Do Large Language Models know who did what to whom?

Joseph M. Denning (UCLA), Xiaohan (Hannah) Guo (University of Chicago), Bryor Snefjella (Arizona State University), Idan A. Blank (UCLA)

Introduction. Large Language Models (LLMs) have achieved unprecedented success at natural language processing, especially in learning syntactic abstractions [e.g., 1,2]. However, LLMs have been criticized for not "understanding" language [e.g., 3,4]. These critiques are hard to evaluate because they conflate "understanding" with logical reasoning and common sense—abilities that, in human minds, are dissociated from language processing per se [5]. Here, we instead focus on a form of understanding that is tightly linked to language: mapping sentence structure onto an event description of "who did what to whom" (thematic roles). Because event semantics might be internally represented even if not evident in LLMs' output (e.g., next word prediction; cf. [6]), we study the hidden representations of these networks in two experiments.

Experiment 1. We conducted a "multivariate pattern analysis" [7] on LLMs' sentence representations. Following an fMRI study [8], we used 94 sets of English sentences, each including a simple, active, transitive "base" sentence (e.g., *the lawyer saved the author*), and 4 edited versions that were similar to their "base" in either (1) meaning (thematic roles) but not syntax (*the author was saved by the lawyer*); (2) syntax but not meaning (*the author saved the lawyer*); (3) both (*the attorney rescued the writer*); or (4) neither (*the attorney was rescued by the writer*). We extracted representations of these sentences from each hidden layer of BERT and GPT2. We compared the representation of the base to the representation of each edited version, using cosine similarity (per layer; Fisher-transformed to improve normality). These similarities were compared across the 4 conditions via Tukey tests in a non-parametric, repeated-measures ANOVA. We also gathered human similarity judgments (N=120) for a subset of these stimuli; each participant saw one sentence pair per condition (each from a different stimulus set), embedded among 4 filler pairs. We found an unexpected pattern: sentence pairs that had opposite (reversed) thematic role assignments, but shared syntax, were represented by LLMs as *more* similar than pairs sharing thematic role assignments but differing in syntax (p<.001; see **Figure 1**). In contrast, human similarity judgments were driven by thematic role assignment rather than by syntax ($p<10^{-16}$; see **Figure 1**). Distributed LLM activity patterns do not robustly represent event structure.

Experiment 2. We tested whether event structure was instead localized to a small subset of hidden units using a probing method, i.e., training a support vector machine on a binary "same/different meaning" classification of sentence pairs (whose representations were concatenated or subtracted). To prevent the use of heuristics on the simple materials from Experiment 1, we created 50 sentence sets where ditransitive base sentences (e.g., *the man gave the milk to the woman*) were edited in 23 ways (e.g., *it was the milk that the woman gave to the man*) such that no obvious "trick" applied across all of them to infer meaning similarity. In addition to testing hidden units, we also tested representations extracted across the attention heads of each layer. We found little evidence that thematic role information was robustly available anywhere among hidden units (performance not above 60%, with one exception; see **Figure 2**). However, some attention heads did reflect thematic roles independently of syntax (see **Figure 3**).

Conclusion. Overall, we find that some components within LLMs do capture thematic role assignments. However, such information appears to exert a much weaker influence on their hidden sentence representations compared to its influence on human similarity judgments. These results demonstrate a gap between humans and machines in the representation of even an extremely basic form of meaning: who did what to whom. We are currently extending our analyses to the more recent models Llama-2 and Persimmon.

Figure 1: Similarity of base sentence to each edited version for BERT, GPT2, and human participants (data shown for the subset of sentences that were rated by human participants; the patterns seen here hold for the full set of sentences). Each dot represents one item (the "human" plot averages across the five participants who saw each item). Error bars show standard errors. Critically, the two center bars (B, C) show opposite patterns between LLMs and humans.

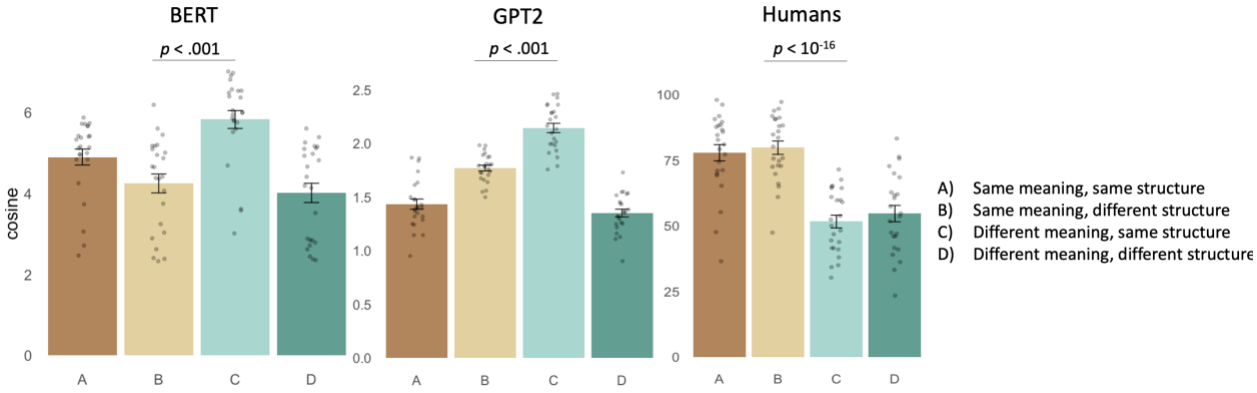


Figure 2. SVM classification accuracies for same vs. different meaning, run on either concatenated or subtracted hidden unit activations for pairs of sentences.

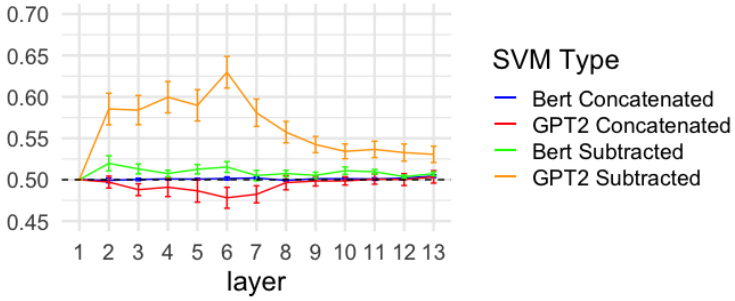
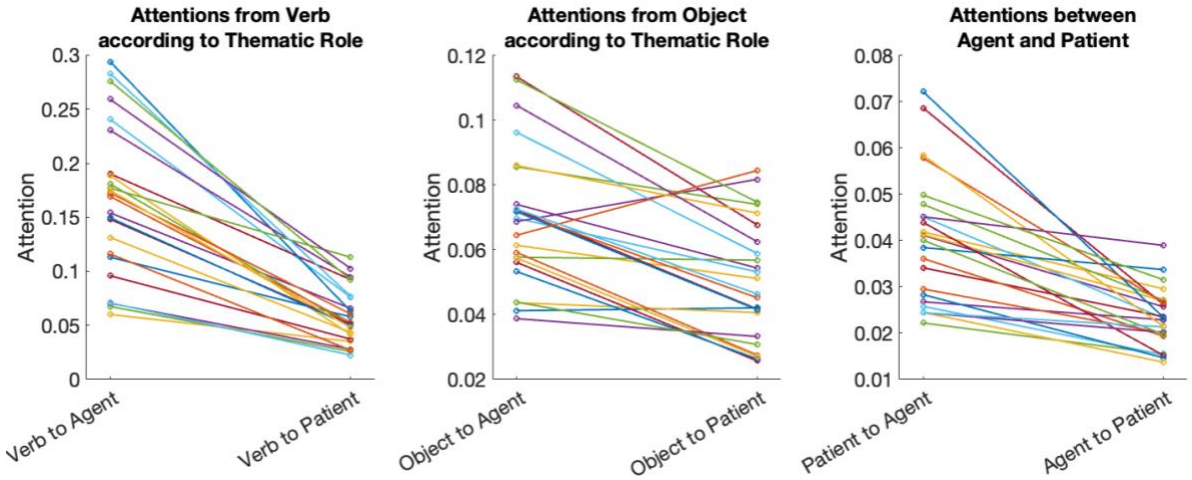


Figure 3: Attention strength between different words for each of the 24 sentence types (colored lines), extracted from BERT's attention head with the highest classification accuracy for same vs. different meaning (Head 5, Layer 11, SVM Accuracy 79%).



References. [1] Rogers et al. (2020). *ACL.* **[2]** Manning et al. (2020). *PNAS*. **[3]** Bender & Koller (2020). *ACL*. **[4]** Marcus (2020). *ArXiv*. **[5]** Mahowald, Ivanova, et al. (2023). *ArXiv*. **[6]** Ettinger (2020). *TACL*. **[7]** Haxby (2012). *NeuroImage*. **[8]** Fedorenko et al. (2020). *Cognition*.