A Scaling laws

We use 7 models to fit the scaling laws of Baichuan 2. The parameter details are shown in Table 10.

N_{hidden}	N_{FFN}	N_{layer}	N_{head}	N_{params} (Millions)
384	1,152	6	6	11.51
704	2,112	8	8	51.56
832	2,496	12	8	108.01
1,216	3,648	16	8	307.60
1,792	5,376	20	14	835.00
2,240	6,720	24	14	1,565.60
2,880	8,640	28	20	3,019.33

Table 10: The model we choose for fitting scaling laws.

The losses of the 7 different models are shown in Figure 8.

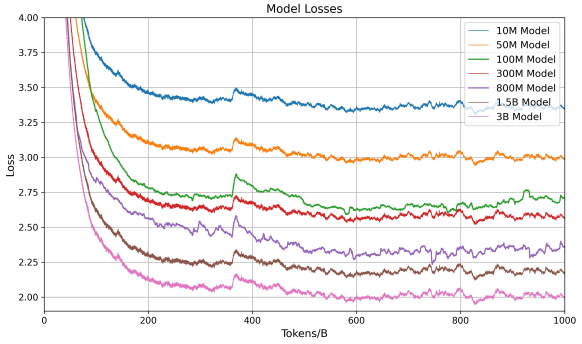

Figure 8: The various training loss of small models for scaling law.

B NormHead

By conducting a word embedding KNN retrieval task, where given a query word the nearest K words are retrieved. We found that the semantic information is mainly encoded by the cosine similarity of embedding rather than L_2 distance. i.e., The KNN results of cosine similarity are words with semantic similarity while the KNN results of L_2 distance are meaningless in some way. Since the current linear classifier computes logits by dot product, which is a mixture of L_2 distance and cosine similarity. To alleviate the distraction of L_2 distance, We propose to compute the logits by the angle only. We normalized the output Embedding so that the dot product is not affected by the norm of embedding.

To validate this operation, we conduct an ablation experiment where we add or remove the normalization before softmax and train a 7B model for 12k steps. All the hyper-parameters and data are the same with Baichuan 2-7B. The training loss is shown in Figure 9. We can see that when removing the *NormHead* the training became very unstable at the beginning, on the contrary, after we normalized the *head* the training became very stable, which resulted in better performance.

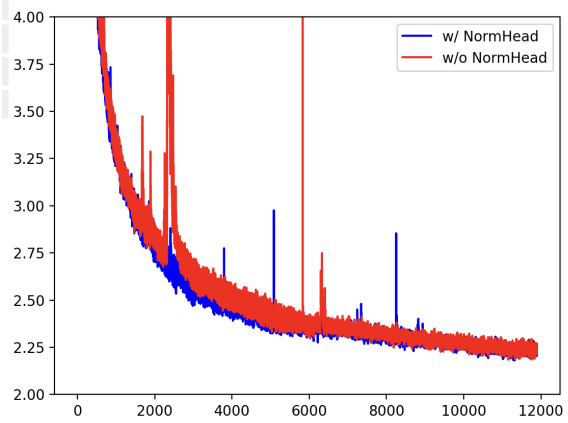

Figure 9: The training loss with and without NormHead operation. The experiments are conducted on 7 billion parameters with the same hyper-parameters (torch random seeds, data flow, batch size, learning rate, etc.)

C Training Dynamics

In this section, we analyze the training dynamics of our model. We save the checkpoints of Baichuan 2-7B and Baichuan 2-13B every 1000 steps. And evaluate those intermediate results on C-Eval development set (Huang et al., 2023), MMLU (Hendrycks et al., 2021a) , CMMLU (Li et al., 2023) , JEC-QA (Zhong et al., 2020), GSM8K (Shi et al., 2022) and HumanEval (Chen et al., 2021). The result is shown in Figure 10.

As shown, both the 7B and 13B models demonstrate substantial gains as training progresses. However, on general benchmarks such as MMLU (Hendrycks et al., 2021a) and C-Eval (Huang et al., 2023), improvements appear to plateau after 2 trillion tokens. In contrast, consistent gains are achieved on the GSM8K math tasks even beyond 2 trillion tokens. This suggests training FLOPs may strongly correlate with improvements in math problem solving, which may be further studied.

D Baichuan Harmless Evaluation Dataset

WARNING: this section contains unsafe, offensive, or upsetting examples of text.

We proposed the Baichuan Harmless Evaluation Dataset (BHED) to evaluate the chat models, as