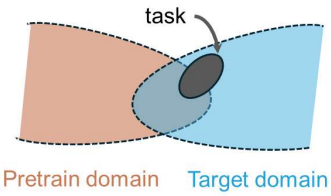
# L3Masking: Multi-task Fine-tuning for Language Models by Leveraging Lessons Learned from Vanilla Models

Yusuke Kimura†, Takahiro Komamizu‡, and Kenji Hatano†
†Doshisha University, ‡Nagoya University, Japan
usk@acm.org

## Gap of pretraining and downstream tasks

Distribution Shift:
domain shift (right fig*), time shift, task shift,…



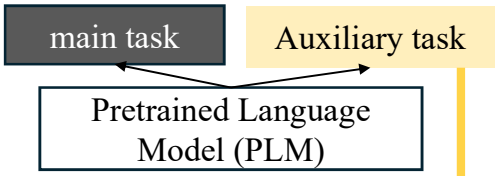

*This figure was created with reference to the following paper.
Gururangan et al. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *In Proceedings of the ACL*, pages 8342–8360.

## Overcoming the gap through Multi-task Fine-tuning

Multi-task fine-tuning (MTL) [1]

- Methods to improve the target task (main task)
- Pretraining tasks are used for auxiliary tasks

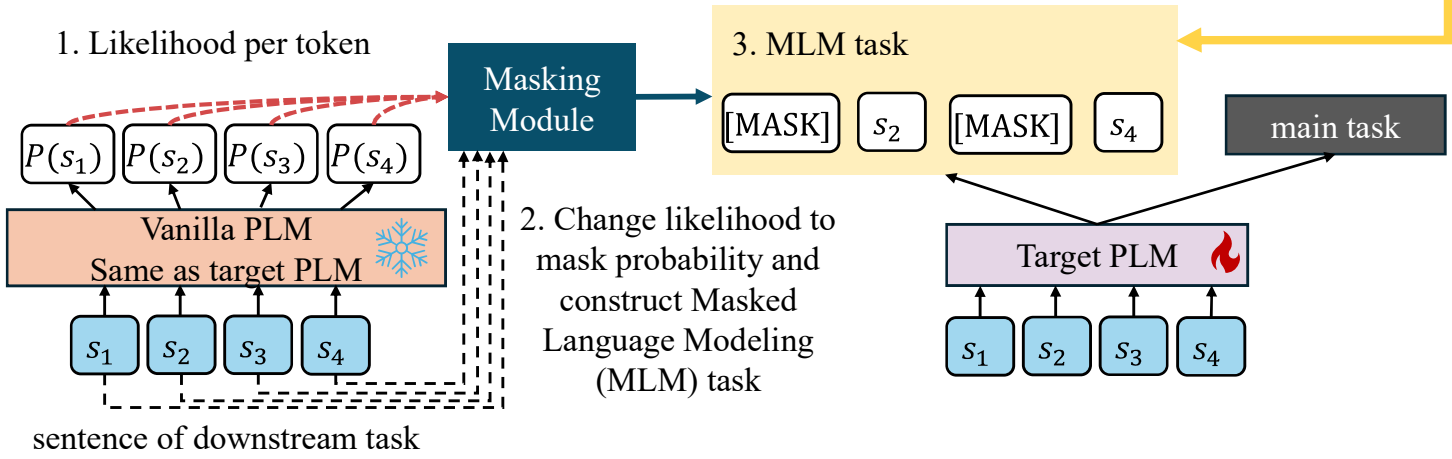
**Key of MTL: Design of auxiliary tasks**




## L3Masking

### Designing an auxiliary task leveraging lessons learned from vanilla models

Idea: The lower likelihood in the vanilla model, the more masking, vice versa




## Evaluation (Text Classification Task)

| Dataset | | ACL-ARC (Computer Science) | | Ohsumed (Medical) | | IMDb (Movie Review) | |
|---|---|---|---|---|---|---|---|
| **Framework** | **Masking** | *Acc* | macro $F_1$ | *Acc* | macro $F_1$ | *Acc* | macro $F_1$ |
| (General Domain) | | RoBERTa-base | | | | | |
| STL | - | 71.73 ± 4.06 | 59.44 ± 6.70 | 70.07 ± 0.54 | 60.92 ± 0.91 | 88.84 ± 0.32 | 88.89 ± 0.30 |
| MTL | RTM | 78.94 ± 1.76 | 70.30 ± 2.20 | 69.92 ± 0.64 | 64.83 ± 0.37 | 91.29 ± 0.27 | **91.30** ± 0.22 |
| MTL | L3Masking | **79.12** ± 1.60 | **73.30** ± 2.90 | **73.38** ± 0.48 | **65.02** ± 0.61 | **91.32** ± 0.15 | 91.13 ± 0.09 |
| (Domain-Specific) | | SciBERT | | ClinicalBERT | | | |
| STL | - | 80.36 ± 2.45 | 71.84 ± 2.73 | 71.02 ± 0.42 | 62.85 ± 0.63 | - | - |
| MTL | RTM | 80.14 ± 1.38 | 70.88 ± 3.06 | 70.75 ± 0.36 | 62.70 ± 0.61 | - | - |
| MTL | L3Masking | **82.50** ± 1.90 | **74.10** ± 2.40 | **71.66** ± 0.78 | **63.70** ± 0.60 | - | - |

STL; Single Task Learning, RTM; Random Token Masking

[1] 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*.