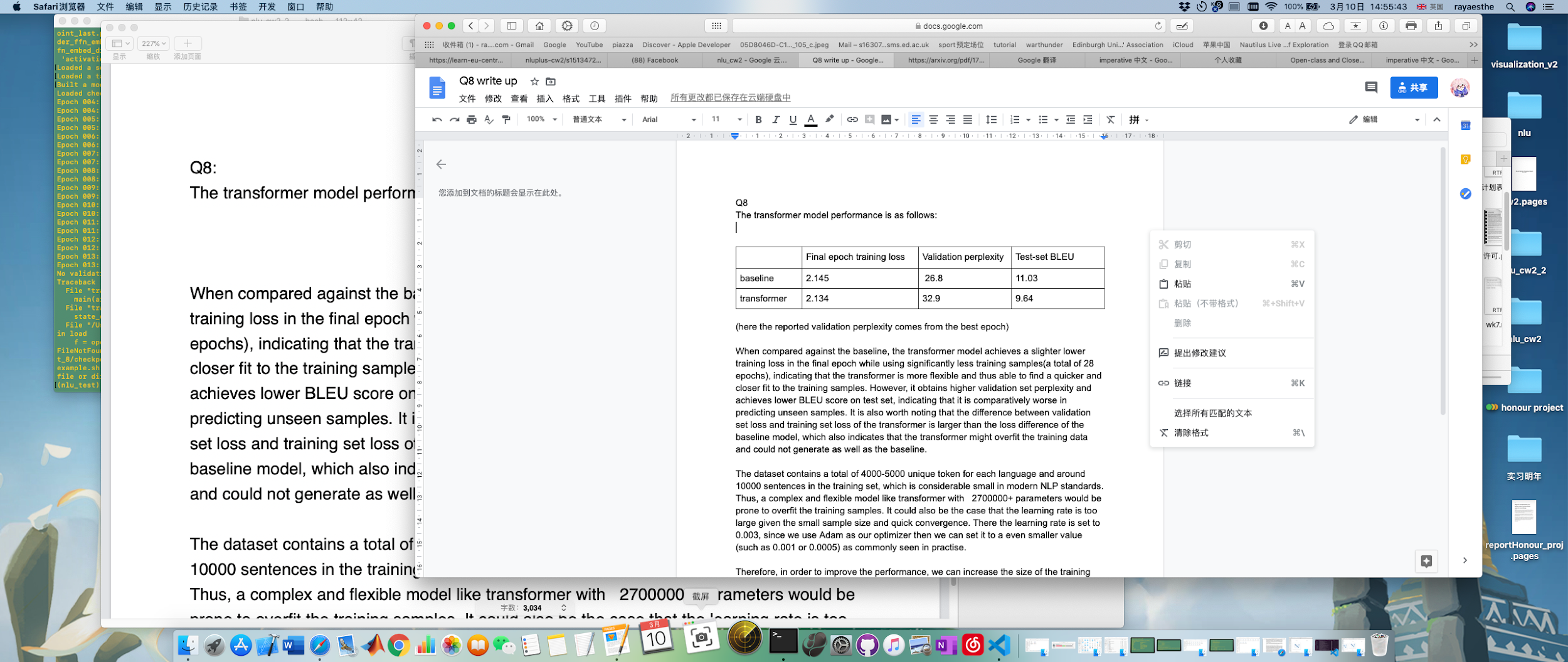
Q8

The transformer model performance is as follows:

|  | Final epoch training loss | Validation perplexity | Test-set BLEU |
| --- | --- | --- | --- |
| baseline | 2.145 | 26.8 | 11.03 |
| transformer | 2.134 | 32.9 | 9.64 |

(here the reported validation perplexity comes from the best epoch)

When compared against the baseline, the transformer model achieves a slighter lower training loss in the final epoch while using significantly less training samples(a total of 28 epochs), indicating that the transformer is more flexible and thus able to find a quicker and closer fit to the training samples. However, it obtains higher validation set perplexity and achieves lower BLEU score on test set, indicating that it is comparatively worse in predicting unseen samples. It is also worth noting that the difference between validation set loss and training set loss of the transformer is larger than the loss difference of the baseline model, which also indicates that the transformer might overfit the training data and could not generate as well as the baseline.

The dataset contains a total of 4000-5000 unique token for each language and around 10000 sentences in the training set, which is considerable small in modern NLP standards. Thus, a complex and flexible model like transformer with 2700000+ parameters would be prone to overfit the training samples. It could also be the case that the learning rate is too large given the small sample size and quick convergence. There the learning rate is set to 0.003, since we use Adam as our optimizer then we can set it to a even smaller value (such as 0.001 or 0.0005) as commonly seen in practise.

Therefore, in order to improve the performance, we can increase the size of the training dataset, for example, by finding other corpuses of English-German sentence pairs. We can also attempt to use a pre-existing attention model with pre-trained weights on related language translation quests. In addition, we could also experiment with hyper-parameters like the learning rate to slow down the training process and mitigate the effect of overfitting.