Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. Udapter: Language adaptation for truly universal dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2302–2315. Association for Computational Linguistics.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. 2023. Skywork: A more open bilingual foundation model. *CoRR*, abs/2310.19341.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. *CoRR*, abs/2109.07684.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 833–844. Association for Computational Linguistics.

Ningyu Xu, Tao Gui, Ruotian Ma, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2022. Cross-linguistic syntactic difference in multilingual BERT: how good is it and how does it affect transfer? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8073–8092. Association for Computational Linguistics.

Ningyu Xu, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2023. Are structural concepts universal in transformer language models? towards interpretable cross-lingual generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13951–13976. Association for Computational Linguistics.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *CoRR*, abs/2306.06688.

Ofir Zafrir, Ariel Larey, Guy Boudoukh, Haihao Shen, and Moshe Wasserblat. 2021. Prune once for all: Sparse pre-trained language models. *CoRR*, abs/2111.05754.

Qingru Zhang, Simiao Zuo, Chen Liang, Alexander Bukharin, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2022. PLATON: pruning large transformer models with upper confidence bound of weight importance. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 26809–26823. PMLR.

Zhong Zhang, Bang Liu, and Junming Shao. 2023. Fine-tuning happens in tiny subspaces: Exploring intrinsic task-specific subspaces of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1701–1713. Association for Computational Linguistics.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *CoRR*, abs/2401.01055.

Michael Zhu and Suyog Gupta. 2018. To prune, or not to prune: Exploring the efficacy of pruning for model compression. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net.

A Languages in Evaluation Corpus

We use evaluation data composed of 30 languages to assess the model's linguistic competence. The 30 languages and their respective token counts (use LLaMA-2 Tokenizer) are as follows: Arabic (4702998), Chinese (2869208), Czech (1362041), Danish (36467), Dutch (3991305), English (1216599), Finnish (372303), French (6755281), German (2884921), Greek (474622), Hungarian (1229433), Indonesian (19226), Italian (6332560), Japanese (501899), Korean (2730794), Malay (5842), Malayalam (1489244), Norwegian (42289), Persian (1736589), Polish (4948702), Portuguese (7598161), Romanian (1381598), Russian (5205716), Spanish (7163860), Swahili (630), Swedish (1450236), Tamil (2920808), Turkish (2484186), Ukrainian (455720), Vietnamese (3606202).

B Core Linguistic Region

The regions are localized from six languages: Arabic, Spanish, Russian, Chinese, Korean, and Vietnamese, respectively. Our work does not alter the embedding layer, as we think it equates to a mapping of tokens, which does not involve modeling linguistic competence.

Region Visualization In Figure 9, we present the distribution of the 'Top' 5% regions in the Attn.o matrix for the LLaMA-2-13B model. The results indicate that across various layers, the core linguistic region on Attn.o matrix is concentrated on different rows. This difference is observed among the 40 various attention heads.

Removal 3% ratio (100K) LLaMA-2 perplexity on 30 languages when the removal ratio is 3% ratio, with $100,000$ samples for each language. Refer to Table 15 for more details.

Removal 3% ratio (10K) LLaMA-2 perplexity on 30 languages when the removal ratio is 3% ratio, with reduced $10,000$ samples for each language. Refer to Table 16 for more details.

Removal 1% and 5% ratio (100K) LLaMA-2-7B perplexity on 30 languages when the removal ratio is changed to 1% and 5% ratio, with $100,000$ equivalent samples for each language. Refer to Table 17 for more details.

C Attention Dimensional Removal

Figure 7 (left) illustrates that the columns of the Attn.k/q/v matrices in the attention layer, as well as the rows of the Attn.o matrix, correspond to different attention head parameters. Conversely, the rows of the Attn.k/q/v matrices and the columns of the Attn.o matrix are closely associated with dimensional features in the representation space.

We remove the 'Top' dimensions in the attention layer, and the results is displayed in Tables 6 and 7. Table 6 reveals that removing the Attention layers' 'Top' dimensions continues to produce more detrimental effects than other dimensions. The visualizations in Figure 2 show that these dimensions are largely concentrated in a few attention heads, suggesting that some attention heads contribute more significantly to the model's linguistic competence. Table 7 indicates that the removals under the second setting cause more damage than the first. Considering that, in the second setting, the 'Top' dimensions in the matrix directly interact with the corresponding dimensional features in the representational space, we can conjecture that these features are tightly linked with the model's linguistic competence.

D Single Parameter Perturbation

In a Transformer block, each column in the Attn.o and the MLP.down matrix of the FFN layer can be considered as the input weights of a neuron. Thus, perturbing a column can be seen as disturbing the

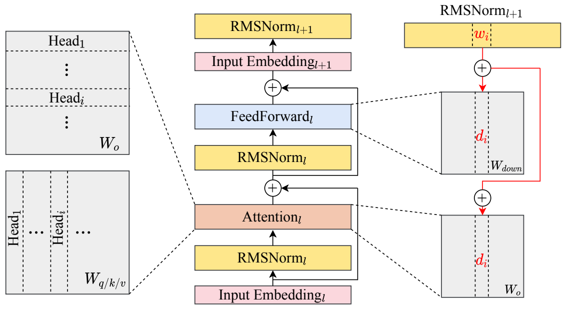

Figure 7: One can see from the left that each row of the Attn.o (W_o) corresponds to a particular attention head, and each column of the Attn.q/k/v ($W_{q/k/v}$) matrix corresponds to one as well. On the right, one can observe the perturbation applied to one weight within RMSNorm, which can be seen as affecting a column of the FFN.down and the Attn.o.

Model Size	# Training Samples	N_d	Attn.o(row), Attn.k/q/v(column)			
			Top	Middle	Bottom	Random
7B	100K	1	9.731	6.448	6.445	6.471
	100K	3	25.82	6.449	6.445	6.474
	100K	5	62.794	6.452	6.446	6.482
	100K	10	875.016	6.456	6.446	6.504
13B	100K	1	10.899	5.857	5.856	5.856
	100K	3	44.384	5.858	5.855	5.98
	100K	5	33.52	5.861	5.856	5.884
	100K	10	118.968	5.863	5.857	5.966
13B	10K	1	8.094	5.856	5.855	5.864
	10K	3	21.561	5.857	5.855	5.866
	10K	5	111.766	5.858	5.856	5.865
	10K	10	108.133	5.861	5.857	5.977

Table 6: Perplexity of LLaMA-2 after removing certain dimensions (zeroed-out) in the attention (Attn) layers. Here, N_d denotes the number of dimensions to remove, 'Top', 'Middle', and 'Bottom' refer to the dimensions with the most, moderate, and least cumulated $\mathcal{I}_\theta$ during further pre-training across six languages, respectively. 'Random' denotes an equivalent number of dimensions chosen at random for comparison.

Model Size	# Training Samples	N_d	Attn.o(column), Attn.k/q/v(row)			
			Top	Middle	Bottom	Random
7B	100K	1	167.804	6.446	6.446	6.446
	100K	3	68554.102	6.446	6.447	6.448
	100K	5	4259.861	6.449	6.447	6.449
	100K	10	68170.25	6.454	6.452	6.449
13B	100K	1	17.609	5.855	5.856	5.856
	100K	3	313.178	5.857	5.856	5.863
	100K	5	526.464	5.858	5.856	5.857
	100K	10	5841.446	5.859	5.858	5.852
13B	10K	1	17.03	5.855	5.856	5.857
	10K	3	206.225	5.856	5.856	5.858
	10K	5	1110.781	5.857	5.856	5.86
	10K	10	9600.097	5.859	5.858	5.874

Table 7: Perplexity of LLaMA-2 after removing certain dimensions in attention (Attn) layers. Different from Table 6, in this table, the columns of the Attn.o and the rows of the Attn.K/Q/V are removed.

input weights of a neuron. Viewed from another angle, if we disturb the output activation value of this neuron, a similar effect should be observed. Within LLaMA, there is a specific module called RMSNorm, where each dimension is associated with a weight. Perturbations to these weights can be regarded as disturbances to the output activation values of the corresponding neurons. In Figure 7 (right), we visually demonstrate how RMSNorm affects a column of the Attn.o and the FFN.down matrix.

Perturbation	Parameter	Perplexity
-	-	5.865
Reset 1	L1-N2100	83224.078
Reset 1	L1-N2800	5.860
Reset 1	L1-N4200	5.858
Mul 10	L1-N2100	4363.462
Mul 10	L1-N2800	5.859
Mul 10	L1-N4200	5.864

Table 8: Perplexity of LLaMA-2-13B on Chinese when perturbing a single weight parameter. Here, 'Reset 1' represents resetting the parameter to 1 (the initial value before pre-training), 'Mul 10' represents multiplying the parameter by 10. 'L1' represents 1-st layers. 'N' represents the 'Input_LayerNorm' module, followed by the perturbed dimension.

E Ablation Study

Tables 9 illustrate the perplexity of LLaMA-2-7B after removing core regions with and without outlier dimensions, respectively.

The ablation experiments reveal that different methods of disruption and varying model sizes exhibit different rates of PPL collapse:

1) Removing according to Attention.Head (attn.k/q/v.col + attn.o.row) results in a slower collapse than according to Dimensional Features (attn.k/q/v.row + attn.o.col). **2)** The 13B model shows a slower rate of collapse. **3)** The abnormal dimension is mainly concentrated in the FFN layer of the "core linguistic region". If preserving outlier dimension, the speed of PPL collapse by removing FFN layers decreases most obviously, while Attention.Head is almost unaffected.

Removal Region	N_d	LLaMA-2-**7B Top(100K)**	
		w/ outlier d	**w/o** outlier d
Attn.o(row) Attn.k/q/v(column) FFN.down(column)	1	848.326	27.265
	3	72594.445	57308.313
	5	48001.992	44730.059
	10	62759.516	73425.438
Attn.o(row) Attn.k/q/v(column)	1	9.731	9.732
	3	25.82	25.822
	5	62.794	23.296
	10	875.016	860.645
Attn.o(column) Attn.k/q/v(row)	1	167.804	9.586
	3	68554.1	136.318
	5	4259.861	688.476
	10	68170.25	431317.863
FFN.up/gate(row) FFN.down(column)	1	20.039	6.727
	3	74905.046	7.672
	5	114725.578	9.946
	10	239015.812	16.913

Table 9: Perplexity of LLaMA-2-7B after removing 'Top' certain dimensions w/ or w/o outlier dimensions respectively. Here, N_d denotes the number of dimensions to remove, 'Top' refers to the dimensions with the most cumulated $\mathcal{I}_\theta$ during further pre-training.

F Monolingual Region

Region Visualization In Figure 10, we present the distribution of the Attn.q matrix for 'Arabic' and 'Vietnamese' in 4 different layers. The results reveal that across various layers, the two monolingual regions are concentrated in different columns of the matrix.

Region Removal Tables 10-14 demonstrate LLaMA-2-7B perplexity after removing Arabic, Spanish, Chinese, Korean, and Vietnamese regions, respectively. The region is obtained by removing the intersections with other languages' respective regions from the 1% 'Top/Bottom' regions, selected from 10,000 or 100,000 sentences during further pre-training according to Equation 4.

Case Study In Figure 8, we use the prompt "*There are 365 days in a year and 12*" to test the model's output in English, Arabic, and Chinese, respectively. The results indicate that removing the monolingual regions causes the model to lose the relative language competence, leading the model to generate repetitive, nonsensical responses rather than correct answers like "*12 months in a year*".