

The Transformer architecture is very popular, so it is potentially impactful to interpret what influences its performance. We test the hypothesis that the model relies on the linguistic properties of a text when working with it. We remove interference with cultural aspects of meaning by using a character-level task with the ByT5 Transformer model. We train ByT5 to decipher sentences encrypted with text ciphers (Vigenère, Enigma). We annotate a sentence dataset with linguistic properties with published NLP tools. On this dataset, we study the relationships between the linguistic properties and the fine-tuned ByT5 decipherment error rate. We analyze correlations, train ML models to predict error rates from the properties and interpret them with SHAP. We find small significant correlations but cannot predict error rates from the properties. We conclude the properties we identified do not give much insight into the performance of the Transformer.