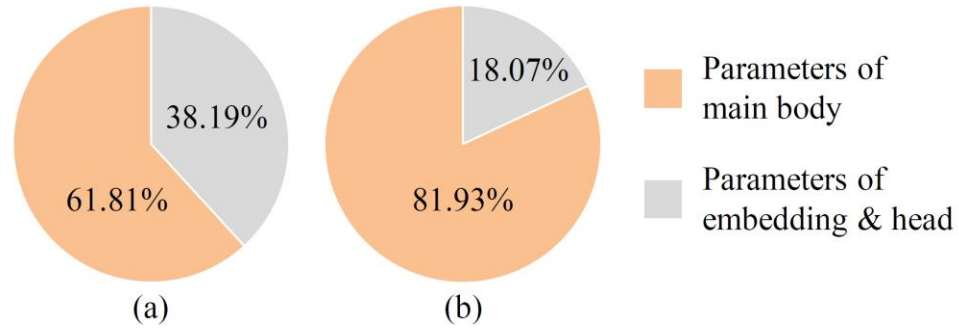
Rethinking Optimization and Architecture for Tiny Language Models

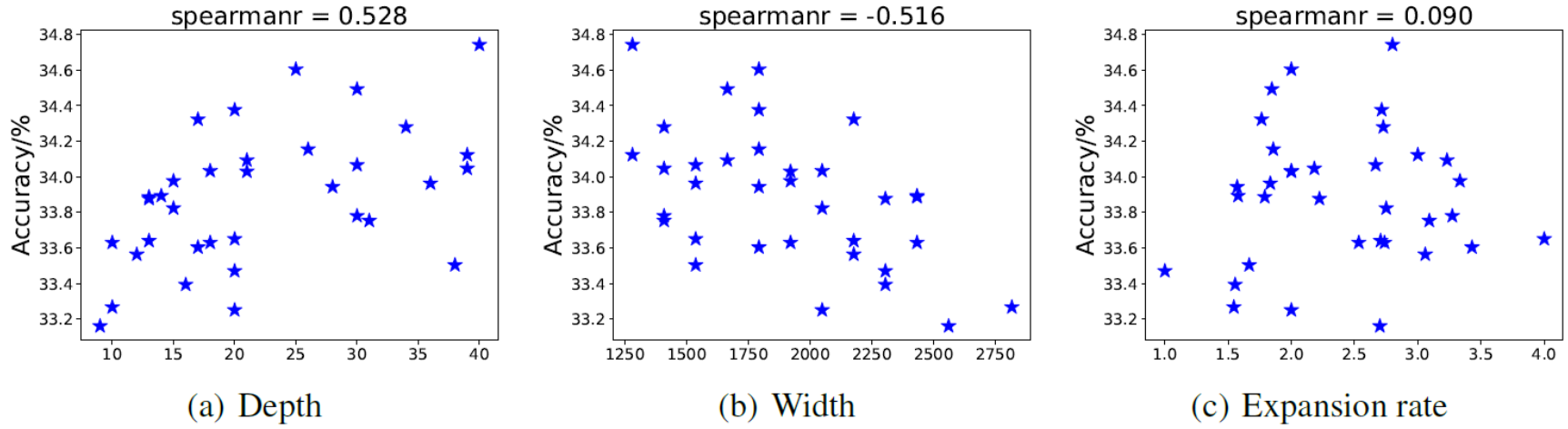
Yehui Tang[1], Kai Han[1], Fancheng Liu[1], Yunsheng Ni[1], Yuchuan Tian[2], Zheyuan Bai[1], Yi-Qi Hu[3], Sichao Liu[3], Shangling Jui[4], Yunhe Wang[1]

[1]Huawei Noah's Ark Lab; [2]Peking University; [3]Consumer Business Group, Huawei; [4]Huawei Kirin Solution.

Neural Architecture

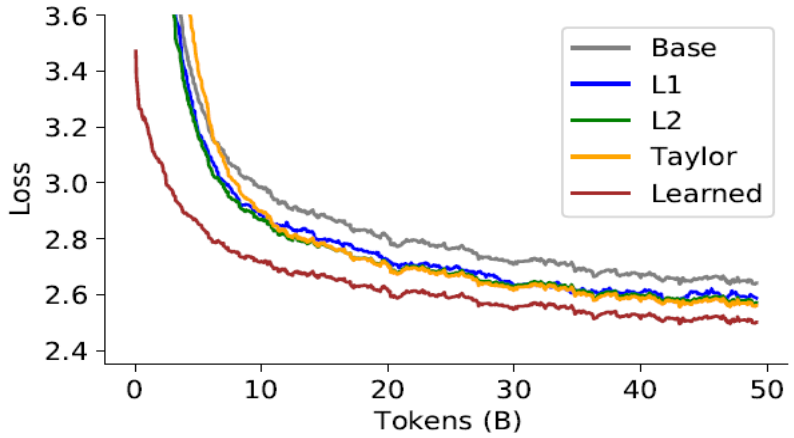


The parameter proportions of model's main body and tokenizer. (a) The large tokenizer inherited from large multilingual models(Wang et al., 2023) (b) Compact tokenizer by removing low-frequency vocabularies



Performance varies w.r.t. model's width, depth and expansion rate. The experiments are conducted on a streamlined dataset comprising 5B tokens. The accuracy is averaged among ARC Easy, HellaSwag and C3. Spearman coefficient is used to measure the correlation between performance and model's configure.

Parameter Initialization

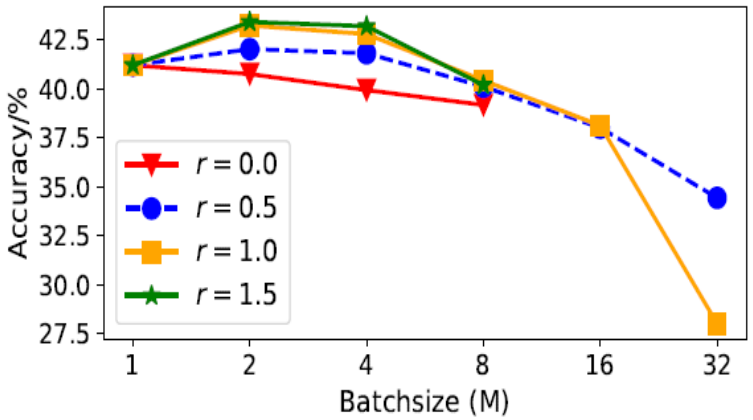


Training loss with different pruning strategies. “Base” denotes training from scratch without inheritance. Inheriting the model parameters with pruning yields a lower loss.

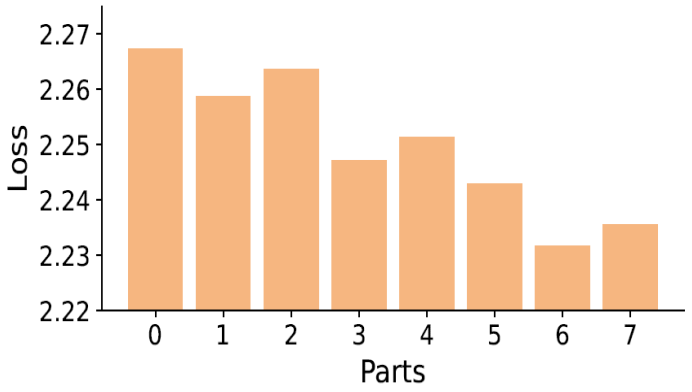
Comparison between different parameter inheritance strategies. “Base” denotes training without inheritance.

Inheritance Strategy	ARC-E	HellaSwag	C3	Avg.
Base	36.68	40.34	49.15	42.06
L1 (Ma et al., 2023)	39.51	47.70	50.96	46.06
L2 (Ma et al., 2023)	41.98	48.33	50.68	47.00
Taylor (Ma et al., 2023)	**43.21**	48.43	**52.05**	47.90
Learned (Xia et al., 2023; Tang et al., 2020)	40.74	**51.77**	51.73	**48.08**

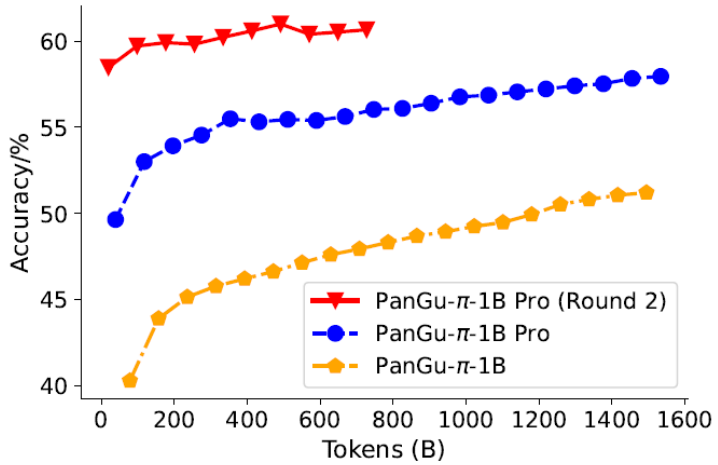
Model Optimization



Using $lr = (bs/bs_0)^r \times lr_0$, where the default batchsize bs0 and learning rate lr0 are set to 1M and 1 × 10−4, respectively. r denotes the increment rate, which is usually set as 0.5 or 1.0 (Krizhevsky, 2014; Goyal et al., 2017).



Loss value varies w.r.t. data on different iterations using a pretrained PanGu-π-1B model. The loss is averaged among batches in each part.



Accuracies of PanGu-π-1B and PanGu-π-1B Pro on HellaSwag during training.

Performance

Comparison with SOTA open-source tiny language models. The best model is listed in bold and second-best is listed in underlined.

Models	Examination				Knowledge	Reasoning		Understanding			Average
	C-Eval	CMMLU	MMLU	AGI-Eval	BoolQ	AX-b	PIQA	EPRSTMT	XSum	C3	
MobileLLaMA-1.4B	23.93	25.10	25.05	18.53	58.75	45.20	71.27	46.25	18.19	37.42	36.97
Sheared-LLaMA-1.3B	24.28	25.10	25.77	18.01	62.39	43.57	72.91	46.25	16.44	35.45	37.02
TinyLLaMA-1.1B	27.85	24.64	25.75	18.54	56.06	45.47	70.62	46.25	20.15	36.71	37.20
MobileLLaMA-2.7B	23.53	25.55	26.63	18.43	54.74	55.80	72.85	46.25	16.96	36.11	37.69
Chinese-LLaMA2-1.3B	28.70	24.78	24.55	19.40	56.79	47.46	56.91	72.50	8.90	43.12	38.31
RWKV-5-1.5B	25.92	25.14	25.66	19.01	62.29	54.05	71.22	46.25	<u>20.67</u>	49.15	39.94
Phi-1.3B	27.78	25.85	44.32	23.42	<u>73.52</u>	44.20	76.99	50.00	14.90	38.96	41.99
PanGu-π-1B	36.85	35.90	35.96	30.77	58.44	43.48	61.92	55.62	15.92	49.21	42.41
Open-LLaMA-3B	27.50	25.42	27.09	20.68	60.58	52.72	77.09	82.50	19.75	43.23	43.66
Phi2-2.7B	31.86	32.18	<u>58.49</u>	28.51	**77.40**	43.57	<u>78.89</u>	46.25	13.66	40.11	45.09
PanGu-π-1B Pro (**Ours**)	46.50	46.56	50.38	41.58	63.43	53.99	64.96	74.38	18.40	52.66	51.28
Qwen-1.8B	**53.60**	**52.12**	46.43	<u>35.83</u>	64.31	<u>57.79</u>	73.83	**88.12**	20.03	<u>58.30</u>	<u>55.04</u>
PanGu-π-1.5B Pro (**Ours**)	<u>52.39</u>	<u>48.51</u>	**62.54**	**44.89**	70.61	**67.93**	**79.55**	<u>86.12</u>	**24.61**	**69.24**	**60.64**

Thank you.