SYNTACTIC INFORMATION DRIVES LANGUAGE MODELS' ALIGNMENT WITH HUMAN COMPREHENSION PROCESSES

Zhengwu Ma (City University of Hong Kong), Chengcheng Wang (City University of Hong Kong) & Jixing Li (City University of Hong Kong)
zhengwuma2-c@my.cityu.edu.hk, cwang495-c@my.cityu.edu.hk, jixingli@cityu.edu.hk

Introduction With the recent success of large language models (LLMs), there has been an increasing interest in understanding whether these models process sentences similarly to humans. Prior studies have suggested a high correlation between transformer-based language models and human behavioral and neural responses during naturalistic reading or listening (e.g., Goldstein et al., 2022; Schrimpf et al., 2021). However, it remains unclear what linguistic information was incorporated by the LLMs during comprehension. In this study, we compared BERT's attention layers with human eye movement and fMRI activity patterns while native English speakers (L1) and non-native learners of English (L2) engaged in self-paced reading tasks during fMRI recording. To assess the contribution of syntactic information, we further removed syntactic dependency extracted from BERT's attention layers and conducted the correlation analysis again.

Methods We used the openly-available Reading Brain dataset (Li et al, 2022), which includes concurrent eye-tracking and fMRI blood-oxygen-level-dependent (BOLD) signals while participants read 5 short English articles. The subjects included 52 L1 speakers and 56 L2 speakers of English. For each subject, we used the eye-fixation timepoints to extract the fMRI signals within a left-lateralized language mask time-locked to each word and constructed an fMRI data matrix for each sentence. We also extracted the saccade pattern for each sentence, and we compared the fMRI and saccade patterns with the attention patterns of BERT using representational similarity analysis (RSA; Kriegeskorte et al., 2008). Following Manning et al. (2020), we extracted the dependency distances of each sentence from BERT's attention layers and we regressed them out from the attention layers. At the group level, we examined the correlation coefficients between the fMRI/saccade patterns and the attention patterns of BERT using a one-sample t-test. Statistical significance was determined by a cluster-based permutation test (Maris & Oostenveld, 2007) with a threshold of $p < 0.05$ family-wise error (FWE) correction.

Results and conclusion Our results showed a significant correlation between BERT's attention patterns and the saccade patterns of both L1 and L2 speakers, with the maximum correlation at layer 15 ($r= 0.29$, $p<.001$ for L1 and $r=0.33$, $p<.001$ for L2). However, after removing syntactic dependency from the model's attention layers, the correlation coefficients decreased substantially for both L1 and L2 speakers (e.g. at Layer 15: $r= 0.29$, $p<.001$ for L1 and $r=0.33$, $p<.001$ for L2; see Figure 1A). The fMRI results revealed a significant cluster at the left middle temporal lobe (LMTL) for both L1 and L2 speakers (see Figure 1B), yet no significant clusters were found after removing syntactic dependency from the attention layers. These results suggest that syntactic parsing plays a major role in LLMs' high alignment with human comprehension processes. Notably, the correlation coefficients between BERT's attention patterns and human eye movement patterns were significantly higher for L2 speakers than for L1 speakers, suggesting that current language models may employ comprehension strategies that differ from those of native speakers.

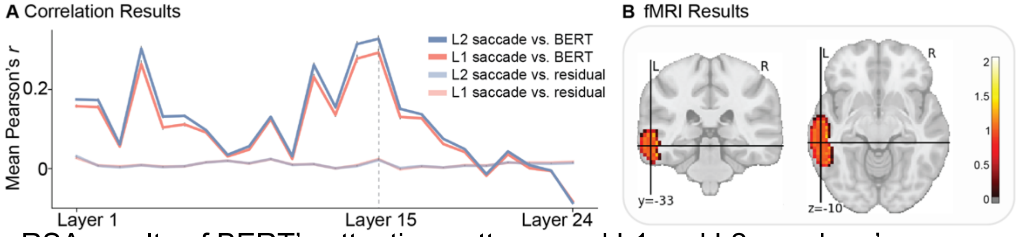

Figure 1. RSA results of BERT's attention patterns and L1 and L2 speakers' eye movement and fMRI activity patterns during naturalistic reading comprehension.