L3Masking: Multi-task Fine-tuning for Language Models by Leveraging Lessons Learned from Vanilla Models

Yusuke Kimura[1], **Takahiro Komamizu**[2], **Kenji Hatano**[1]

[1] **Doshisha University, Japan**, [2] **Nagoya University, Japan**

{usk, taka-coma, hatano}@acm.org

Abstract

When distributional differences exist between pre-training and fine-tuning data, language models (LMs) may perform poorly on downstream tasks. Recent studies have reported that multi-task learning of downstream task and masked language modeling (MLM) task during the fine-tuning phase improves the performance of the downstream task. Typical MLM tasks (e.g., random token masking (RTM)) tend not to care tokens corresponding to the knowledge already acquired during the pre-training phase, therefore LMs may not notice the important clue or not effective to acquire linguistic knowledge of the task or domain. To overcome this limitation, we propose a new masking strategy for MLM task, called L3Masking [1], that leverages lessons (specifically, token-wise likelihood in a context) learned from the vanilla language model to be fine-tuned. L3Masking actively masks tokens with low likelihood on the vanilla model. Experimental evaluations on text classification tasks in different domains confirms a multi-task text classification method with L3Masking performed task adaptation more effectively than that with RTM. These results suggest the usefulness of assigning a preference to the tokens to be learned as the task or domain adaptation.

1 Introduction

Language Models (LM) pre-trained on generic documents such as BERT (Kenton and Toutanova, 2019) or GPTs (e.g., GPT-4 (Achiam et al., 2023)) may perform poorly on downstream tasks when the vocabulary or context used in the documents in each pre-training and downstream task differs (Gururangan et al., 2020; Shi et al., 2024). To bridge the domain gap between pre-training and fine-tuning, continual pre-training is used. Continual pre-training re-trains a model by applying the pre-training task again on the task or domain data (Xie et al., 2023). Recent studies have reported that multi-task learning (MTL) of downstream and pre-training tasks (e.g., masked language modeling (MLM)) during the fine-tuning phase can improve the performance of downstream tasks in comparison with continual pre-training (Dery et al., 2022, 2023; Kimura et al., 2023).

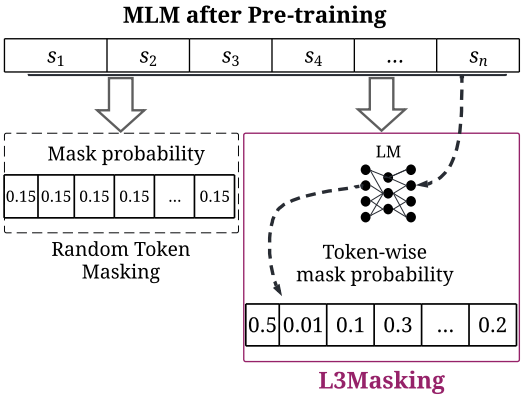

Figure 1: L3Masking vs. Random Token Masking. L3Masking determines masking tokens based on the pseudo-likelihood calculated through the vanilla model.

Existing task or domain adaptation methods for the encoder of Transformer architecture (Vaswani et al., 2017) typically utilize MLM, and random token masking (RTM) is mostly used masking strategy (Kenton and Toutanova, 2019; Liu et al., 2019). MLM is expected to use to adaptively learn the distribution of task and domain data from the learned distribution of the pre-training corpora. The MLM with simple strategy treats all tokens equally. However, existing MLMs ignore the linguistic knowledge already acquired by the language model, and, to learn the distribution properly, it requires large amount of time and data. Beside the fact that the amount of data for fine-tuning is limited, the more efficient masking strategy for MLM task is desired.

To overcome this, we propose a new masking

[1]The code is available at https://github.com/usk-Kimura/L3Masking

Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U), pages 53–62

strategy for MLM called L3Masking (Leveraging Lessons Learned from vanilla model) as an effective task or domain adaptation. Figure 1 highlights the difference between L3Masking and a popular and simple masking strategy Random Token Masking. L3Masking identifies tokens with low likelihoods as task- or domain-specific tokens that appear less frequently in the similar contexts in the generic documents, and it actively masks them so that LM learns these tokens during fine-tuning.

Unlike causal language modeling which computes the likelihood of a token in a sentence only from the preceding tokens, MLM can compute the likelihood of the token conditional on both preceding and subsequent tokens. This difference has led to variations in the idea of a sentence's likelihood and has been noted in what is called the pseudo-log-likelihood (PLL) (Kauf and Ivanova, 2023). Based on the PLL, this study defines a token-wise pseudo-likelihood in the downstream task sentence and actively mask tokens with low pseudo-likelihood.

In consequence, the contributions of this paper can be summarized as follows:

- **L3Masking**: This paper proposes a new masking strategy called L3Masking for MLM task in the multi-task text classification, which set token-wise mask probabilities for task or domain adaptation, enhancing the adaptability of LMs to new domains and tasks.
- **Validation**: Experimental evaluations reported in this paper validate the effectiveness of L3Masking through three text classification tasks in different domains, highlighting its improvement from the simply fine-tuned models and its superiority over random token masking in the comparison of masking strategy.
- **Efficiency**: This paper also demonstrate that L3Masking not only improves the text classification performance of models but also increases the efficiency of training in text classification tasks. By selectively masking task- and domain-specific tokens, L3Masking reduces the number of training epochs required while maintaining or improving accuracy.

2 Related Studies

This section describes the task or domain adaptation methods that have been studied in contexts of continual pre-training and fine-tuning.

2.1 Adaptation in Continual Pre-training

Continual pre-training is a method of continuing further pre-training with additional data to adapt a vanilla LM pre-trained by generic corpora to a specific task or domain (Gururangan et al., 2020; Xie et al., 2023). A fundamental assumption of the method, known as the Selective Language Modeling (SLM) (Lin et al., 2024), is that all tokens are not equally useful for adaptation. Specifically, a reference model is first prepared that is continually pre-trained on high-quality data for the domain in question. Then, from low-quality data containing many tokens that are not included in the documents of downstream tasks in the domain concerned, tokens with the necessary knowledge are identified and actively learned, thereby enabling effective and efficient continual pre-training.

The difference between SLM and L3Masking is the quality of the target documents. SLM relies on high-quality data from the domain to determine whether a token corresponds to that linguistic knowledge, therefore, the cost of collecting high-quality data is high. L3Masking differs from SLM in that it determines task- or domain-specific tokens based only on the data of downstream task.

2.2 Adaptation in Fine-Tuning

META-TARTAN (Dery et al., 2022) is an effective task or domain adaptation method that brings pre-training tasks into fine-tuning phase, and it is a multi-talk learning besides of downstream tasks. META-TARTAN performs the MTL with the downstream and the pre-training tasks as auxiliary tasks and dynamically weights the loss values of each task to increase the accuracy of the validation data in the downstream task based on meta-learning. META-TARTAN employ RTM, which masks tokens in a uniform random manner (Gururangan et al., 2020), in analogous with continual pre-training in MLMs.

Many masking strategies for the MLM task have been proposed, such as Knowledge Masking, PMI-Masking, and InforMask (Sun et al., 2019; Levine et al., 2021; Sadeq et al., 2022). These methods use PMI, which depends on the frequency of token occurrence and co-occurrence, to increase the probability of collocation being masked. However, as the size of the dataset in the post-training phase is limited compared to the pre-training corpus, these methods may be less effective with small amounts of data where the co-occurrence pattern of tokens

is less pronounced.

Using RTM in META-TARTAN may be not effective because the masking target masks tokens with regardless of the linguistic knowledge acquired in pre-training. In order to adapt a data distribution for downstream tasks that is different from pre-training, masking more tokens that are not plausible for the LM in a certain context may effectively lead to the acquisition of linguistic knowledge in the task or domain. Based on this idea, L3Masking identifies tokens that are not plausible in context based on the likelihood of each token in the context on the vanilla models.

3 L3Masking: the proposed method

This paper propose a new masking strategy of MLM task for task or domain adaptation, called L3Masking. The basic idea is to improve task and domain adaptability by actively masking tokens in sentences that are not well trained during pre-training. Figure 2 depicts the overview of L3Masking. L3Masking captures the tokens that are most likely to represent task- or domain-specific linguistic knowledge based on a token-wise likelihood. Since the likelihood cannot be calculated simply in bidirectional LM, L3Masking calculates the token-wise pseudo-likelihood and then masks more tokens with lower the pseudo-likelihood.

3.1 Pseudo-log-likelihood of a Sentence

In unidirectional LM, the log-likelihood of a sentence can be calculated by summation of the logarithm of the predicted probability of the tokens based on the preceding tokens. However, as MLM takes the tokens behind a token when predicting it into account, it expands the interpretation of the likelihood that can utilize the subsequent tokens in addition to the preceding tokens. Therefore, the following three methods are proposed to compute the pseudo-log-likelihood of a sentence in an MLM, namely, PLL-original (Salazar et al., 2020), PLL-word-l2r (Kauf and Ivanova, 2023), and PLL-whole-word (Kauf and Ivanova, 2023).

In the previous study, the PLL score calculated by PLL-word-l2r is considered the best pseudo-log-likelihood for a sentence (Kauf and Ivanova, 2023). PLL-word-l2r (PLL_{12r}) is based on word as a unit for masking and tokens of a word on the right (future direction) are not aware via masking during inference. This idea is formulated as follows:

$$\mathrm{PLL}_{12r}(S) := \sum_{w=1}^{|S|} \sum_{t=1}^{|w|} \log P_{\mathrm{MLM}}(s_{w_t} | S_{\setminus s_{w_{t' \geq t}}}) \tag{1}$$

where the t-th token s_{w_t} is subject to calculate a probability in a context represented as $S_{\setminus s_{w_{t' \geq t}}}$. For inference, the context is constructed by substituting the token sub-sequence of a word w, where the t-th token s_{w_t} is a part of, from s_{w_t} to the last token $s_{w_{t'}}$ of w. In other words, $S_{\setminus s_{w_{t' \geq t}}}$ is denoted as $(s_0, s_1, \ldots, s_{t-1}, \texttt{[MASK]}, ..., \texttt{[MASK]}, s_{t'+1}, \ldots, s_n)$.

3.2 Token-wise Pseudo-likelihood

In this study, the pseudo-likelihood (PL) of each token is calculated based on PLL_{l2r} (Eqn. (1)). In this study, this pseudo-likelihood of token s in a sentence S is called the PL of Token (PLT) and is defined as follows:

$$\begin{aligned} &\mathrm{PLT}\left(X = s_{w_t} \mid S_{\setminus s_{w_{t' \geq t}}}\right) \\ &\qquad = P_{\mathrm{MLM}}\left(s_{w_t} \mid S_{\setminus s_{w_{t' \geq t}}}\right) \end{aligned} \tag{2}$$

where X refers to the token for which the PLT is to be calculated.

For instance, given a sentence "The quick brown fox jumps over the lazy dog," suppose to calculate the pseudo-likelihood of the token "jump." Here, "jumps" is assumed to consist of two subwords: "jump" and the suffix "s." The context $S_{\setminus s_{w_{t' \geq t}}}$ is "The quick brown fox `[MASK]` `[MASK]` over the lazy dog." Using this context, the probability of "jump" is calculated $\mathrm{PLT}(X = \text{"jump"} \mid S_{\setminus s_{w_{t' \geq t}}})$.

3.3 Convert PLT to Mask Probability

In L3Masking, the PLT is the pseudo-likelihood itself, that is, the probability of the token in a context. Since our idea is to mask more tokens with lower likelihood, we take the complementary probability, PLT^c (Eqn. (3)), as the mask probability.

$$\begin{aligned} &\mathrm{PLT}^c\left(X = s_{w_t} \mid S_{\setminus s_{w_{t' \geq t}}}\right) \\ &\qquad = 1 - \mathrm{PLT}\left(X = s_{w_t} \mid S_{\setminus s_{w_{t' \geq t}}}\right) \end{aligned} \tag{3}$$

The existing study has discussed that a significantly high mask probability for the MLM task can degrade the performance of downstream tasks (Wettig et al., 2023). Therefore, in this study, we define a modified PLT (mPLT) that is controlled to prevent the PLT^c used as the mask probability from

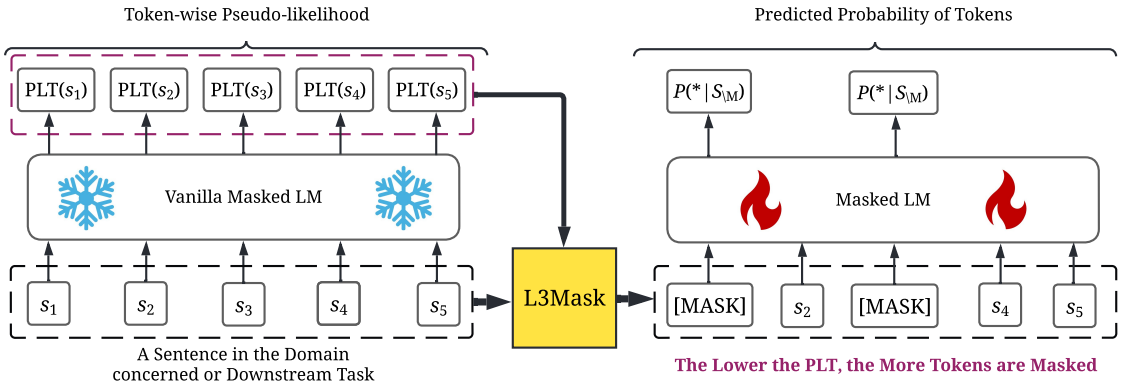

Figure 2: Overview of L3Masking. PLT denotes the token-wise pseudo-likelihood. $P(*|S_{\backslash M})$ represents the prediction probability for each token in the vocabulary of the language model at the masked position in a context $S_{\backslash M}$ excluding the masked token.

becoming too high. In particular, mPLT is calculated so that the mask probabilties in a sentence to a specified value $\bar{p}$. Formally, given a PLT^c sequence $P = (p_1, p_2, \ldots, p_n)$ corresponding with a n-length token sequence of a sentence and a specified average mask probability $\bar{p}$, find a constant α such that $\frac{1}{n}\sum_{p_i \in P} \alpha p_i = \bar{p}$. From this equation, $\alpha = \frac{n\bar{p}}{\sum_{p_i \in P} p_i}$ can be easily derived. By using this α, mPLT for each token t in a sentence S can be calculated as follows:

$$\begin{aligned} &\text{mPLT}\left(X = s_{w_t} \mid S_{\backslash s_{w_{t' \geq t}}}\right) \\ &\quad = \alpha \cdot \text{PLT}^c\left(X = s_{w_t} \mid S_{\backslash s_{w_{t' \geq t}}}\right) \end{aligned} \quad (4)$$

Note that in some cases when n is large and/or the summation of PLT^{c}s is too small (since PLT^c is token-wise probability, $\sum_{p_i \in P} p_i = 1$ does not hold.), it may theoretically happen mPLT values become more than 1. To assure mPLT to be probabilistic and $\bar{p}$ consistent, when an mPLT value exceeds 1, the exceeded value is equally distributed to all the other tokens in the sentence.

3.4 Mask Strategy

In the proposed masking strategy, the value `[MASK]` calculated by Equation (4) is the mask probability of the token in each sentence. The process for constructing the MLM task, such as replacing tokens and random tokens, follows the strategy in RoBERTa (Liu et al., 2019). The tokens to be manipulated are determined based on the probabilities calculated for each token. Of these, the token is replaced with the `[MASK]` token with a probability of 80%, and the token is replaced by a random token with the 10% probability, and 10% probability of leaving the token as is. Also, unlike the masking strategy of BERT, the `[MASK]` positions are re-calculated for each mini-batch.

4 Evaluation Experiment

To evaluate L3Masking, we conducted experiments to answer the following four questions:

- ***Q1***: *Does LM learn their own shortcomings from the vanilla model to improve their classification performance compared with the fine-tuned model of the vanilla model only on downstream tasks?*
- ***Q2***: *Does L3Masking improve classification performance by adaptively learning task- or domain-specific tokens effectively?*
- ***Q3***: *Does L3Masking mask task or domain-specific tokens more frequently than random token masking, and how does this change the model's adaptability?*
- ***Q4***: *Does L3Masking improve the efficiency of training by focusing more on masking important tokens in the data for the relevant domain and downstream tasks?*

4.1 Settings

Datasets and Metrics. In our experiments, we use three datasets ACL-ARC (Jurgens et al., 2018), Ohsumed (Hersh et al., 1994), and IMDb (Maas et al., 2011), to evaluate existing methods and our method. The basic statistics for each dataset are shown in Table 1. As for the text classification problem, we set the evaluation metrics as macro F_1 score and accuracy in the confusion matrix.

Table 1: Basic statistics of the three datasets used in the evaluation experiments. $|D_{\text{train}}|$, $|D_{\text{valid}}|$ and $|D_{\text{test}}|$ represent the numbers of instances in the training, validation, test data, respectively, and $|C|$ is the number of classes.

Domain	Task	Type of Supervised Label	$\lvert D_{\text{train}}\rvert$	$\lvert D_{\text{valid}}\rvert$	$\lvert D_{\text{test}}\rvert$	$\lvert C\rvert$
Computer Science	ACL-ARC (Jurgens et al., 2018)	citation intent	1,688	114	139	6
Medical	Ohsumed (Kringelum et al., 2016)	category classification	3,022	4,043	4,043	23
Movie Review	IMDb (Luan et al., 2018)	sentiment classification	25,000	2,500	22,500	2

Table 2: Experimental Settings

Parameter	Value
Optimizer	AdamW
Learning Rate	1e-4
Token Length	128
Batch Size	64
Dropout Rate	0.10
Average Mask Probability	0.15
Number of Epochs	150
Early Stopping Patience (epochs)	3

Comparison Methods. To demonstrate the usefulness of L3Masking, we implemented L3Masking into the multi-task learning text classification framework (MTL) of META-TARTAN[2] (Dery et al., 2022) instead of RTM for the MLM task. To answer Q1, L3Masking is compared with a simple fine-tuned model without any auxiliary task, and we call it STL (Single Task Learning). To show a comparison of the impact of MLM on classification performance due to different masking strategies in Q2 and Q3, RTM and L3Masking were used as auxiliary tasks in META-TARTAN framework, and we call MTL methods with these masking strategies as RTM and L3Masking for short, respectively. Note that MLM tasks, including L3Masking and RTM, were applied to data of the text classification task. To answer Q4, we recorded the number of training epochs of the META-TARTAN framework when using RTM or L3Masking, respectively.

Implementations. The hyper-parameters of META-TARTAN were set as Table 2, and the same hyper-parameters were used for L3Masking and the baseline methods. Our experimental evaluation selected the vanilla models pre-trained in the generic corpora, BERT-base[3] (Kenton and Toutanova, 2019) and RoBERTa-base[4] (Liu et al., 2019), as LM for META-TARTAN in our experimental evaluation to confirm task and domain adaptability. In addition, the vanilla models pre-trained on the dedicated domain corpora, SciBERT[5] (Beltagy et al., 2019) and ClinicalBERT[6] (Wang et al., 2023), were used to check task adaptability to the computer science domain (ACL-ARC task) and medical domain (Ohsumed task). To optimize task weights of META-TARTAN, objective metrics were aligned to the evaluation metrics (i.e., accuracy or macro F_1). For instance, when evaluating the performance of RTM or L3Masking by the accuracy metric, the task weights of META-TARTAN are optimized based on accuracy scores in the validation data.

4.2 Results

Table 3 showcases the results of this experiment. Overall, our method, L3Masking, demonstrated improvements across a range of datasets compared to the baseline methods. In particular, L3Masking performed superior or comparable to Baseline and RTL in ACL-ARC and Ohsumed, regardless of the language models. However, in general domain dataset IMDb, L3Masking and RTM showed superior performance to STL, while the gap between L3Masking and RTM are limited. This result indicates that advantages of L3Masking are more emphasized in domain-specific contexts.

On the ACL-ARC dataset, L3Masking showed varying degrees of improvement across different general domain LM compared to RTM; L3Masking on both BERT-base and RoBERTa-base showed improvements in the macro F_1 and the accuracy scores, especially RoBERTa-base benefited to a greater extent. In particular, L3Masking improved accuracy by 0.18 points and macro F_1 score by

[2] https://github.com/ldery/TARTAN/tree/main, accessed on October 13, 2024

[3] google-bert/bert-base-uncased, https://huggingface.co/google-bert/bert-base-uncased, accessed on October 13, 2024

[4] FacebookAI/roberta-base, https://huggingface.co/FacebookAI/roberta-base, accessed on October 13, 2024

[5] allenai/scibert_scivocab_uncased, https://huggingface.co/allenai/scibert_scivocab_uncased, accessed on October 13, 2024

[6] medicalai/ClinicalBERT, https://huggingface.co/medicalai/ClinicalBERT, accessed on October 13, 2024

Table 3: Comparison of Accuracy and Macro F_1 of text classification between STL, RTM, and L3Masking in percentages. The average values and standard deviations of 10 trials are reported. The highest average values for each language model and for each Accuracy and Macro F_1 is in **bold**.

Dataset		**ACL-ARC**		**Ohsumed**		**IMDb**	
Framework	**Masking**	*Acc*	F_1	*Acc*	F_1	*Acc*	F_1
(General Domain)		BERT-base (Kenton and Toutanova, 2019)					
STL	-	**71.34** $\pm$ 0.35	63.07 $\pm$ 0.69	76.69 $\pm$ 3.41	**68.76** $\pm$ 3.47	88.05 $\pm$ 0.05	87.15 $\pm$ 0.56
MTL	RTM	70.77 $\pm$ 0.86	62.15 $\pm$ 0.48	**76.98** $\pm$ 2.03	67.47 $\pm$ 2.40	88.05 $\pm$ 0.05	**88.19** $\pm$ 0.08
MTL	L3Masking	71.31 $\pm$ 0.98	**63.15** $\pm$ 0.90	76.81 $\pm$ 1.49	66.10 $\pm$ 3.50	**88.10** $\pm$ 0.21	88.08 $\pm$ 0.08
(General Domain)		RoBERTa-base (Liu et al., 2019)					
STL	-	71.73 $\pm$ 4.06	59.44 $\pm$ 6.70	70.07 $\pm$ 0.54	60.92 $\pm$ 0.91	88.84 $\pm$ 0.32	88.89 $\pm$ 0.30
MTL	RTM	78.94 $\pm$ 1.76	70.30 $\pm$ 2.20	69.92 $\pm$ 0.64	64.83 $\pm$ 0.37	91.29 $\pm$ 0.27	**91.30** $\pm$ 0.22
MTL	L3Masking	**79.12** $\pm$ 1.60	**73.30** $\pm$ 2.90	**73.38** $\pm$ 0.48	**65.02** $\pm$ 0.61	**91.32** $\pm$ 0.15	91.13 $\pm$ 0.09
(Domain-Specific)		SciBERT (Beltagy et al., 2019)		ClinicalBERT (Wang et al., 2023)			
STL	-	80.36 $\pm$ 2.45	71.84 $\pm$ 2.73	71.02 $\pm$ 0.42	62.85 $\pm$ 0.63	-	-
MTL	RTM	80.14 $\pm$ 1.38	70.88 $\pm$ 3.06	70.75 $\pm$ 0.36	62.70 $\pm$ 0.61	-	-
MTL	L3Masking	**82.50** $\pm$ 1.90	**74.10** $\pm$ 2.40	**71.66** $\pm$ 0.78	**63.70** $\pm$ 0.60	-	-

3.00 points in RoBERTa-base compared to RTM. However, in the BERT-base, L3Masking performed comparably to STL. The SciBERT model exhibited the most substantial improvement with L3Masking, achieving an accuracy of 82.50 and a macro F_1 score of 74.10, surpassing RTM by 2.36 points in accuracy and 3.22 points in the macro F_1 score.

On the Ohsumed dataset, L3Masking's classification performance varied. In the general domain model, the BERT-base was slightly lower than RTM in both the macro F_1 and accuracy scores. For BERT-base, accuracy was similar for STL, RTM, and L3Masking, and STL had the best macro F_1 score. However, for the RoBERTa-base, L3Masking performed better than STL and RTM, especially in accuracy, which was 3.46 points better than RTM. ClinicalBERT with L3Masking achieved an accuracy of 71.66 and F_1 score of 63.70, outperforming STL and RTM.

On the IMDb dataset, L3Masking's impact was generally limited across the general domain LM. For both BERT-base and RoBERTa-base, L3Masking did not show much difference from baseline or RTM. These results suggest that L3Masking's effect may be less pronounced in general domains such as movie reviews.

4.3 Analysis

In this section, we analyze the effectiveness and efficiency of the L3Masking by examining the types of tokens that were frequently masked and their impact on model performance. We also assess influences on the training process in terms of both accuracy and the number of epochs required.

Types of tokens masked by L3Masking. The L3Masking reveals significant insights into domain-specific adaptation by assigning higher masking probabilities to tokens that carry essential linguistic and domain-specific information. Tables 4 and 5 show the results of part-of-speech (POS) analysis on ACL-ARC training data conducted using NLTK[7] in Python, along with the average mask probability by L3Masking for each POS tag. It is important to note that while POS analysis is performed on a word-by-word basis, L3Masking assigns mask probabilities per token. Therefore, in this analysis, POS tags are assigned to each token, including subwords, and the results are then aggregated by POS tag.

As observed in Table 4, foreign words (FW) and plural nouns (NNS) exhibit the highest masking probabilities in both SciBERT and BERT models within the ACL-ARC dataset. This suggests that L3Masking effectively identifies tokens contributing to the domain's unique linguistic patterns, facilitating more effective knowledge transfer during fine-tuning.

In contrast, general grammatical tokens such as wh-pronouns (WP) and base form verbs (VB) consistently show lower masking probabilities (Table 5), indicating that these elements contribute less to domain-specific adaptations. This distinction underscores L3Masking's ability to prioritize

[7]Natural Language Toolkit (Version 3.8.1) , https://www.nltk.org/, accessed on October 13, 2024

Table 4: The top 5 POS tags with the highest masking probability for RoBERTa and SciBERT in the training data of the ACL-ARC dataset using L3Masking. The masking probability listed in the table is the average of the masking probability for each token. POS tags that occur less than 10 times have been removed from the table.

Rank	L3Masking (RoBERTa)			L3Masking (SciBERT)		
	POS Tag	Description	Mask Probability	POS Tag	Description	Mask Probability
1	FW	Foreign word	0.2023	POS	Possessive ending	0.3133
2	)	Closing parenthesis	0.1907	FW	Foreign word	0.2131
3	(	Opening parenthesis	0.1883	NNS	Noun, plural	0.2017
4	NNS	Noun, plural	0.1832	)	Closing parenthesis	0.2009
5	NNP	Proper noun, singular	0.1647	(	Opening parenthesis	0.1737

Table 5: Worst 5 POS tags with lowest masking probability for RoBERTa and SciBERT in training data of ACL-ARC dataset using L3Masking.

Rank	L3Masking (RoBERTa)			L3Masking (SciBERT)		
	POS Tag	Description	Mask Probability	POS Tag	Description	Mask Probability
1	WP	Wh-pronoun	0.0265	JJS	Adjective, superlative	0.0267
2	VB	Verb, base form	0.0465	WP	Wh-pronoun	0.0281
3	.	Punctuation mark	0.0633	EX	Existential there	0.0317
4	CD	Cardinal number	0.0664	CD	Cardinal number	0.0492
5	VBN	Verb, past participle	0.0686	RBS	Adverb, superlative	0.0503

learning relevant language patterns while minimizing the focus on general linguistic features.

Efficiency. We also found that L3Masking is not only effective for the META-TARTAN framework, but also efficient. Figure 3 illustrates the differences in the number of training epochs and accuracy between RTM and L3Masking across BERT, RoBERTa, SciBERT, and ClinicalBERT. As shown in Figure 3, L3Masking applied to BERT and RoBERTa achieved superior or comparable accuracy in fewer epochs on average than RTM, reducing training time while maintaining or enhancing model performance. This efficiency is particularly advantageous for language models trained on general domain documents, such as BERT and RoBERTa, where computational resources and time are often constrained.

In contrast, while L3Masking in SciBERT and ClinicalBERT improved classification performance over RTM, it did not reduce the number of epochs required. This discrepancy can be attributed to the inherent nature of domain-specific LMs like SciBERT and ClinicalBERT, which are already finely tuned to their respective domains during pretraining. As a result, these models benefit more from L3Masking's ability to refine domain-specific knowledge, leading to improved accuracy. However, because these models are already adapted to their domains, the room for efficiency gains in terms of reduced training time is limited.

These results indicate that L3Masking can effectively decrease training time for models based on generic corpora, like BERT and RoBERTa. However, for models like SciBERT and ClinicalBERT, which are trained on specialized domains, L3Masking primarily enhances task performance without reducing training duration.

4.4 Lessons Learned

As shown in the experimental results above, L3Masking's ability to selectively mask task- or domain-specific tokens significantly enhances the model's performance and adaptability in text classification, confirming its effectiveness over RTM in this context. In summary, questions raised in this section are answered in the rest of this section.

Q1 — Yes, language models (LMs) that learn about their own shortcomings (lessons) demonstrate better classification performance than those that only focus on downstream tasks. Specifically, using L3Masking, models actively learn domain-specific knowledge essential for downstream tasks by focusing on tokens with low pseudo-likelihood and masking them. This approach strengthens areas where the model is underperforming, enabling it to effectively apply learned knowledge. The method helps bridge the distributional differences between

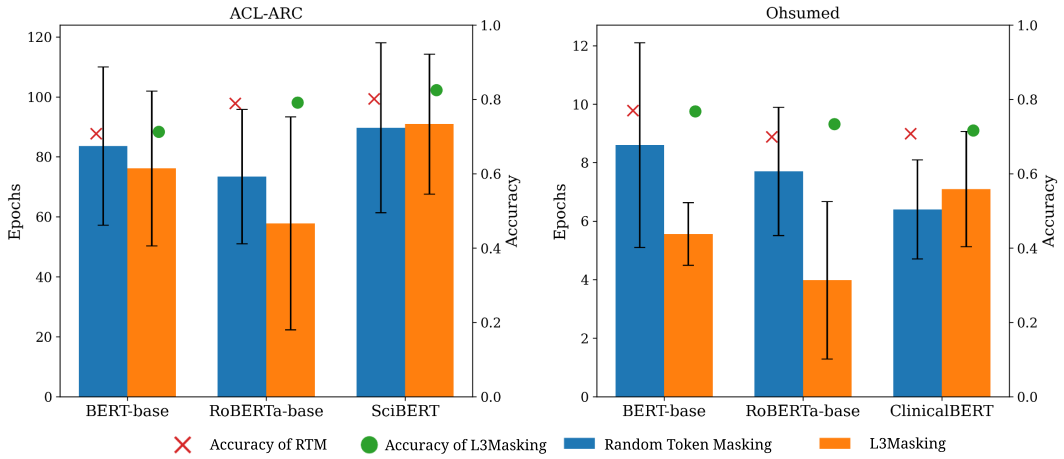

Figure 3: Difference in the number of training epochs for ACL-ARC and Ohsumed using RTM or L3masking for each language model. The plot is based on the average and standard deviation of 10 experiments.

pre-training and fine-tuning tasks, thereby enhancing task adaptability. It has been shown that such learning leads to improved classification performance, particularly in specialized domains.

Q2 — Yes, L3Masking adapts more effectively than Random Token Masking (RTM) and improves classification performance. Notably, in domain-specific language models such as SciBERT and ClinicalBERT, L3Masking demonstrates superior accuracy and F_1 scores compared to RTM. L3Masking identifies and prioritizes tokens specific to the task or domain, leading to more effective task adaptation. Unlike RTM, where tokens are treated uniformly, L3Masking overcomes this limitation by promoting the learning of language patterns relevant to the task. This targeted masking strategy enhances the model's understanding and application of domain-specific knowledge.

Q3 — Yes, L3Masking masks task or domain-specific tokens more frequently than RTM, significantly enhancing the model's adaptability in text classification tasks. By prioritizing the masking of tokens such as foreign words (FW) and plural nouns (NNS), which are crucial in domain-specific contexts like those found in the ACL-ARC and Ohsumed datasets, L3Masking facilitates a deeper understanding of domain-specific language patterns. This strategic focus enables the model to capture better essential linguistic features required for accurate domain-specific classification.

Moreover, this targeted approach enhances the model's adaptability by allowing it to concentrate on tokens that carry significant domain-specific information. As a result, models equipped with L3Masking outperform those using RTM in terms of performance metrics, particularly in domain-specific classification tasks.

Q4 — Yes, L3Masking improves the efficiency of training by strategically focusing on masking important tokens that are crucial for the relevant domain and downstream tasks. By prioritizing task- and domain-specific tokens during the masking process, L3Masking enables the model to concentrate its learning on the most relevant and informative aspects of the data. This targeted approach leads to a reduction in the number of training epochs required to achieve comparable or superior accuracy, particularly in general-domain models like BERT and RoBERTa.

5 Conclusion

This paper introduced L3Masking as a novel masking strategy for fine-tuning of Masked Language Models to text classification. Our method leverages likelihood scores from the vanilla models to actively mask task- or domain-specific tokens. For calculating mask probability on the bidirectional MLMs, token-by-token pseudo-likelihood scores are used. Our method focuses more on tokens that are underrepresented in the pre-training corpus but are crucial for downstream tasks. Through the experimental evaluation of three text classification tasks from different domains, we demonstrated that L3Masking outperforms traditional random token masking, particularly in domain-specific language models such as SciBERT and ClinicalBERT.

Future work will focus on refining the token selection algorithm to handle diverse datasets better

and exploring L3Masking's potential in other NLP tasks beyond text classification. Additionally, applying L3Masking to the continual pre-training of large language models (LLMs) represents a significant future direction. By leveraging L3Masking in LLMs, we aim to achieve more accurate domain adaptation, task-specific learning, and effective utilization of large-scale datasets, ultimately enhancing LLMs' overall performance and applicability in various specialized and general domains.

Limitations

Despite the promising results, our study has several limitations. Firstly, our experiments were primarily focused on text classification tasks. Although these tasks provide a good benchmark for evaluating multi-task classification methods, it remains to be unveiled how L3Masking performs in other NLP tasks, such as named entity recognition, machine translation, or text generation. Future research should extend the evaluation of L3Masking to a wider range of tasks to fully understand its capabilities and limitations.

Secondly, the computational overhead associated with calculating token-by-token pseudo-likelihood scores can be substantial. However, we emphasize that the calculation of the mask probability for L3Masking only needs to be performed once per dataset. Although L3Masking can still be computationally expensive, the results presented in this paper suggest that it is worth considering as a replacement for random token masking as an auxiliary task.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, and Others. 2023. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3615–3620. Association for Computational Linguistics.

Lucio M. Dery, Paul Michel, Mikhail Khodak, Graham Neubig, and Ameet Talwalkar. 2023. AANG: Automating auxiliary learning. In *The Eleventh International Conference on Learning Representations*.

Lucio M. Dery, Paul Michel, Ameet Talwalkar, and Graham Neubig. 2022. Should We Be Pre-training? An Argument for End-task Aware Training as an Alternative. In *The Tenth International Conference on Learning Representations*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

William Hersh, Chris Buckley, T. J. Leone, and David Hickam. 1994. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, page 192–201, Berlin, Heidelberg. Springer-Verlag.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Carina Kauf and Anna Ivanova. 2023. A Better Way to Do Masked Language Model Scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yusuke Kimura, Takahiro Komamizu, and Kenji Hatano. 2023. An automatic labeling method for subword-phrase recognition in effective text classification. *Informatica (Slovenia)*, 47(3):315–326.

Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. 2016. ChemProt-3.0: a global chemical biology diseases mapping. *Database: The Journal of Biological Databases and Curation*, 2016.

Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. 2021. PMI-MASKING: PRINCIPLED MASKING OF CORRELATED SPANS. In *The Ninth International Conference on Learning Representations*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,

Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Nafis Sadeq, Canwen Xu, and Julian McAuley. 2022. InforMask: Unsupervised Informative Masking for Language Model Pretraining. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5866–5878, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712. Association for Computational Linguistics.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *Preprint*, arXiv:2404.16789.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced Representation through Knowledge Integration. *Preprint*, arXiv:1904.09223.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, Kanmin Xue, Xiaoying Li, and Ying Chen. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should You Mask 15% in Masked Language Modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000. Association for Computational Linguistics.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2023. Efficient continual pre-training for building domain specific large language models. *Preprint*, arXiv:2311.08545.