Interpreting Reflexive Binding in Transformer Language Models through Feed-Forward Networks

Jieun Kim
University of Ulsan
kimje@ulsan.ac.kr

Transformer language models (LMs) exhibit increasingly human-like behavior on syntactic phenomena, yet little is known about how abstract grammatical constraints are internally implemented. While previous studies have demonstrated Principle-A-sensitive behavior using surprisal measures, the underlying neural computations remain largely unexplored. This study investigates how reflexive binding is implemented inside transformer language models using mechanistic interpretability.

Experiments were conducted on **Pythia-6.9B** using a linguistically controlled dataset of **100 English sentence sets (600 sentences)**. Each set consists of three grammatical conditions—**True Antecedent**, **False Antecedent**, and **Corrupted**—combined with animate and inanimate distractor variants. Unlike standard minimal-pair evaluations, the dataset orthogonally manipulates structural accessibility and cue-based plurality while independently varying distractor animacy, allowing the contributions of structural licensing and retrieval cues to be examined separately.

Behaviorally, surprisal at the reflexive pronoun confirms that the model reliably prefers structurally licensed antecedents while exhibiting cue-based interference patterns consistent with psycholinguistic findings.

To investigate the underlying computation, we perform neuron-level feed-forward network (FFN) analysis based on sub-update decomposition. Pairwise effect analysis identifies neurons whose contributions differ systematically across grammatical contrasts, after which the corresponding value vectors are projected into vocabulary space to characterize the semantic information encoded by individual neurons.

The analysis reveals that Principle-A-sensitive computation is concentrated within a relatively small subset of recurrent FFN neurons exhibiting highly consistent computational behavior across grammatical contrasts. Vocabulary projection further identifies two functionally distinct neuron populations: one projects strongly toward plural lexical fields (e.g., *they*, *their*, *themselves*), whereas another exhibits equally stable computational effects without obvious lexical specialization, suggesting more abstract computational roles.

These findings provide one of the first mechanistic accounts of reflexive binding in transformer language models and demonstrate how linguistically controlled datasets combined with mechanistic interpretability can bridge behavioral evidence and internal neural computation.