantly outperforms the CE-only alternative in general reasoning tasks, particularly in complex and abstract reasoning scenarios.

Furthermore, we visualize the gradient magnitudes of both the KL divergence-based and CE-based training frameworks during the training process, as presented in Figure 3 (**B**). While the two curves exhibit a similar overall trend, the gradients from the KL divergence-based framework remain consistently smaller than those from the CE-based alternative throughout the training process, until convergence is reached. Upon convergence, the two gradient curves become sufficiently close, indicating that the gradient contribution from the KL divergence gradually approaches that of the CE counterpart. This observation is particularly noteworthy, given that the KL divergence is computed with respect to the base model, which implies that the initial gradient from KL divergence is close to zero. As training progresses, the KL divergence begins to exert a more significant influence, leading to a gradual increase in its gradient contribution. This dynamic behavior resembles a negative feedback mechanism, similar to those identified in adaptive learning systems (Zhao et al., 2018). Consequently, the introduction of KL divergence within the MOL framework elegantly ensures the preservation of

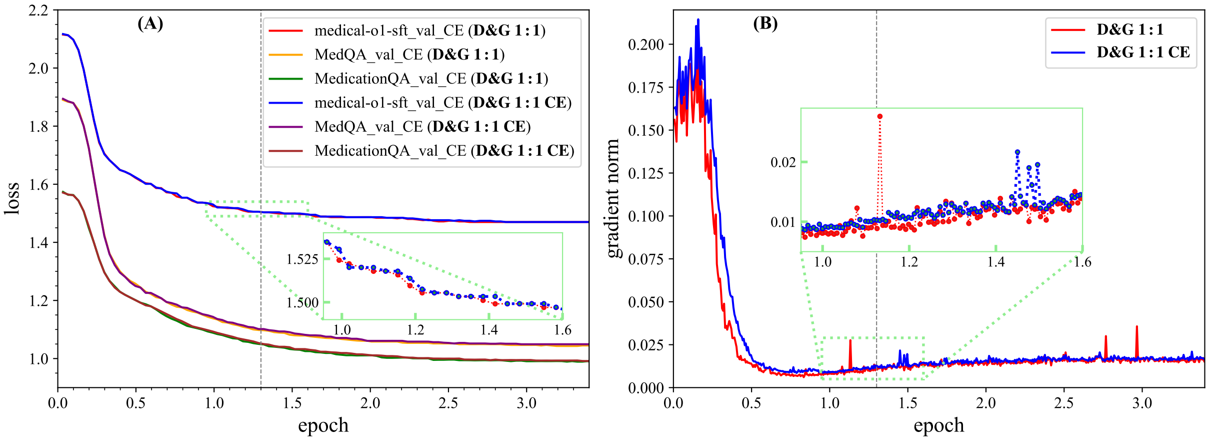

Figure 3: **(A)** Comparison of CE loss across various validation sets during training on Qwen3-8B. The main plot depicts the loss trends across different domain-corpus, including medical-o1-sft, MedQA and MedicationQA, while the inset offers a magnified view focusing on the performance of the medical-o1-sft under the two configurations. Specifically, **D&G 1:1** refers to the setup described in Table 1, while **D&G 1:1 CE** denotes the alternative configuration that replaces KL divergence with CE throughout the training. The "_val" indicates the validation set, and "_CE" denotes CE loss. **(B)** Comparison of gradient norm during training. The main plot illustrates the gradient norm dynamics of the **D&G 1:1** and **D&G 1:1 CE** configurations across training epochs. While both configurations exhibit similar temporal evolution patterns, the **D&G 1:1** consistently shows smaller magnitudes than **D&G 1:1 CE** until convergence, after which the two curves align closely. The inset provides a zoomed-in view centered at 1.3 training epochs.

general reasoning capabilities, thereby demonstrating its necessity.

5.2.2 Influence of Coefficient α

Furthermore, while the efficacy of KL divergence has been demonstrated, it is essential to justify whether the hyperparameter α in Equation 4 and 5 significantly influences the final outcomes. To investigate this, we configure α to 0.5, effectively assigning equal importance to domain-specific and general corpora. Our experiments reveal that a near-zero α value consistently yielded superior performance. The quantitative validation supporting this observation is detailed in Table 3.

6 Disscussion

This work proposes an MoL framework to address the dual challenges of continual learning in LLMs, by decoupling loss functions for domain knowledge and general knowledge. This aligns with lifelong learning principles (Zheng et al., 2025), enabling LLMs to dynamically integrate specialized knowledge without catastrophic forgetting.

The 1:1 general-domain corpus ratio, empirically validated as optimal, reflects a natural balance observed in real-world systems. For instance, in Retrieval-Augmented Generation (RAG), the ratio between retrieved external knowledge and pre-trained model priors often mirrors this equilibrium. Similarly, in agent-environment interactions, the proportion of environmental feedback (domain-specific) and prior knowledge (general) typically aligns with 1:1-like dynamics in context-action pairs. This inherent balance simplifies deployment, offering a scalable solution for adaptive LLMs in evolving domains.

7 Conclusion

This study introduces the MoL framework, a dual-loss architecture that synergistically preserves general language capabilities while enhancing domain-specific performance through decoupled optimization. By applying CE loss for domain-corpus training and KL divergence for general-corpus alignment, the framework mitigates catastrophic forgetting in foundational skills while systematically integrating specialized knowledge. The 1:1 domain-to-general corpus ratio is empirically validated as optimal, demonstrating its ability to prevent overfitting while avoiding laborious and computationally intensive hyperparameter tuning processes. These contributions establish MoL as a principled, scalable solution for multi-domain language model training, offering both theoretical insights and practical deployment advantages in real-world heterogeneous scenarios.

Limitations

While the MoL framework achieves notable improvements, one limitation lies in the unexpected enhancement of AIME reasoning and IF capabilities through general KL divergence alignment on general-corpus. This phenomenon, though empirically observed, lacks a systematic analysis of its underlying mechanism. More research is required to clarify its role in bridging general and specialized performance within the MoL framework.

References

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 25–29. IOS Press.

Mohamed Abdelhamid and Abhyuday Desai. 2024. Balancing the scales: A comprehensive study on tackling class imbalance in binary classification. *arXiv preprint arXiv:2409.19751*.

Aviv Adler, Jennifer Tang, and Yury Polyanskiy. 2021. Quantization of random distributions under kl divergence. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 2762–2767. IEEE.

AIME. 2025. AIME problems and solutions.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Alex Carriero, Kim Luijken, Anne de Hond, Karel GM Moons, Ben van Calster, and Maarten van Smeden. 2025. The harms of class imbalance corrections for machine learning based prediction models: a simulation study. *Statistics in Medicine*, 44(3-4):e10320.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 3(5).

Andrea Cossu, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, Tinne Tuytelaars, and Davide Bacciu. 2024. Continual pre-training mitigates forgetting in language and vision. *Neural Networks*, 179:106492.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

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

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. 2021b. Lifelong pretraining: Continually adapting language models to emerging corpora. *arXiv preprint arXiv:2110.08534*.

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

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. 2023. An empirical investigation of the role of pre-training in lifelong learning. *Journal of Machine Learning Research*, 24(214):1–50.

Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. 2023. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, Yuanxing Zhang, Xu Tan, Jie Fu, et al. 2024. D-cpt law: Domain-specific continual pre-training scaling law for large language models. *Advances in Neural Information Processing Systems*, 37:90318–90354.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ulrich Schollwöck. 2005. The density-matrix renormalization group. *Reviews of modern physics*, 77(1):259–315.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Kaitao Song, Hao Sun, Xu Tan, Tao Qin, Jianfeng Lu, Hongzhi Liu, and Tie-Yan Liu. 2020. Lightpaff: A two-stage distillation framework for pre-training and fine-tuning. *arXiv preprint arXiv:2004.12817*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8968–8975.

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

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, et al. 2025. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *arXiv preprint arXiv:2503.10460*.

Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan Fang Li, Guilin Qi, and Gholamreza Haffari. 2022. Pretrained language model in continual learning: A comparative study. In *International Conference on Learning Representations 2022*. OpenReview.

Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang, Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang Xie, and Yuyin Zhou. 2024. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *arXiv preprint arXiv:2409.15277*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.

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

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

Shengbin Yue, Shujun Liu, Yuxuan Zhou, Chenchen Shen, Siyuan Wang, Yao Xiao, Bingxuan Li, Yun Song, Xiaoyu Shen, Wei Chen, et al. 2024. Lawllm: Intelligent legal system with legal reasoning and verifiable retrieval. pages 304–321.

Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. 2018. Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proceedings of the 24th ACM*

SIGKDD international conference on knowledge discovery & data mining, pages 1040–1048.

Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. 2025. Towards lifelong learning of large language models: A survey. *ACM Computing Surveys*, 57(8):1–35.

A Optimization of KL Divergence

We introduce an additional regularization scheme by aggregating the probability values of low-probability tokens across both the base LLM and the CPT model. The aggregated probability distribution is then used for calculating the reverse KL divergence. This approach focuses the optimization process on aligning the high-probability regions of the base model's generation behavior, sharing conceptual parallels with physical systems analysis where attention is directed towards low-energy states that dominate system behavior (Schollwöck, 2005).

For example, the vocabulary of Qwen3 exceeds 150,000 tokens (Yang et al., 2025). However, we employ a dynamic probability truncation strategy to reduce the vast vocabulary size by focusing on the most probable tokens at each generation step. Specifically, given the base model's probability distribution, we retain the top n most probable tokens (denoted as set T), preserving their individual probabilities. The remaining probabilities outside of T are aggregated into a single residual probability mass,

$$\begin{aligned} p'_{0,\,residual} &= \sum_{t \notin T} p_0(t), \\ p'_{\theta,residual} &= \sum_{t \notin T} p_\theta(t), \end{aligned} \tag{6}$$

resulting in a reduced distribution p' that maintains the total probability mass of 1. In our work, we set the parameter n to 31.

B Analysis of C-Eval Results

We observe that the C-eval benchmark exhibits lower evaluation results for the Qwen3-8B model. Analysis of the evaluation results 1 and 2 (already translated into English) reveals that even in non-thinking mode with 5-shot examples, the model may sequentially analyze options, leading to incorrect identification of the first capital letter as the final answer selection. The trained model demonstrates improved IF capabilities, enabling more accurate output generation based on provided examples.

C More Results

We evaluate the MoL framework using Qwen2.5-7B-Instruct as the base model, training on a 330M-token internal domain corpus and a 300M-token general corpus (100M-token Magpie-built corpus repeated three times). This configuration achieves a domain-to-general token ratio of approximately 1:1 while maintaining sufficient training scale. The general corpus was explicitly constructed using the Magpie framework (Xu et al., 2024). Full-parameter training was implemented with a learning rate of 1e-5 and a batch size of 1024, while all other hyperparameters were aligned with those specified in the main text.

The results showed significant performance improvements on the internal domain evaluation set compared to the baseline, with no degradation in general linguistic capabilities. The model retained consistent performance on standard benchmarks, demonstrating the framework's ability to preserve foundational language skills during domain-specific training. Table 4 summarizes these findings.

Qwen3-8B Model Outputs on C-eval with 5-Shot Examples

Wrong Result:

```
{
    "prompt": [
        ...(previous shots)
        {
            "role": "HUMAN",
            "prompt": ...(omitted prompt),
        },
        {
            "role": "BOT",
            "prompt": "D",
        },
        {
            "role": "HUMAN",
            "prompt": "The following is a single-choice question from a logic exam in China...\nAnswer: "
        }
    ],
    "origin_prediction": "This question tests an argument support type logical reasoning question. We need to find an option that can most effectively support the conclusion in the original text...### Option analysis:\n\nA...",
    "predictions": "A",
    "references": "C"
}
```

CPT Model Outputs on C-eval with 5-Shot Examples

Right Result:

```
{
    "prompt": [
        ...(previous shots)
        {
            "role": "HUMAN",
            "prompt": ...(omitted prompt),
        },
        {
            "role": "BOT",
            "prompt": "D",
        },
        {
            "role": "HUMAN",
            "prompt": "The following is a single-choice question from a logic exam in China..."\nAnswer: "
        }
    ],
    "origin_prediction": "C",
    "predictions": "C",
    "references": "C"
}
```

		Qwen2.5-7B-Instruct	**+ D&G 1:1**
Domain	Concept-explanation	58.33	**67.96**
	Summarize	39.84	**53.50**
	Simple-QA	48.67	**57.70**
	Ops-FAQ	17.37	**63.59**
	Product-FAQ	23.34	**41.32**
General	C-Eval	**79.10**	79.06
	MMLU	74.27	**74.39**
	CMMLU	78.67	**78.68**
	BBH	**69.70**	67.04
	HellaSwag	81.87	**81.91**
Coding	MBPP	**66.60**	64.80
	HumanEval	**81.10**	**81.10**
Math	MATH	**57.60**	57.56
	Gsm8k	85.14	**85.15**

Table 4: Performance comparison of various models across different task categories, including Domain, Business-related, Coding, General, and Math tasks. The **D&G 1:1** corresponds to the definition provided in Table 1.