Evidence Against the Modularity of Syntax in Large Language Models

Thomas A. McGee and Idan A. Blank, (UCLA)

Background. Transformer language models have recently demonstrated exceptional performance in a variety of linguistic tasks [1]. To represent sentences, transformers use components called "attention heads" to capture how words in a text input relate to one another by assigning numerical weights linking different words [2]. Some attention heads assign weights that accurately encode meaningful linguistic features [3-4] including, importantly, heads that appear specialized for identifying particular syntactic dependencies [3-6]. These heads are a main candidate for "encapsulated" syntactic computations that are impenetrable to non-syntactic information. Such computations would be strikingly different from those of the human mind, where non-syntactic information sources (e.g., semantics) influence parsing from the earliest moments of online processing [7-9] and where syntax and semantics are tightly linked in the mental lexicon [10-12]. Therefore, we tested whether the activity of syntax-specialized attention heads in transformers is modulated by one type of semantic information: plausibility.

Materials. We studied 5 dependency types from the Stanford Dependencies for English [13]: direct object (*dobj)*, passive auxiliary (*auxpass*), nominal subject (*nsubj*), indirect object (*iobj*), and object of a preposition (*pobj*). For each dependency, we constructed 50 pairs of sentences (all with the same structure) such that in one sentence the dependency was semantically plausible (in the broader context of the sentence), and in the other—implausible (Table 1). For the first 4 dependencies, plausibility was manipulated by swapping two words in the sentence. For *pobj*, it was manipulated by replacing the prepositional object with a new word. Each sentence also contained a "lure" word that was not part of the critical dependency but, in the implausible sentences, was a semantically plausible target for that dependency (Table 1).

Predictions. We predicted that "syntax-specialized" heads would not be encapsulated from plausibility information. Thus, compared to plausible sentences, in implausible sentences less attention would be allocated to the syntactically correct target, and more attention to/from the lure.

Procedure. We analyzed the BERT-base-uncased model [14], the GPT-2 small model [15] (both trained on word prediction) and are currently analyzing Llama2. We first identified attention heads in BERT that were specialized for each dependency, replicating [4]; then, we extended the approach to GPT-2 (where such heads have not been previously identified). To this end, across a subset of the Penn Treebank 2 [16], for each dependency we found the attention head that most consistently allocated more attention between words in that dependency than to other words in the sentence. Once attention heads of interest were identified (one per dependency), we analyzed the weights they assigned between critical words in our stimuli. We tested BERT on all 5 dependencies, and GPT-2 on 4 dependencies. Because attention weights are bounded in [0,1], we logit-transformed them prior to linear, mixed-effects modeling. We used linear regression to model attention weight as a function of plausibility, with word frequency as a control. One analysis modeled attention within the critical dependency, and another—to/from the "lure" words (Bonferroni corrected for the number of dependencies). We included a random intercept by sentence pair unless the model did not converge (and was reduced to fixed effects only).

Results and Discussion. Average attention strength between critical words in BERT was significantly higher in plausible vs. implausible sentences for *dobj*, *nsubj*, *iobj,* and *pobj*, but not *auxpass* ($t_{(48.65)}$=2.54, p=0.0143) (Fig. 1). Additionally, attention directed to/from lure words was significantly greater in implausible sentences for all dependencies except for *dobj*, ($t_{(45.03)}$=-0.077, p=0.94). In GPT-2, attention strength between critical words was significantly higher in plausible vs. implausible sentences for all dependencies (Table 2). Attention directed to/from lure words was higher in the implausible condition for *dobj* and *nsubj*, but not for *iobj* ($t_{(97)}$=-1.060, p=0.292) or *pobj* ($t_{(48.02)}$=-2.07, p=0.044). Therefore, even in attention heads that are most strongly selective for specific syntactic dependencies, computations are influenced by semantics. This result mirrors the penetrability of the syntactic parser to semantics in the human mind.

Table 1 Example sentence pairs for BERT. Words in bold constitute the critical dependency. Attention is directed from the underlined word to the other bolded word. Words in red are "lures".

Dependency	Plausible / likely	Implausible / unlikely
Direct object	The guide **showed** the visitor a **sculpture.**	The guide **showed** the sculpture a **visitor.**
Passive Auxiliary	The car that was never repaired **was** severely **damaged**.	The car that was never damaged **was** severely **repaired**.
Nominal Subject	The **chef** in the kitchen **cooked** a stew.	The **kitchen** in the chef **cooked** a stew.
Indirect object	The guide **showed** the **visitor** a sculpture.	The guide **showed** the **sculpture** a visitor.
Object of a preposition	It was the **ladder** in the tree that the cat climbed **up**.	It was the **simplicity** in the tree that the cat climbed **up**.

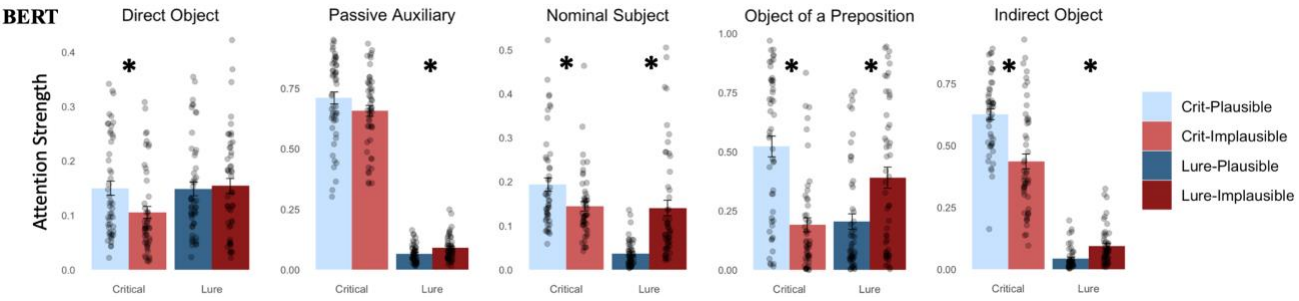

Figure 1. Attention strength between words in a critical dependency for the five dependency types in BERT. Each plot shows data from a different attention head (the one most "specialized" for the corresponding dependency). Dots represent the attention strength for individual sentences. Error bars show ± 1 standard error of the mean. Bright bars show attention for the critical, syntactically correct dependency ("crit"), and dark bars show attention to/from semantic lures.

Table 2 Statistical results for GPT-2. Results for the critical word analyses are shown in the "crit" column, while results for the lure word analyses are shown in the "lure" column.

Dependency	Crit β_1	Crit SE	Crit p-value	Lure β_1	Lure SE	Lure p-value
Dobj	0.800	0.126	$<10^{-4}$	-0.388	0.115	0.0014
Nsubj	0.564	0.148	0.0004	-2.553	0.221	$<10^{-4}$
Iobj	0.602	0.160	0.0003	-0.177	0.167	0.292
Pobj	0.158	0.057	0.0074	-0.023	0.011	0.044

References

[1] Chang & Bergen (2023). *arXiv preprint.* **[2]** Vaswani et al. (2017). *Adv. Neural Inf. Process. Sys.* **[3]** Vig, J., & Belinkov, Y. (2019). *arXiv preprint.* **[4]** Clark et al. (2019). *ACL*. **[5]** Voita et al. (2019). *arXiv*. **[6]** Raganato & Tiedemann (2018). *EMNLP*. **[7]** Trueswell et al. (1994). *J. mem. Lang.* **[8]** Garnsey et al. (1997), J. Mem. Lang. **[9]** Traxler et al. (2002), J. Mem. Lang. **[10]** Goldberg, (1995). U. Chicago Press. **[11]** Jackendoff (2007), Brain Res. **[12]** Bybee (2010), Cambridge U. Press **[13]** De Marneffe et al. (2006). **[14]** Devlin et al. (2018). *ArXiv preprint*. **[15]** Radford et al. (2019). *OpenAI blog.* **[16]** Marcus et al. (1995).