Hyperbolic Representations for Prompt Learning

Nan Chen[1,2,3], **Xiangdong Su**[1,2,3], **Feilong Bao**[1,2,3](✉)
[1]College of Computer Science, Inner Mongolia University, Hohhot, China
[2]National & Local Joint Engineering Research Center of
Intelligent Information Processing Technology for Mongolian
[3]Inner Mongolia Key Laboratory of Mongolian Information Processing Technology
chennannlp@gmail.com, cssxd@imu.edu.cn, csfeilong@imu.edu.cn

Abstract

Continuous prompt tuning has gained significant attention for its ability to train only continuous prompts while freezing the language model. This approach greatly reduces the training time and storage for downstream tasks. In this work, we delve into the hierarchical relationship between the prompts and downstream text inputs. In prompt learning, the prefix prompt acts as a module to guide the downstream language model, establishing a hierarchical relationship between the prefix prompt and subsequent inputs. Furthermore, we explore the benefits of leveraging hyperbolic space for modeling hierarchical structures. We project representations of pre-trained models from Euclidean space into hyperbolic space using the Poincaré disk which effectively captures the hierarchical relationship between the prompt and input text. The experiments on natural language understanding (NLU) tasks illustrate that hyperbolic space can model the hierarchical relationship between prompt and text input. We release our code at `https://github.com/myaxxxxx/Hyperbolic-Prompt-Learning`.

Keywords: prompt learning, hierarchical structure, pre-trained model

1. Introduction

The remarkable achievements of pre-trained large language models (Devlin et al., 2018; Yang et al., 2019; Raffel et al., 2019) empower the development of lightweight fine-tuning techniques (Li and Liang, 2021; Qin and Eisner, 2021) for downstream tasks. In lightweight fine-tuning, the majority of the pre-trained model is frozen, while small trainable modules are available. This strategy effectively addresses the issue of catastrophic forgetting (Houlsby et al., 2019) often encountered during full parameter training. Therefore, lightweight fine-tuning not only achieves remarkable results but also significantly reduces training costs.

There are multiple avenues for exploring lightweight fine-tuning techniques (Rebuffi et al., 2017; Li and Liang, 2021; Chen et al., 2021; Zheng et al., 2021). One such approach is prompt learning (Shin et al., 2020; Ding et al., 2021) which freezes all parameters of the pre-trained model and utilizes natural language prompts to query a language model. There are two types of prompt tokens, including discrete prompts (Gao et al., 2020; Shin et al., 2020) and continuous prompts. Typically, discrete prompts are composed of a task description and/or a series of canonical examples. However, the use of discrete prompts can result in suboptimal performance in many cases compared to fine-tuning. Different from discrete prompts (Shin et al., 2020; Gao et al., 2020), continuous prompts tuning (Liu et al., 2021b; Li and Liang, 2021; Zhong et al., 2021) is an idea to tune continuous representations, it only updates continuous prompts parameters during training. Properly optimized continuous prompt tuning can be comparable to fine-tuning universally across various NLU tasks (Zhong et al., 2021; Lester et al., 2021).

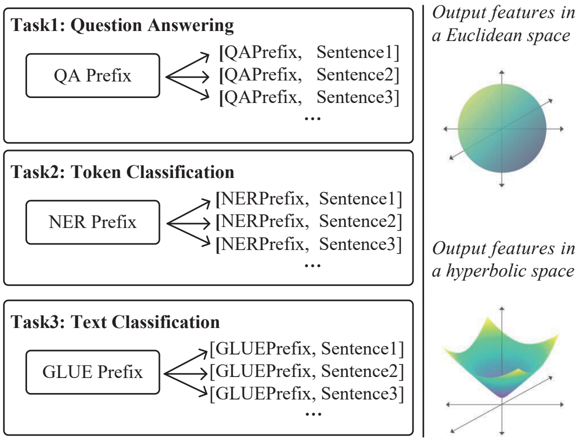

Figure 1: Hierarchical structure of prompt learning (Left). Euclidean space and hyperbolic space (Right).

This work is supported by National Natural Science Foundation of China (62066033); Outstanding Youth Project of Inner Mongolia Natural Science Foundation (2022JQ05); Young science and technology talents cultivation project of Inner Mongolia University (21221505); Applied Technology Research and Development Program of Inner Mongolia Autonomous Region(2021GG0158).

LREC-COLING 2024, pages 8487–8492
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

In prompt learning, the prefix prompt is closely related to the subsequent text inputs. It guides the downstream language model, which can be modeled as a tree relationship: prompt as the root node and subsequent inputs as the leaf node. As shown in Fig. 1(left), the class number of leaf nodes surpasses the count of root nodes and this type of tree can be viewed as a hierarchical structure. This observation motivates us to reconsider the relationship between prompt and subsequent text inputs from a hierarchical perspective.

In this paper, our focus is on exploring the hierarchical relationship between prompts and subsequent text inputs. Traditionally, prompts and text inputs were represented in Euclidean space. However, inspired by Desai et al. (2023); Ge et al. (2022); Atigh et al. (2022); Gulcehre et al. (2018), we propose leveraging the Poincaré disk to project prompts and text inputs from Euclidean space into hyperbolic space. By doing so, we effectively capture the hierarchical relationship that exists between them. We conducted comprehensive experiments on several natural language understanding (NLU) tasks, including question answering, named entity recognition, and sentence classification. The experimental results consistently demonstrate that prompt tuning in hyperbolic space enhances the performance across all NLU tasks and this observation highlights the generality and effectiveness of our approach.

Overall, our main contributions can be summarized as follows:

- We investigate the hierarchical structure between prompts and downstream task inputs, and propose the utilization of the Poincaré disk hyperbolic space to model and substantiate this relationship.
- Experiments on sentence classification, question answering, and token classification tasks demonstrate the effectiveness of our proposed approach.

2. Approach

In section 2.1, we briefly introduce the background of the Poincaré ball model. In section 2.2, we describe our proposed hyperbolic approach in detail.

2.1. Background: The Poincaré Ball Model

There are various isometric models of hyperbolic space. The n-dimensional hyperbolic space $\mathbb{H}^n$ is a Riemannian manifold of constant negative curvature. We select the Poincaré ball model ($\mathbb{D}_c^n$, $g^{\mathbb{D}}$) with the curvature parameter c.

This model is realized as a pair of an n-dimensional ball ($\mathbb{D}^n = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|^2 < 1, c \geq 0\}$) with the Riemannian metric $g^{\mathbb{D}}$ = $\lambda_c^2 g^E$. $\lambda_c = \frac{2}{1-c\|\boldsymbol{x}\|^2}$ is the conformal factor and $g^E = \mathbf{I}_n$ is Euclidean metric tensor which means local distances are scaled by the factor λ_c approaching infinity near the boundary of the ball. So, the *space expansion* property emerges naturally in hyperbolic spaces.

In the Euclidean spaces, however, the volume of an object of a diameter r scales polynomially in r, in the hyperbolic space, such volumes scale exponentially with r. Intuitively, We regard it as a continuous analogue of trees with a branching factor m. and $O\left(m^n\right)$ nodes on the level n, which in this case serves as a discrete analogue of the radius. Similar to hierarchical data, prefix prompts, and the subsequent text inputs also form a tree structure in which continuous prefix prompts prepend different subsequent inputs. This property allows hyperbolic space to efficiently model prompt tuning tasks.

2.2. Hyperbolic Representations of Pre-trained Model

We first describe the Poincar ball model projection. To map from the Euclidean tangent space to the hyperbolic space, networks operate on the Poincaré ball. The projection of a Euclidean vector $\boldsymbol{x}$ onto the Poincaré ball is given by the exponential map with anchor v:

$$\exp_v^c(\boldsymbol{x}) = v \oplus_c \left(\tanh\left(\sqrt{c}\frac{\lambda_v^c\|\boldsymbol{x}\|}{2}\right)\frac{\boldsymbol{x}}{\sqrt{c}\|\boldsymbol{x}\|}\right) \quad (1)$$

with $\oplus_c$ the Möbius addition:

$$v \oplus_c w = \frac{\left(1+2c\langle v, w\rangle+c\|w\|^2\right)v+\left(1-c\|v\|^2\right)w}{1+2c\langle v, w\rangle+c^2\|v\|^2\|w\|^2} \quad (2)$$

In practice, v is commonly set to the origin, simplifying the exponential map to:

$$\exp_0(\boldsymbol{x}) = \tanh((\sqrt{c}\|\boldsymbol{x}\|)(\boldsymbol{x}/(\sqrt{c}\|\boldsymbol{x})) \quad (3)$$

As shown in Fig. 2, we take BERT$_{large}$ model as an example. Given the trainable continuous embeddings [$\boldsymbol{p_1}$, $\boldsymbol{p_2}$,..., $\boldsymbol{p_n}$] as prefix representations, the prompt representation and input text are fed into the BERT$_{large}$ model:

$$\text{outputs}_{feature} = \text{BERT}_{large}([\boldsymbol{P_e}; \boldsymbol{X_e}]) \quad (4)$$

The output of the pre-trained model is to project to hyperbolic space:

$$\text{outputs}_{hy} = \exp_0(\text{outputs}_{features}) \quad (5)$$

Finally, according to per-task-specific settings, the outputs of hyperbolic space are fed into a linear classifier to get final logits:

$$\text{logits} = \text{BERT}_{linear}(\text{outputs}_{hy}) \quad (6)$$

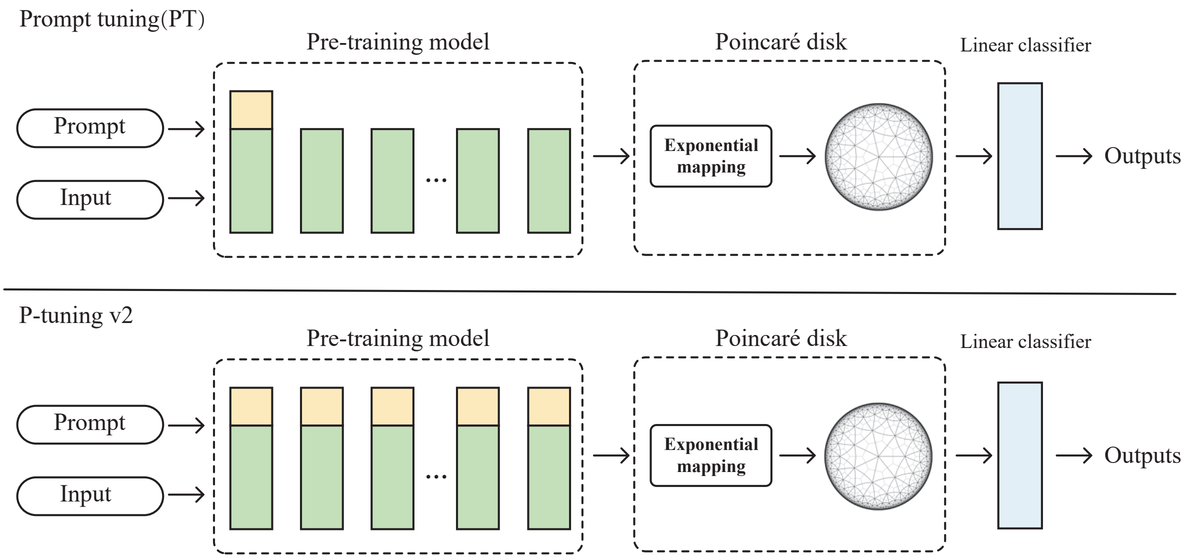

Figure 2: Overview of our proposed approach. Yellow blocks refer to trainable prompt embeddings. Green blocks are frozen pre-trained language models.

	BoolQ		CB		COPA		MultiRC (F1a)	
	PT	PT-2	PT	PT-2	PT	PT-2	PT	PT-2
BERT$_{large}$	63.2	75.2	73.2	94.6	62.0	75.0	**59.6**	**70.6**
BERT$_{large}$ + Poincaré	**66.7**	**76.0**	**75.0**	**96.0**	**71.0**	**79.0**	59.0	70.2
RoBERTa$_{large}$	62.2	**84.5**	69.6	94.6	63.0	88.2	**59.9**	**82.5**
RoBERTa$_{large}$ + Poincaré	**62.6**	**84.5**	**73.2**	**96.4**	**65.0**	**91.0**	59.3	82.0
	ReCoRD (F1)		RTE		WiC		WSC	
	PT	PT-2	PT	PT-2	PT	PT-2	PT	PT-2
BERT$_{large}$	**44.2**	**72.8**	53.5	**78.3**	56.9	71.0	**63.5**	66.3
BERT$_{large}$ + Poincaré	44.0	72.6	**65.0**	77.6	**65.1**	**73.2**	**63.5**	**67.3**
RoBERTa$_{large}$	**46.3**	**89.3**	54.5	87.0	57.8	69.0	**63.5**	**63.4**
RoBERTa$_{large}$ + Poincaré	**46.3**	89.0	**57.4**	**88.4**	**71.3**	**71.0**	**63.5**	**63.4**

Table 1: Results on SuperGLUE development set. (PT: Prompt tuning Lester et al. (2021); PT-2: P-tuning v2 Liu et al. (2021a); **bold**: the best).

3. Experiments

We conduct comprehensive experiments on various widely used pre-trained models and natural language understanding (NLU) tasks to evaluate the efficacy of hyperbolic representations in prompt learning. Except for fine-tuning, all methods are implemented with frozen language model backbones, following the experimental setup of (Liu et al., 2021a).

Tasks. We employ datasets from SuperGLUE (Wang et al., 2019) to evaluate overall NLU task performance. Furthermore, we introduced a set of sequence labeling tasks, such as named entity recognition (Sang and De Meulder, 2003; Weischedel et al., 2013) and extractive question answering (Rajpurkar et al., 2016).

Baselines. In our experiments, we employed the BERT$_{large}$ (Devlin et al., 2018) and RoBERTa$_{large}$ (Liu et al., 2019) pre-trained models as backbone language models (LM). We integrated the Poincaré sphere model into the following prompting methods:

- Prompt tuning (Lester et al., 2021): This method involves the addition of virtual tokens exclusively at the embedding layer of pre-trained language models.

- Prefix-tuning (Liu et al., 2021a): This method utilizes deep continuous prompts by inserting virtual tokens at the beginning of all key-value pairs within the attention layers of pre-trained language models.

	CoNLL03		CoNLL04		SQuAD 1.1		SQuAD 2.0	
	PT	PT-2	PT	PT-2	PT	PT-2	PT	PT-2
BERT	82.5	82.2	71.2	82.2	63.0/75.3	**82.1**/**89.4**	50.8/52.6	67.6/71.4
+ Poincaré	**84.2**	**83.4**	**72.8**	**84.1**	**62.8/75.0**	82.0/89.3	**51.0/53.0**	**68.2/73.0**
RoBERTa	87.1	86.9	76.2	86.2	72.5/78.4	88.0/94.0	67.5/71.4	81.1/84.5
+ Poincaré	**88.8**	**92.1**	**78.2**	**89.0**	**73.0/78.6**	**88.0/94.2**	**68.6/72.3**	**81.5/85.0**

Table 2: Results on named entity recognition (NER) and question answering (QA). (PT: Prompt tuning Lester et al. (2021); PT-2: P-tuning v2 Liu et al. (2021a); **bold**: the best).

Implementations. All our models are trained on a single NVIDIA A100 Tensor Core GPU. We implement prompt tuning and prefix-tuning following (Liu et al., 2021a) settings. Other hyperparameter settings can be found in our code.

3.1. Main Results

Table 1 presents the performances of hyperbolic representations in the SuperGLUE task. With the exception of the MultiRC and ReCoRD tasks, our approach exhibits better performance compared to the existing PT and PT2 baselines. Particularly in the RTE task, the PT-based hyperbolic BERT model exhibits remarkable superiority over the corresponding baselines.

Table 2 presents the performance of the named entity recognition (NER) and question answering (QA) tasks, respectively. It is observed that the NER experiments demonstrate significant improvements compared to the QA experiments which highlights the beneficial role of hyperbolic representations in effectively modeling token classification tasks.

3.2. Ablation Study & Inference Speed

We perform ablation experiments on four tasks. In Fig. 3, we show the effect of different curvatures for low and high-resource tasks. For high-resource tasks, we can observe that the effect of the curvature value is negligible, with only minor changes in performance even for large curvature differences (e.g., 1 to 4). A different observation can be made with low-resource tasks, we see a significant performance fluctuation for different curvatures. which suggests that the hyperparameter c is particularly sensitive to low-resource tasks.

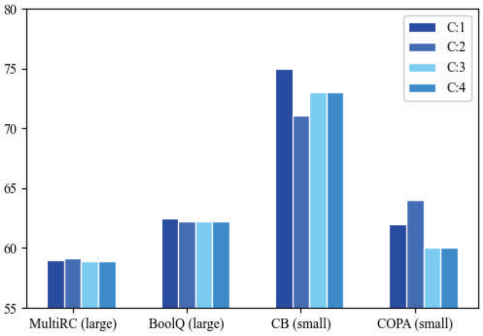

Figure 3: Comparison of curvature for high and low resources datasets. C:1 represents the curvature is set to 1.

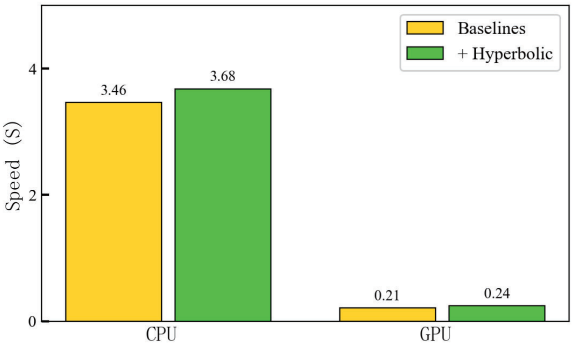

Figure 4: Comparison of inference speed for named entity recognition task (CoNLL 2003 dataset).

Based on the findings depicted in Fig. 4, we conduct an analysis of the inference time between employing Poincaré model and abstaining from its usage. The experiments are conducted using a single A100 for GPU inference. The results reveal that the incorporation of the Poincaré model led to a marginal increase of 0.03 seconds in the inference time, compared to the absence of the Poincaré model. Similarly, there is an observed increase of 0.24 seconds for CPU inference. These experimental outcomes indicate that the utilization of the Poincaré model has a negligible impact on the inference time.

4. Conclusion

In this paper, we explore the Poincaré disk hyperbolic representations of pre-trained models in NLU tasks, projecting representations from Euclidean space into hyperbolic space to model the hierarchical relationship between the prompt and input text. With high accuracy and efficiency, hyperbolic representations can be an effective supplement to prompt learning.

5. References

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. 2022. Hyperbolic Image Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4443–4452, New Orleans, LA, USA. IEEE.

Xiang Chen, Xin Xie, Ningyu Zhang, Jiahuan Yan, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021. Adaprompt: Adaptive prompt-based finetuning for relation extraction. *arXiv preprint arXiv:2104.07650*.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. 2023. Hyperbolic Image-Text Representations. ArXiv:2304.09172 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv e-prints*.

Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Hai-Tao Zheng, Zhiyuan Liu, Juanzi Li, and Hong-Gee Kim. 2021. Prompt-learning for fine-grained entity typing. *arXiv preprint arXiv:2108.10604*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. 2022. Hyperbolic Contrastive Learning for Visual Representations beyond Objects. ArXiv:2212.00653 [cs].

Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, and Nando de Freitas. 2018. Hyperbolic Attention Networks. ArXiv:1805.09786 [cs].

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021a. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *arXiv:2103.10385*.

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. *arXiv e-prints*.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, volume 30, pages 506–516. Curran Associates, Inc.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *NeurIPS 2019*, pages 3261–3275.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yanan Zheng, Jing Zhou, Yujie Qian, Ming Ding, Jian Li, Ruslan Salakhutdinov, Jie Tang, Sebastian Ruder, and Zhilin Yang. 2021. Fewnlu: Benchmarking state-of-the-art methods for few-shot natural language understanding. *arXiv preprint arXiv:2109.12742*.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [mask]: Learning vs. learning to recall. *arXiv preprint arXiv:2104.05240*.