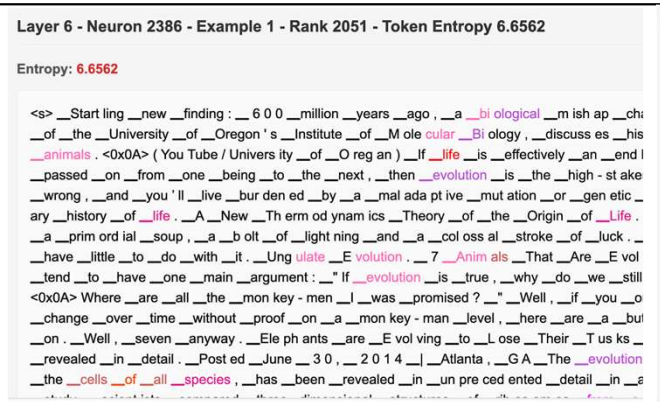
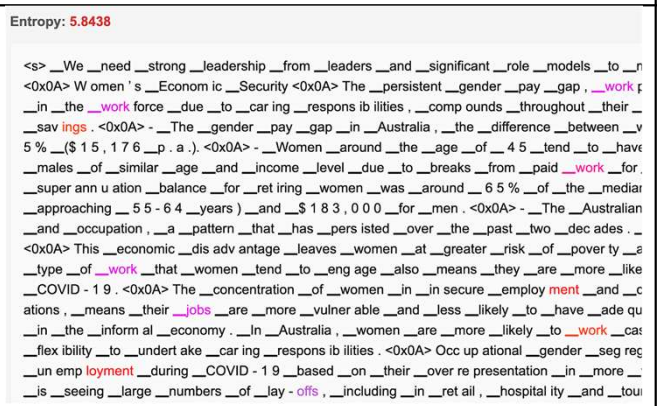
ABCI 3.0開発加速利用 （2024年度）成果概要（公開用）

課題名：Towards Interpretable Foundation Models: Sparse Auto-Encoder-Based Transformer Architectures	実施時期：2025年１月２０日から2025年3月31日まで 所属機関名：理化学研究所革新知能統合研究センター 自然言語理解チーム 代表者氏名：乾健太郎

成果概要：Transformer-based neural network architectures are the key technology underlying the rapid progress of generative AI. However, the internal computational processes of these models is difficult to understand and interpret. Previous work has predominantly relied on post-hoc interpretation methods, such as training sparse auto-encoders (SAEs). SAEs disentangle internal representations into more easily interpretable features but suffer from a low degree of internal validity. Here, we developed a SAE-like component that can inserted at points of interest in the Transformer architecture. This SAE-based Transformer is then trained from scratch, resulting in a model that is inherently interpretable at the insertion points.

成果のポイント：Our main result achieved during the accelerated development program is a proof-of-concept which shows that it is possible to insert sparse autoencoder-like components directly during pretraining of transformer-based neural networks. Since naïve insertion of the SAE-like components degraded performance, we developed strategies for improving and speeding up the convergence the training of SAE-based Transformer architectures. The resulting highly interpretable models incurred only a small performance penalty that we were able to offset by slightly increasing the model size. Similarly to SAEs, the features learned by our SAE-like components appear to encode interpretable concepts such as “words related to biology and evolution” (left figure) or “words related to work and employment” (right figure). In addition to interpreting the outputs of the fully trained model, the sparsity of the activations in the SAE-like components also allows tracking how these features develop and gradually specialize during training.





成果についてより詳細な情報を提供しているWebページ、発表論文などの情報：n/a