先端科学技術研究科　修士論文要旨

所属研究室 （主指導教員）	自然言語処理学 （渡辺　太郎　（教授））		
学籍番号	2411404	提出日	令和 8年 1月 19日
学生氏名	長谷川　遼		
論文題目	Knowledge Editing Induces Underconfidence in Language Models		

要旨

As language models continue to scale, the demand for knowledge editing, a retraining-free knowledge update method, has increased. However, since knowledge editing directly alters token prediction probabilities acquired during pretraining, the probabilities may diverge from the empirical distribution. In this study, we analyze the impact of knowledge editing to compare the alignment between token prediction probabilities and task accuracy by calculating confidence calibration before and after knowledge editing. Our results reveal that, for tasks requiring semantic understanding, the range of increase in token prediction probabilities tends to be smaller than that of accuracy improvement, suggesting that knowledge editing methods lead to less confidence in prediction.