UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Do Large language Models know who did what to whom?

Permalink

https://escholarship.org/uc/item/35x5m9cm

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Denning, Joseph
Guo, Xiaohan (Hannah)
Snefjella, Bryor
et al.

Publication Date

2024

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Do Large language Models know who did what to whom?

Joseph M. Denning[1] (josephdenning@gmail.com)
Xiaohan (Hannah) Guo[2] (hannahguo@uchicago.edu)
Bryor Snefjella[1]
Idan A. Blank[1] (iblank@psych.ucla.edu)
[1] Department of Psychology, 502 Portola Plaza, Los Angeles, CA 90095 USA
[2]Department of Psychology, 5848 S University Ave, Chicago, IL 60637 USA

Abstract

Large Language Models (LLMs), which match or exceed human performance on many linguistic tasks, are nonetheless commonly criticized for not "understanding" language. These critiques are hard to evaluate because they conflate "understanding" with reasoning and common sense—abilities that, in human minds, are dissociated from language processing per se. 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). Whereas LLMs can be directly trained to solve to this task, we asked whether they naturally learn to extract such information during their regular, unsupervised training on word prediction. In two experiments, we evaluated sentence representations in two commonly used LLMs—BERT and GPT-2. Experiment 1 tested hidden representations distributed across all hidden units, and found an unexpected pattern: sentence pairs that had opposite (reversed) agent and patient, but shared syntax, were represented as more similar than pairs that shared the same agent and same patient, but differed in syntax. In contrast, human similarity judgments were driven by thematic role assignment. Experiment 2 asked whether thematic role information was localized to a subset of units and/or to attention heads. We found little evidence that this information was available in hidden units (with one exception). However, we found attention heads that reflected thematic roles independent of syntax. Therefore, some components within LLMs capture thematic roles, but such information exerts a much weaker influence on their sentence representations compared to its influence on human judgments.

Keywords: large language models; comprehension; thematic roles; representational similarity

Introduction

Large Language Models (LLMs) have achieved unprecedented success at natural language processing. Their success demonstrates the power of statistical learning over strings of linguistic forms (Contreras Kallens et al., 2023; Piantadosi, 2023): by merely learning to predict the next word in a text, LLMs develop the ability to produce texts that conform to the syntactic rules of a language (e.g., McCoy et al., 2023). Indeed, the internal representations and next-word predictions of LLMs suggest that these systems have acquired many complex grammatical generalizations (for a review, see: Linzen & Baroni, 2021). Nonetheless, humans do not use grammar as an end, but rather as an intermediate step in mapping linguistic input forms onto meaning (semantics). Can LLMs extract meaning from their input? More specifically: what kinds of meaning can be acquired merely from learning to predict the next (or a missing) word?

Whereas the behavior and internal activity of LLMs exhibit some signatures of semantic representations (for a review, see Pavlick, 2022), these models are often criticized for not truly "understanding" language (e.g., Bender & Koller, 2020). The critiques use various definitions of "understanding": Some critics claim that LLMs do not have "grounded" knowledge that links linguistic meaning to non-linguistic experience (Bender & Koller, 2020; Bisk et al., 2020); others claim that LLMs lack common sense, i.e., intuitive theories about how the world works (Sinha et al., 2019; Ullman, 2023); yet others claim that LLMs do not reason logically (Ettinger, 2020; Wu et al., 2023). All such criticisms conclude that LLMs do not understand language like humans do.

However, these critiques are valid only to the extent that they rely on accurate notions of language comprehension in humans. Whereas humans can relate linguistic input to non-linguistic experiences, evaluate it against prior knowledge, and use it for logical inferences, all these capacities are kinds of "thinking", not language (Mahowald, Ivanova, et al., 2024): in the human mind, the cognitive systems that support sensorimotor or affective processes, common sense reasoning, and logical inferences are functionally distinct from the system that analyzes linguistic input (Fedorenko et al., 2024; Fedorenko & Varley, 2016). Whereas linguistic processing is a prerequisite for, e.g., evaluating whether a sentence is consistent with common sense, these two steps are dissociable, and a failure to carry out the latter does not demonstrate a failure to carry out the former. Given that the mind dedicates a system to linguistic processing per se, a fair yet critical bar for LLMs to pass is semantic processing that is language-internal or, at least, closely tied to language (e.g., Piantadosi & Hill, 2022, but see Jackendoff, 2012).

Here, we test such a case of understanding: using the structure of a sentence to figure out "who did what to whom", a process called "*thematic role assignment*" (Rissman & Majid, 2019). The mapping of grammatical positions (e.g., subject, object) onto thematic roles (e.g., an action's agent, patient) is variable across syntactic structures: in an active sentence like "the pilot punched the chef", the mapping is subject (pilot)=agent, object (chef)=patient; but in a passive sentence, like "the pilot was punched by the chef", the roles are reversed: the grammatical subject is still the pilot, but it

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

is now the patient being punched. Thus, two sentences with different constructions (active vs. passive) can have the same thematic role assignments. Alternatively, two sentences with the same construction can have opposite assignments, e.g., "the pilot punched the chef" vs. "the chef punched the pilot". Therefore, inferring who did what to whom requires combining several types of syntactic information (linear order, construction), at least in the absence of prior knowledge about which noun is a more plausible agent (Caramazza & Zurif, 1976; Mahowald et al., 2023).

Thematic role assignment is an important component in psycholinguistic theories (for a review, see: Rissman & Majid, 2019), unlike many tasks used to test LLMs, such as logical entailment or common-sense reasoning. However, it appears to be dissociable from syntactic processing per se (Caramazza & Miceli, 1991; Chatterjee et al., 1995), so evidence for syntactic abilities in LLMs does not trivially predict successful thematic role assignment. In the human brain, thematic role assignment engages the Core Language Network (Ivanova et al., 2021), a system selective for high-level language processing (Fedorenko et al., 2011, 2024), including the extraction of sentence meaning (Fedorenko et al., 2016). Whereas this process also recruits a-modal regions outside the Core Language Network, their involvement relies on task demands because those regions are overall not sensitive to linguistic meaning (Ivanova, 2022; see also Frankland & Greene, 2020; Wang et al., 2016). Thus, understanding "who did what to whom" in a sentence appears to predominately rely on *linguistic* computations, and thus provides an appropriate test for LLMs.

In two experiments, we test whether training LLMs on word prediction results in representations that reflect thematic roles. We use LLMs that are pre-trained on this objective, without any further fine-tuning on other objectives. Four matters about this rationale are worth emphasizing. First, LLMs can be directly trained on thematic role assignment via supervised learning. However, our question is whether the broader objective of word prediction suffices for this purpose. Word prediction is the consensus objective for (pre-)training LLMs and such LLMs are treated as "general language processors" that are compared to human behavior and brain activity (e.g., Schrimpf et al., 2021).

Second, many state-of-the-art LLMs, like ChatGPT (OpenAI, 2022) or GPT4 (OpenAI, 2023), are not only (pre)trained on word prediction but are also fine-tuned using "reinforcement learning from human feedback" (RLHF; Christiano et al., 2017; Ouyang et al., 2022). In RLHF, an LLM receives input about human preferences and learns to align its responses with those preferences. Such input about human preferences constitutes non-linguistic information, so LLMs trained with RLHF are outside the scope of this work.

Third, prior work has demonstrated that LLMs represent thematic role information (e.g., Tenney, Das, et al., 2019; Tenney, Xia, et al., 2019). Yet, this ability is often tested using corpora derived from natural text (e.g., Carreras & Màrquez, 2005; Pradhan et al., 2013), which likely contains few challenging examples: in most sentences, thematic roles might be assigned based on heuristics (Mahowald et al., 2023). We instead specifically design stimuli that are less susceptible to heuristics: "reversible" sentences where both agent and patient are equally likely to produce an action, and (in Experiment 2) in a wide range of constructions. We chose this approach because other linguistic capacities that seem robust when tested on corpora "from the wild" can break down when tested on carefully crafted stimuli (e.g., Chaves & Richter, 2021; Glockner et al., 2018; McCoy et al., 2019; Rosenman et al., 2020; Sinha et al., 2021).

Fourth, testing whether LLMs map syntax onto thematic roles assumes that a syntactic representation is available to LLMs. However, syntactic processing in LLMs still falls short of humans (Marvin & Linzen, 2018); and even when LLMs do capture the structure of sentences, they might rely on "tricks" that differ from the rich, systematic linguistic principles guiding humans (Chaves & Richter, 2021; McCoy et al., 2019; Sinha et al., 2021). Still, such characterizations of LLMs are often based on sentences with quite complex structures. There is wide implicit agreement that LLMs do capture the structure of simple sentences like "the pilot punched the chef", which are the stimuli used in Experiment 1. Additionally, LLMs appear to be able to represent syntactic information "when it matters" (Papadimitriou et al., 2022).

Our approach is therefore appropriate for testing whether LLMs implicitly assign roles like "agent" and "patient" to event participants. All our analyses rely on the following paradigm: we feed sentences to LLMs, extract the resulting representations—i.e., activity patterns in hidden layers—and quantitively characterize to what extent they are influenced by thematic role information. Specifically, we generate sentences that either (i) share thematic role assignments but differ in syntax, or (ii) have opposite thematic role assignments but share syntax. We test whether sentence pairs of type (i) are more similar to one another than pairs of type (ii), which would be expected if word prediction suffices for LLMs to learn something akin to thematic roles. This is akin to representational similarity analyses in fMRI (Kriegeskorte et al., 2008; Norman et al., 2006). We also test to what extent human similarity ratings of the same sentence pairs reflect thematic role assignments. If LLMs are good models of human language processing, then the influence of thematic role information on their representations should be as strong as its influence on human judgments.

Experiment 1

Methods

Stimuli. Stimuli were based on Fedorenko et al., (2020) and generated in two steps: First, we created 94 "base" active, transitive sentences describing a two participant event, such as "*the lawyer saved the author*". Then, we edited each base sentence to create four versions that changed its thematic role assignments (and hence the meaning) and/or its structure: (A)

same meaning, same structure (SEM_s-$SYNT_s$): a "control" version where nouns and verbs are replaced by near-synonyms, maintaining the active structure and the thematic role assignments (*the attorney rescued the writer*); (B) same meaning, different structure (SEM_s-$SYNT_d$): the sentence is converted to passive while maintaining its base words and thematic role assignments (*the author was saved by the lawyer)*; (C) different meaning, same structure (SEM_d-$SYNT_s$): the agent and the patient are swapped, thus changing thematic role assignments while maintaining the base words and active structure (*the author saved the lawyer*); and (D) different meaning, different structure (SEM_d-$SYNT_d$): another "control" version with different words (near-synonyms), reverse thematic role assignments, and a passive structure, i.e., a sentence that is maximally different from the base (*the attorney was rescued by the writer*).

If LLMs represent sentence structure and use it to assign thematic roles, they would represent sentence pairs with the same thematic role assignments as more similar to one another than pairs with opposite assignments: the base sentence would be most similar to condition SEM_s-$SYNT_s$, followed by SEM_s-$SYNT_d$, then SEM_d-$SYNT_s$, then SEM_d-$SYNT_d$. The critical comparison is between the similarity of the base to SEM_s-$SYNT_d$ vs. its similarity to SEM_d-$SYNT_s$. If, however, LLMs fail to infer event meaning, similarities would only / mostly reflect whether sentences share structure, regardless of "who did what to whom" (i.e., similarity of the base to SEM_d-$SYNT_s$ would be higher than to SEM_s-$SYNT_d$).

Large language models. We used BERT and GPT-2[1] (Devlin et al., 2018; Radford et al., 2019) as implemented in HuggingFace. These LLMs are frequently studied and, unlike more recent LLMs, their hidden representations are accessible. The versions of these LLMs we studied each have 12 layers, each with 768 hidden units and 12 attention heads.

Evaluating representational similarities. For each layer in each LLM, we extracted a representation of each sentence: this was the distributed patterns of activity across hidden units for the [CLS] token in BERT, and the '.' token in GPT-2 (Schrimpf et al., 2021). Then, for each sentence set, we compared the representation of the base sentence and each other condition via the cosine similarity measure. Because such similarities might be influenced by a small subset of units with, e.g., very high activations across all sentences (Timkey & van Schijndel, 2021), we first normalized each hidden unit's activations relative to that unit's average and standard deviation across a large set of sentences (COCA; Davies, 2009). Similarities were Fisher-transformed to render their distribution closer to Gaussian and ameliorate bias in averaging them across stimuli (Silver & Dunlap, 1987). We contrasted the four conditions in terms of their similarity to the base sentence using a non-parametric, one-way repeated-measures ANOVA (Friedman test). Significant results were followed by pairwise post-hoc, two-tailed Wilcoxon signed rank tests (Bonferroni corrected).

Behavioral judgments. Human judgments about our stimuli are required as a standard against which to evaluate LLMs (Arana et al., 2023). It remains unclear how automatically and accurately we can infer thematic roles based on grammatical cues alone (versus e.g. plausibility cues) without careful contemplation or explicit instructions to closely attend to sentence structure (Ferreira & Lowder, 2016).

We collected behavioral judgments from 120 participants, recruited via UCLA's participant recruitment system (n=3 removed due to missing responses). The study was approved by UCLA's Institutional Review Board.

In an online experiment, participants rated pairs of sentences for their similarity, using a sliding scale between 1 (completely different) and 100 (identical). To minimize the chances that participants detect the distinctions between our conditions and use an artificial strategy for solving the task, each participant made only one judgment per condition (i.e., rated the similarity between a single sentence from that condition and its corresponding base sentence). Stimuli across the four conditions came from distinct sets (no base sentence was read more than once by a participant). We created 24 experimental lists, each consisting of 4 sentence pairs and shown to 5 participants. To mask the purpose of the study, these pairs were interleaved among 5 other pairs where similarity did not require close attention to sentence structure and event roles, and could instead be derived from general common sense (e.g., sentences on a shared topic vs. distinct topics). We z-scored similarities within each participant.

Both humans and LLMs may succeed in this task. Alternatively, LLMs and humans might err in similar ways, with both failing to reliably assign thematic roles. But if machines and humans diverge in their performance patterns, it would suggest that linguistic representations in LLMs are, in some crucial ways, different from those in human minds.

Results

We report results for the last layer in each model, but the critical findings hold across layers. Below, we use "similar" to mean "similar to one another".

BERT. We found a significant difference between cosine similarities across conditions ($\chi^2_{(3)}$=223.0, $p<10^{-47}$; Figure 1A). Post-hoc tests revealed that: (1) the two control conditions differed from one another as expected (SEM_s-$SYNT_s$ > SEM_d-$SYNT_d$, z=5.11, $p=10^{-5}$); (2) a pair with different syntax but shared meaning was more similar than the maximally different control pair (SEM_s-$SYNT_d$ > SEM_d-

[1] Analysis of more recent, larger models, Llama 2-7B and Persimmon-8B yielded qualitatively the same pattern of results.

$SYNT_d$, (z=8.34, $p = 10^{-15}$) but, surprisingly, also more similar than the maximally similar control pair (SEM_s-$SYNT_d$ > SEM_s-$SYNT_s$, z=6.34, $p = 10^{-8}$); (3) a pair with the same syntax but different meaning was surprisingly more similar than the maximally similar pair (SEM_d-$SYNT_s$ > SEM_s-$SYNT_s$, z=8.41, $p = 10^{-15}$) but, as expected, also more similar than the maximally different pair (SEM_d-$SYNT_s$ > SEM_d-$SYNT_d$, z=z=8.41, $p = 10^{-15}$); and (4) most critically, a pair with the same meaning (but different syntax) was *less* similar than a pair with opposite meanings (but the same syntax) (SEM_s-$SYNT_d$ < SEM_d-$SYNT_s$, z=8.34, $p = 10^{-15}$). Thus, syntax exerts a stronger influence on BERT representations than thematic roles do.

GPT-2. We found a significant difference between cosine similarities across conditions ($\chi^2_{(3)}$=243.17, $p<10^{-51}$; Figure 1B). Post-hoc tests revealed that: (1) the two control conditions differed from one another as expected (z=4.50, $p<10^{-4}$); (2) a pair with different syntax but shared meaning was more similar than the maximally different control pair (SEM_s-$SYNT_d$ > SEM_d-$SYNT_d$, z=8.33, $p<10^{-16}$) but, surprisingly, also more similar than the maximally similar control pair (SEM_s-$SYNT_d$ > SEM_s-$SYNT_s$, z=8.19, $p<10^{-14}$); (3) a pair with the same syntax but different meaning was surprisingly more similar than the maximally similar pair (SEM_d-$SYNT_s$ > SEM_s-$SYNT_s$, z=8.33, $p<10^{-15}$) but, as expected, also more similar than the maximally different pair (SEM_d-$SYNT_s$ > SEM_d-$SYNT_d$, z=8.32, $p<10^{-15}$); and (4) most critically, a pair with the same meaning (but different syntax) was *less* similar than a pair with opposite meanings (but the same syntax) (SEM_s-$SYNT_d$ < SEM_d-$SYNT_s$, z=8.23, $p<10^{-14}$). Thus, syntax exerts a stronger influence on GPT-2 representations than thematic roles do.

Human Judgments. We attempted to fit a linear, mixed-effects model predicting sentence similarity from condition with random intercepts by participant and/or stimulus set. These models did not converge and showed little variance across participants (due to z-scoring) and across sets. We therefore ran a fixed-effects model predicting similarity between sentences from condition. This model had an adjusted R^2 of 0.25, $F_{(4,458)}$=40.01, p<.001 (Figure 1C). Post-hoc tests found that: (1) the two control conditions differed from one another in the expected direction (z=7.83, $p<10^{-12}$) (2) a pair with the same meaning but different syntax was more similar than the maximally different control pair (SEM_s-$SYNT_d$ > SEM_d-$SYNT_d$, z=8.13, $p<10^{-15}$), and did not differ from the maximally similar control pair (SEM_s-$SYNT_d$ vs. SEM_d-$SYNT_d$, z=0.29, p=1); (3) unlike in LLMs, but consistent with a strong influence of thematic roles, a pair with different meanings but shared syntax was less similar than the maximally similar pair (SEM_d-$SYNT_s$ < SEM_s-$SYNT_s$, z=8.96, $p<10^{-15}$). Despite sharing syntax, this pair did not differ from the maximally different pair (SEM_d-$SYNT_s$ vs. SEM_d-$SYNT_d$, z=1.15, p=1); and (4) most critically, a pair with the same meaning (but different syntax) was *more* similar than a pair with opposite meanings (but the same syntax) (SEM_s-$SYNT_d$ > SEM_d-$SYNT_s$, z=9.27, $p<10^{-15}$). This pattern is the opposite of what was found for LLMs, and demonstrates that thematic roles exert a stronger influence on human similarity judgments than syntax does.

We did not directly compare human to LLM data, because the similarity judgments in these two datasets are on different scales (Likert vs. cosine). However, we analyzed LLM data for the subset of 24 stimulus sets for which we collected behavioral data, and the comparison between the two critical conditions was still significant (p<.001 for both models). Even without a direct comparison, we emphasize that human similarity judgments are governed by thematic role assignments and go in the opposite direction from LLM representational similarities, which are governed by syntax.

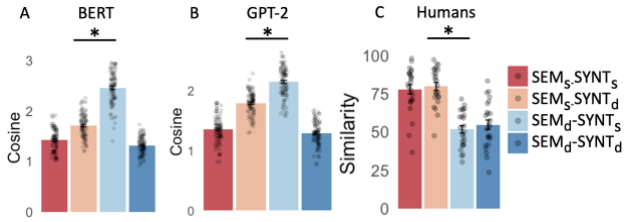

Figure 1: Similarity of the base sentences to sentences in each condition for (A) BERT, (B) GPT-2, and (C) humans. For LLMs, similarities are Fisher-transformed. Each dot represents one item (panel C averages across the five participants who saw each item).

Experiment 2

Any cognitively plausible representation of sentences should feature thematic roles as a main component; for this reason, Experiment 1 quantified LLM representations as distributed activity patterns across all hidden units. However, thematic roles might instead be encoded by a small subset of units (with others representing unrelated information, e.g., lexical semantics). Thus, although LLMs do not emphasize thematic roles, perhaps they still extract this information, i.e., possess similar representational capacities to those of humans.

Experiment 2 thus asked: are thematic roles represented *anywhere* among LLM units, even in a small subset of them? We extracted activity patterns across all units for pairs of sentences and fed them to an algorithm that tried to find any information, in any set of units, that could classify whether that pair shared common thematic role assignments or not. We also asked whether thematic role information was available in components of LLMs other than hidden units, namely, in the attention heads. We thus extracted attention weights between content words from pairs of sentences and followed the same classification procedure.

Methods

Stimuli. Classifying whether two sentences share common thematic role assignments cannot be done with sentences

from Experiment 1, where a solution might rely on simple "tricks" due to the limited number of syntactic structures. Therefore, we created ditransitive sentences with a variety of structures (e.g., "*the man gave the milk to the woman*", *"it was the woman that was given the milk by the man*"), where no global "trick" can infer whether sentences share meaning or not. A single stimulus set used 12 structures, each with the two versions having opposite agent-patient assignments, for a total of 24 sentences. Structures varied in whether they were active or passive, double- or prepositional-object, and had or did not have a cleft. We generated 50 sets of these 24 sentences, for a total of 1,200 sentences.

Evaluating LLM representations. Prior to training a classifier on pairs of sentences, we performed the same analysis as in Experiment 1, computing cosine similarities for every pair of sentences within each stimulus set. We tested whether similarities for pairs that shared a meaning was higher than for pairs with different meanings, and split the analysis by whether their respective structures differed in 0, 1, 2, or 3 of the syntactic features described above.[2]

For the main analysis, we trained a support vector machine (SVM) to distinguish between "same meaning" vs. "different meaning" sentence pairs based on their distributed representations (each layer analyzed separately). In one analysis the representations of the two sentences were concatenated; in another, they were subtracted. In a 66-fold cross-validation, a separate SVM was trained on each combination of 20 of the 24 sentence structures, holding out the two versions of each of the remaining 2 structures for testing. Training excluded pairs consisting of two versions of the same structure, as they always had different meanings.

Behavioral judgments. 120 participants were recruited online through UCLA's participant recruitment system. They rated pairs of sentences for their similarity as in Experiment 1. Each participant judged only two critical pairs of sentences that either had the same or different thematic role assignments. These two trials were from different sets. Filler trials were included as in Experiment 1. Due to a coding error, we only sampled 12 out of the 50 stimulus sets.

Evaluating attention heads. For each attention head in each layer, we studied attention patterns assigned between entities in each sentence. The relative position of words varied across sentence structures, which sometimes necessitated "forward-looking" attention (i.e., from a previous to a future word). However, attention in GPT-2 is only backward-looking, so our analysis was limited to BERT, which has bidirectional attention. We extracted each sentence's attention weights between every pair of the following words: subject (which, depending on sentence structure, was the agent or patient), indirect object (patient or agent), verb, and direct object; we excluded attention from the verb to the direct object and vice versa because these involved neither agent nor patient. Using these vectors of 10 attention weights per sentence, the same SVM analyses described above were conducted to classify sentences pairs with same vs. different meaning.

Results

BERT. We found limited evidence of robust representation of thematic roles: for sentence pairs differing in one syntactic feature, pairs with the same meaning were *more* similar than pairs with different meanings (z=2.42, p<.05). However, pairs with no differences in syntactic features showed the opposite pattern (z=15.17, p<.001). For pairs differing in 2 or 3 features, similarity did not differ as a function of whether the pair had the same meaning (z=0.71, p=.47; z=.85, p=.39).

GPT-2. We found limited evidence of robust representation of thematic roles: for sentence pairs differing in one syntactic feature, pairs with the same meaning were *more* similar than pairs with different meanings (z=10.21, p<.001). However, pairs with no differences in syntactic features, as well as pairs different in 2 or 3 features, showed the opposite pattern (z=12.79, p<.001; z=8.17, p<.001; z=10.29, p<.001).

SVM. Figure 2 shows the SVM accuracies across layers for SVMs trained on either concatenated or subtracted representations of sentence pairs. Most SVMs fail to reach significance above 50% chance level except, notably, when trained on subtracted GPT-2 activations. In one case, an SVM reached above 60% accuracy (layer 5).

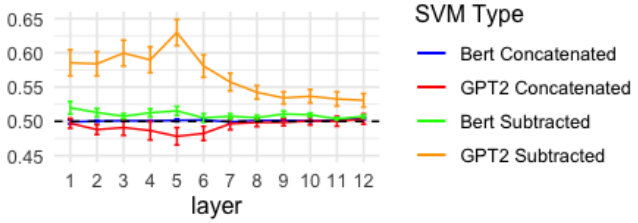

Figure 2: Classification accuracies for SVMs predicting whether two sentences had the same vs. different meanings. SVMs were trained per layer of BERT and GPT-2, on representations of pairs of sentences that were either concatenated or subtracted.

Human Judgments. We attempted to fit a linear, mixed-effects model predicting sentence similarity from condition (same vs. different meaning) with random intercepts by participant and/or stimulus set, but encountered the same issues as in Experiment 1. Therefore, we ran a fixed-effects

[2] An example pair of sentences with 0 changes and the same thematic role assignments: "*it was the lawyer who assigned the architect the homework*" and "*it was the homework that the lawyer assigned the architect*" (both are active + direct object + cleft).

model: adjusted R^2=0.299, $F_{(2,228)}$=48.63, p<.001. Specifically, sentence pairs that shared meaning were rated as more similar than pairs that did not (z=4.391, p<.001).

Attention heads. SVM classification of sentence pairs (same vs. different meaning) had high accuracy for several attention heads. We highlight here head 5 in layer 11, which had an accuracy of 79% (when concatenating attention weights for a pair of sentences). To characterize this head's function (Figure 3), we contrasted its attention patterns (1) from the verb to the agent vs. patient; (2) from the direct object to the agent vs. patient; and (3) from agent to patient vs. vice versa. Each comparison was carried out in a linear, mixed-effects model. Attention weights, which are restricted between [0,1], were logit-transformed (this did not change the results) and modeled with a fixed effect of direction (towards the agent vs. patient) and random intercepts and slopes by stimulus set and by structure. The verb ($t_{(51.48)}$=13.76, p<10^{-16}) and direct object ($t_{(36.48)}$=5.18, p<10^{-5}) both allocated more attention to the agent than to the patient, and patients directed more attention to agents than vice versa ($t_{(35.60)}$=7.76, p<10^{-8}). These patterns held across most sentence structures regardless of the grammatical positions of agents and patients, so they reflect thematic roles, not syntax.

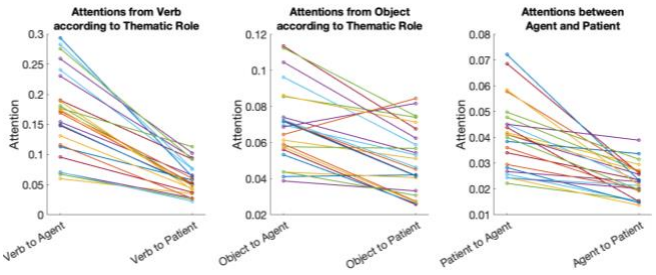

Figure 3: Attention patterns for BERT's head 5 in layer 11 for each of the 24 sentence types, averaged across stimulus sets.

Notably, the classification accuracy of this attention head exceeds humans: only 56.9% of participants rated sentence pairs that shared thematic role assignments as more similar than pairs that did not (and 11.7% rated both pairs equally). On the sentence structures that participants viewed, this attention head was 83% accurate (compared to an average accuracy of 49.79% across all other attention heads).

Discussion

This study asked whether LLMs understand sentences in the minimal sense of representing "who did what to whom". In Experiment 1, we found that the overall geometry of LLM distributed activity patterns failed to capture this information, as similarities between sentences reflected whether they shared syntax more than whether they shared thematic role assignments. Human judgments, in contrast, were strongly driven by this aspect of meaning. In Experiment 2, we found limited evidence that thematic role information was available even in a subset of hidden units (with one exception). However, it was available in some attention heads, even for sentences that human participants struggled with.

These results are important because event semantics are tightly linked to understanding linguistic input as such, unlike aspects of comprehension like common-sense reasoning or logical inference which, while frequently studied in LLMs, reflect non-linguistic thinking (Mahowald, Ivanova, et al., 2024). Even in our relatively simple task of mapping sentence structure onto thematic roles, LLMs do not give meaning the prominent role that humans do, despite possessing the capacity to extract this information. Training LLMs on word prediction is sufficient for learning what thematic roles are, but perhaps not for representing them in a human-like way.

Our findings are consistent with the broader claim that the success of LLMs in syntactic processing does not guarantee similar success in semantic processing (Weissweiler et al., 2022). Indeed, despite the impressive syntactic capabilities of LLMs (e.g., Manning et al., 2020; McCoy et al., 2023; Wilcox et al., 2021; for a review, see: Linzen & Baroni, 2021) prior work has demonstrated that LLMs trained on word prediction alone have limited understanding: they struggle with tracking the state of entities in a text (Kim & Schuster, 2023), sometimes refer to entities that do not exist (Schuster & Linzen, 2022), and make predictions that are only weakly sensitive to event roles (Ettinger, 2020).

As stated in the introduction, LLMs can perform thematic role assignment if they are directly fine-tuned on this task. However, our study asked whether robust representations of thematic roles could result from training on word prediction exclusively, to understand how a general linguistic objective affects LLM capabilities. Our work thus complements the existing literature characterizing models that lack fine-tuning (for a review, see: Chang & Bergen, 2023). Probing such models is crucial as they are the ones that are commonly compared to human behavior and brain activity (e.g., Schrimpf et al., 2021). For instance, our findings suggest that whatever brain activity can be predicted from the hidden activations of LLMs, it does reflect thematic roles.

Our study used a specific similarity metric (cosine) and rather simple classifiers. It is possible that a different metric, or an algorithm that can learn more complex classifiers, could detect a stronger representation of thematic roles in LLMs. We leave such investigations for future work. We also do not wish to suggest that LLMs could never represent thematic roles in more human-like ways (which might be the case for larger models, other architectures, other training corpora, or models exposed to reinforcement learning from human feedback). Nonetheless, our findings emphasize that it is vital to test the ability of LLMs to understand *language* (cf. exhibit thinking) using carefully crafted materials inspired by psycholinguistics, in order to ensure that the seemingly meaningful text that LLMs generate reflects comprehension rather than non-linguistic "tricks" that have little to do with human language processing (McCoy et al., 2019).

References

Arana, S., Hagoort, P., Schoffelen, J.-M., & Rabovsky, M. (2023). Perceived similarity as a window into representations of integrated sentence meaning. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02129-x

Bender, E. M., & Koller, A. (2020). *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*. 5185–5198. https://doi.org/10.18653/V1/2020.ACL-MAIN.463

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). *Experience Grounds Language* (arXiv:2004.10151). arXiv. https://doi.org/10.48550/arXiv.2004.10151

Caramazza, A., & Miceli, G. (1991). Selective impairment of thematic role assignment in sentence processing. *Brain and Language*, *41*(3), 402–436. https://doi.org/10.1016/0093-934X(91)90164-V

Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language*, *3*(4), 572–582. https://doi.org/10.1016/0093-934X(76)90048-1

Carreras, X., & Màrquez, L. (2005). *Introduction to the CoNLL-2005 shared task: Semantic role labeling*. 152–164.

Chang, T. A., & Bergen, B. K. (2023). *Language Model Behavior: A Comprehensive Survey*.

Chatterjee, A., Maher, L. M., Rothi, L. J. G., & Heilman, K. M. (1995). Asyntactic Thematic Role Assignment: The Use of a Temporal-Spatial Strategy. *Brain and Language*, *49*(2), 125–139. https://doi.org/10.1006/brln.1995.1024

Chaves, R. P., & Richter, S. N. (2021). Look at that! BERT can be easily distracted from paying attention to morphosyntax. *Proceedings of the Society for Computation in Linguistics*, *4*(1), 28–38.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cognitive Science*, *47*(3), e13256. https://doi.org/10.1111/cogs.13256

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48.

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, *108*(39), 16428–16433.

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *203*, 104348–104348. https://doi.org/10.1016/J.COGNITION.2020.104348

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 1–24.

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, *113*(41), 6256–6262.

Fedorenko, E., & Varley, R. (2016). Language and thought are not the same thing: Evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, *1369*(1), 132–153. https://doi.org/10.1111/NYAS.13046

Frankland, S. M., & Greene, J. D. (2020). Two ways to build a thought: Distinct forms of compositional semantic representation across brain regions. *Cerebral Cortex*, *30*(6), 3838–3855.

Glockner, M., Shwartz, V., & Goldberg, Y. (2018). Breaking NLI systems with sentences that require simple lexical inferences. *arXiv Preprint arXiv:1805.02266*.

Ivanova, A. A. (2022). *The role of language in broader human cognition: Evidence from neuroscience*.

Jackendoff, R. (2012). *A user's guide to thought and meaning*. OUP Oxford.

Kim, N., & Schuster, S. (2023). *Entity Tracking in Language Models* (arXiv:2305.02363). arXiv. http://arxiv.org/abs/2305.02363

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.

Mahowald, K., Diachek, E., Gibson, E., Fedorenko, E., & Futrell, R. (2023). Grammatical cues to subjecthood are redundant in a majority of simple clauses across languages. *Cognition*, *241*, 105543. https://doi.org/10.1016/j.cognition.2023.105543

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv Preprint arXiv:1808.09031*.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv Preprint arXiv:1902.01007*.

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. (2023). How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN. *Transactions of the Association for Computational Linguistics*, *11*, 652–670. https://doi.org/10.1162/tacl_a_00567

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

OpenAI. (2023). *GPT-4 Technical Report*. https://arxiv.org/abs/2303.08774v3

OpenAI, T. B. (2022). Chatgpt: Optimizing language models for dialogue. *OpenAI*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Papadimitriou, I., Futrell, R., & Mahowald, K. (2022). When classifying grammatical role, BERT doesn't care about word order... Except when it matters. *arXiv Preprint arXiv:2203.06204*.

Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics*, *8*, 447–471.

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, Lingbuzz*, *7180*.

Piantadosi, S. T., & Hill, F. (2022). *Meaning without reference in large language models* (arXiv:2208.02957). arXiv. https://doi.org/10.48550/arXiv.2208.02957

Pradhan, S., Moschitti, A., Xue, N., Ng, H. T., Björkelund, A., Uryupina, O., Zhang, Y., & Zhong, Z. (2013). *Towards robust linguistic analysis using ontonotes*. 143–152.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9–9.

Regev, T. I., Casto, C., Hosseini, E. A., Adamek, M., Brunner, P., & Fedorenko, E. (2022). Intracranial recordings reveal three distinct neural response patterns in the language network. *bioRxiv*, 2022–12.

Rissman, L., & Majid, A. (2019). Thematic roles: Core knowledge or linguistic construct? *Psychonomic Bulletin & Review*, *26*(6), 1850–1869.

Rosenman, S., Jacovi, A., & Goldberg, Y. (2020). Exposing shallow heuristics of relation extraction models with challenge data. *arXiv Preprint arXiv:2010.03656*.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Schuster, S., & Linzen, T. (2022). *When a sentence does not introduce a discourse entity, Transformer-based models still sometimes refer to it* (arXiv:2205.03472). arXiv. http://arxiv.org/abs/2205.03472

Sinha, K., Jia, R., Hupkes, D., Pineau, J., Williams, A., & Kiela, D. (2021). Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv Preprint arXiv:2104.06644*.

Sinha, K., Sodhani, S., Dong, J., Pineau, J., & Hamilton, W. L. (2019). CLUTRR: A diagnostic benchmark for inductive reasoning from text. *arXiv Preprint arXiv:1908.06177*.

Tenney, I., Das, D., & Pavlick, E. (2019). *BERT Rediscovers the Classical NLP Pipeline* (arXiv:1905.05950). arXiv. https://doi.org/10.48550/arXiv.1905.05950

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., & Pavlick, E. (2019). *What do you learn from context? Probing for sentence structure in contextualized word representations* (arXiv:1905.06316). arXiv. https://doi.org/10.48550/arXiv.1905.06316

Timkey, W., & van Schijndel, M. (2021). All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. *arXiv Preprint arXiv:2109.04404*.

Ullman, T. (2023). *Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks* (arXiv:2302.08399). arXiv. https://doi.org/10.48550/arXiv.2302.08399

Wang, J., Cherkassky, V. L., Yang, Y., Chang, K. K., Vargas, R., Diana, N., & Just, M. A. (2016). Identifying thematic roles from neural representations measured by functional magnetic resonance imaging. *Cognitive Neuropsychology*, *33*(3–4), 257–264. https://doi.org/10.1080/02643294.2016.1182480

Weissweiler, L., Hofmann, V., Köksal, A., & Schütze, H. (2022). *The Better Your Syntax, the Better Your Semantics? Probing Pretrained Language Models for the English Comparative Correlative* (arXiv:2210.13181). arXiv. https://doi.org/10.48550/arXiv.2210.13181

Wilcox, E. G., Vani, P., & Levy, R. P. (2021). A Targeted Assessment of Incremental Processing in Neural LanguageModels and Humans. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 939–952. https://doi.org/10.18653/v1/2021.acl-long.76

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., & Kim, Y. (2023). *Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks* (arXiv:2307.02477). arXiv. https://doi.org/10.48550/arXiv.2307.02477