Punctuation Restoration for Linguistic Structure Learning

Michael Zhou[z] **Markus Frohmann**[f] **Junghyun Min**[m]
[z]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[f]Johannes Kepler University, Linz, Austria
[m]Georgetown University, Washington, DC, USA
[z]mjzhou@andrew.cmu.edu

Introduction. While modern language models (LMs) already perform well in most NLP tasks, there is room for improvement in understanding structure in text, as LMs trained on token prediction do not fully capture the hierarchical structure present in linguistic input (Manikandan et al., 2023). Punctuation often encodes vital structural boundaries in discourse, syntax, and meaning (Dale, 1991), and can serve as implicit sturctural cues in LM pre-training. Previous work has shown that additional representation learning via continual pre-training on punctuation restoration (PR) improves language models' performance in structure-related NLP tasks like information extraction (OIE), named entity recognition (NER), and semantic role labeling (Min et al., 2025).

However, the effects of PR are not fully understood, as previous work only investigates T5 (Raffel et al., 2020) on structure-related tasks. In addition, there may be unintended effects of extended pre-training, such as overtraining (Springer et al., 2025) that may lead to degradation in future task-specific fine-tuning, and model misalignment (Betley et al., 2026). In this paper, we perform a wide-scale, systematic evaluation into the effects of additional pre-training on PR for structure learning and natural language understanding. Working with English, we incorporate additional model architectures, a wider range of evaluation tasks, and systematic hyperparameter ablation.

Methods. In PR training, given a sentence stripped of casing and punctuation mark (e.g. `lets eat grandma`), models learn to predict the fully punctuated and cased sentence (e.g. `Let's eat, Grandma!`). To study its effects, we experiment with three transformer models: BERT (Devlin et al., 2019), GPT2 (Radford et al., 2019), and T5, representing decoder-only, encoder-only, and encoder-decoder architectures, respectively. Off-the-shelf models representing each architecture is continually pre-trained on punctuation restoration. We compare the resulting 3 models to their off-the-shelf counterparts (baselines) by measuring their performance on structure-related and general langauge understanding downstream tasks after task-specific fine-tuning. We evaluate our models on GLUE (Wang et al., 2018) in addition to those employed in Min et al. (2025), for a total of 10 tasks.

Results. Our initial results in table 1 show that GPT2 benefits the most from PR, and BERT the least. For all 3 architectures, PR yields improvement in Winograd natural language inference (WNLI) task, which measures a model's ability to resolve the referent of a pronoun in an NLI setting and requires an understanding of sentence structure and coreference. In contrast, additional training on PR hurts performance on STS-B, a task focusing on semantic textual similarity, indicating that PR may distract models from semantic information. We also find that gradient accumulation is a crucial hyperparameter due to the sparsity of signal in PR; only a few punctuation marks and uppercase letters exist in each sentence. We plan to continue our analysis across model size, investigating whether PR benefits larger models too.

Task	GPT2 Δ	T5 Δ	BERT Δ
SST-2	▼ 1.8	▲ 1.2	▼ 0.7
QQP	▲ 0.2	0.0	▼ 2.8
QNLI	▼ 0.8	▼ 2.7	▼ 3.9
RTE	▲ 1.1	▼ 20.1	▼ 5.4
WNLI	▲ 2.8	▲ 4.6	▲ 7.0
CoLA	▲ 11.1	▲ 0.5	▼ 7.3
MRPC	▲ 0.2	▼ 5.9	▼ 6.4
STS-B	▼ 0.5	▼ 10.0	▼ 6.7
NER	▲ 1.8	▲ 0.4	▼ 17.2
OIE	▲ 1.3	▲ 0.9	0.0

Table 1: Difference in performance with and without additional training on PR, measured on GLUE tasks (Wang et al., 2018), NER (Sang and De Meulder, 2003) and OIE (Xue et al., 2016).

References

Jan Betley, Niels Warncke, Anna Sztyber-Betley, Daniel Tan, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. 2026. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649(8097):584–589.

Robert Dale. 1991. Exploring the role of punctuation in the signalling of discourse structure. In *Proceedings of a workshop on text representation and domain modelling: ideas from linguistics and AI*, pages 110–120. Technical University of Berlin Berlin, Germany.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and 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.

Hariharan Manikandan, Yiding Jiang, and J Zico Kolter. 2023. Language models are weak learners. *Advances in Neural Information Processing Systems*, 36:50907–50931.

Junghyun Min, Minho Lee, Lee Woochul, and Yeonsoo Lee. 2025. Punctuation restoration improves structure understanding without supervision. In *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, pages 120–130.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 142–147.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. 2025. Overtrained language models are harder to fine-tune. *Preprint*, arXiv:2503.19206.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.