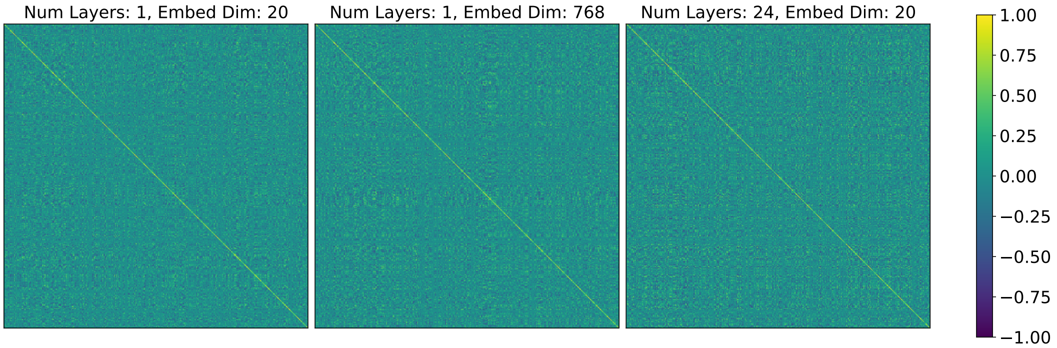

Figure 1: A heat map of cosine similarity between token embeddings after 3000 epochs training under different settings. All the other configurations are the same as default (in particular, the vocabulary size is 800). The above figures show that the token embeddings are nearly-orthogonal for different number of layers and different embedding dimensions, even for one-layer transformer (as mentioned in Question 5 by Reviewer yXHq) or an embedding dimension much smaller than the vocabulary size (as mentioned by multiple reviewers and clarified in the global response).

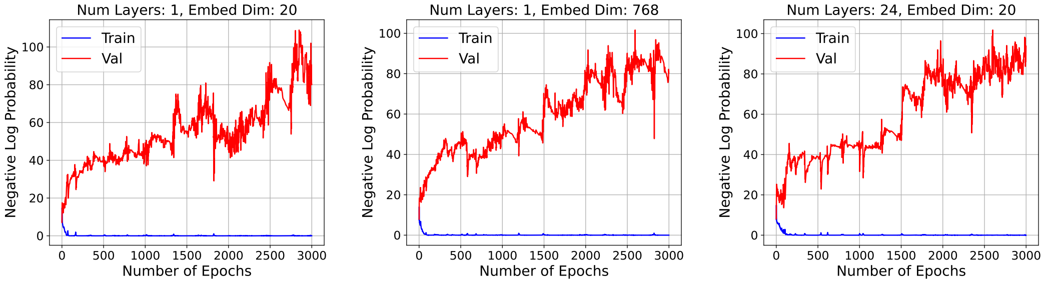

Figure 2: Training and validation loss for the reversal curse under the same settings as Figure 1. Similarly, the above figures show that the reversal curse can still be observed for different number of layers and different embedding dimensions.

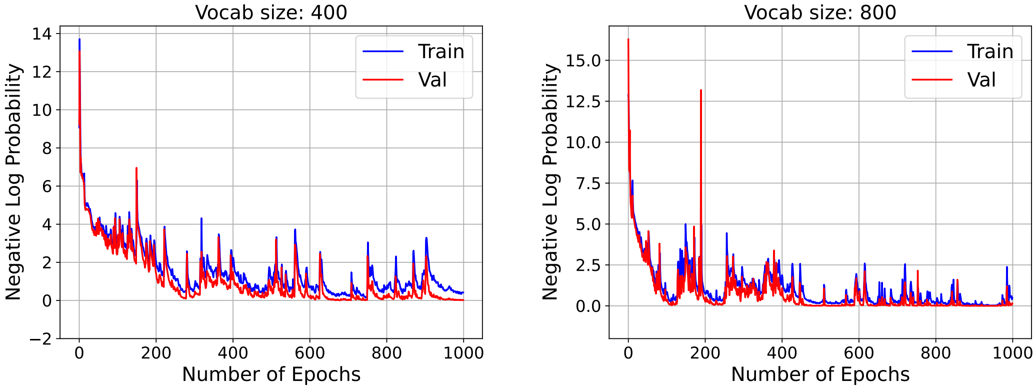

Figure 3: Training and validation loss for in-context learning (ICL). All sentences consist of seven tokens and have the form of “$\mathtt{A}_i \to \mathtt{B}_j \Leftrightarrow \mathtt{B}_j \leftarrow \mathtt{A}_i$”. The loss is calculated on the last token. The above figures show that ICL could help to mitigate the reversal curse (as mentioned both by Reviewers bwDn and 82qZ).