Evidence Against Syntactic Encapsulation in Large Language Models

Thomas McGee (thomasmcgee@g.ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095 United States

Idan Blank (iblank@psych.ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095 United States

Abstract

Transformer-based large language models (LLMs) have recently demonstrated exceptional performance in a variety of linguistic tasks. The main mechanism through which LLMs combine words in a sentence is called "attention heads": these components assign numerical weights linking different words in the input to one another, capturing different relationships between these words. Some attention heads automatically learn to assign weights that accurately encode meaningful linguistic features including, importantly, heads that appear specialized for identifying particular syntactic dependencies. Are syntactic computations in such heads "encapsulated", i.e., impenetrable to the influence of non-syntactic information? Such encapsulated 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, and where syntax and semantics are tightly linked in the mental lexicon. Here, we tested whether the activity of syntax-specialized attention heads in transformer-based LLMs is modulated by one type of semantic information: plausibility. In each of two LLMs (BERT and GPT-2), we first identified attention heads specialized for various dependency types; in six of the seven cases tested, we found that implausible semantic information reduces attention between the words that constitute the dependency for which a head is specialized. These results demonstrate that, even in attention heads that are the best candidates for syntactic encapsulation, syntactic information is penetrable to semantics. These data are broadly consistent with the integration of syntax and semantics in human minds.

Keywords: Modularity; syntax; semantics; sentence processing; artificial neural networks; large language models

Introduction

In recent years, transformer-based large language models (LLMs) have achieved remarkable performance in a variety of language tasks (e.g., Brown et al., 2020; Chang & Bergen, 2023; Contreras Kallens, Kristensen-McLachlan, & Christiansen, 2023; OpenAI, 2022; Piantadosi, 2023; Vaswani et al., 2017). The text that LLMs generate, as well as their internal representations, indicate that they have mastered many (but not all) non-trivial syntactic abstractions that underlie the structure of sentences (e.g., Manning et al., 2020; McCoy et al., 2023; Wilcox, Vani, & Levy, 2021; for a review, see: Linzen & Baroni, 2021). To the extent that LLMs have human-like syntactic knowledge, they may constitute good models of human language processing. However, this latter claim depends on whether LLMs, beyond "having syntactic knowledge" in the broad sense, also align with humans in the finer details of how they represent and process syntax (e.g., Arehalli, Dillon, & Linzen, 2022; van Schijndel & Linzen, 2021).

One of the central properties of human syntactic processing is that it is rapidly influenced by external information sources, such as semantics or visual referents (McRae, Spivey-Knowlton, & Tanenhaus, 1998; Tanenhaus et al., 1995; Trueswell, Tanenhaus, & Garnsey, 1994). In other words, syntax is not *encapsulated*. Rather than proceeding independently of other processes, the parser appears to be penetrable to various non-syntactic information sources, and it opportunistically uses them as soon as they become available (or, at least, as soon as we can measure). The issue of encapsulation is fundamental because it is one of the main characteristics of cognitive modules[1] (Fodor, 1983), and whether a processing domain (such as syntax) is modular is a basic architectural property of a computing system.

To demonstrate the distinction between a modular and a non-modular syntactic parser, consider a sentence "We met the owner of the company that was sold yesterday". From a syntactic point of view—ignoring word meaning—the

[1] Encapsulation is often confused with the notion of domain-specificity, but these are two different properties. A system can be domain-specific—that is, exclusively process and output representations in a particular domain—while still being penetrable to external systems that can influence its operations (several systems in the human mind appear to behave this way).

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.*

sentence is structurally ambiguous, because the clause "that was sold yesterday" can describe either the owner or the company (this may be clearer in the sentence "we met the owner of the company that everyone knows", where the clause "that everyone knows" is semantically compatible with the owner or the company). However, this ambiguity can be resolved using semantic information: owners (animate entities) are rarely sold, but companies are often sold. Hence, the clause "that was sold yesterday" likely describes the company. A parser that is initially encapsulated would be unable to use this semantic information in its early processing stages to resolve the structural ambiguity (just like in the sentence with "that everyone knows"). In contrast, a parser that is penetrable from the earliest moments would not encounter any ambiguity, immediately determining the correct structure.

One might argue that syntactic processing in LLMs is encapsulated because these models can generate next-word predictions that obey the rules of syntax even for nonsense sentences that are "devoid of semantics" (e.g., "the colorless green ideas that I ate with the chair *sleep*" rather than *sleeps*) (Gulordava et al., 2018). Similar reasoning has been applied to the human mind: given that we can judge the grammaticality of such nonsense sentences, the mental parser must be encapsulated. However, such conclusions about encapsulation do not logically follow: even if a parser *were* penetrable to, e.g., lexico-semantic information, it is reasonable to assume that it could still process syntactic features in the absence of coherent meaning. In other words, a parser being penetrable to semantics does not entail that semantics are necessary for its proper operation.

Moreover, linguistic analysis and behavioral evidence from humans suggest that abstract syntactic constructions are not devoid of meaning, but rather are associated with abstract semantic meaning (e.g., "colorless green ideas sleep furiously" has a structure that roughly means "things with a specific type of a certain property do something in a particular way") (Goldberg, 1995, 2019; see also Bencini & Goldberg, 2000; Casenhiser & Goldberg, 2005; Johnson & Goldberg, 2013). Similarly, studies of LLMs suggest that their internal representations of phrase-level syntactic structure are associated with meaning (Li et al., 2022). In addition, the quality of syntactic representations in LLMs is, to some extent, reliant on semantics: it deteriorates for nonsense sentences, as well as for "Jabberwocky" sentences in which content words are replaced with nonwords (Arps et al., 2023; Maudslay & Cotterell, 2021).

Encapsulation in Attention Heads?

It is not surprising to find syntactic representations that are penetrable to non-syntactic information *somewhere* within LLMs—during language processing, different information sources must eventually combine and interact with one another. But such penetrable representations in one part of a LLM might co-exist alongside encapsulated representations in other parts. If one were to look for syntactic encapsulation, what would be a good place to start?

To answer this question, consider the architecture of transformer-based LLMs. Like other types of LLMs, transformers represent every token (usually, a word) in a sentence in a distributed manner across hidden units, and this representation is gradually transformed as it passes through the model's layers. The main feature that distinguishes transformers from previous LLMs is that the transformation of a token's representation from one layer to the next is heavily influenced by components called "*attention heads*" (Vaswani et al., 2017). These heads capture how tokens in a text input relate to one another by assigning numerical weights from each token to other tokens, directing tokens to "attend" to one another. In bidirectional transformers, each token can attend to all tokens, both preceding and following it, as well as itself; in unidirectional transformers, each token can only attend to itself and all preceding tokens but not to those following it. Each token's representation is transformed into a "mix" of all the tokens it attends to via a weighted linear combination, wherein tokens that are attended to more strongly are represented more prominently. Each attention head assigns different patterns of weights across tokens, producing a distinct "mixed" representation. These representations are combined across all attention heads within a layer (and further elaborated upon).

Thus, the hidden layers themselves are perhaps expected to have representations where syntactic information is somewhat entangled with non-syntactic (e.g., semantic) information: as representations are transformed between a given layer *n* and the next layer *n+1*, tokens can attend to one another in diverse ways across attention heads, such that a token's resulting representation in layer *n+1* contains information about the representations of many tokens from layer *n*, for myriad reasons, both syntactic and otherwise. In contrast, we reasoned that *if* encapsulated syntactic computations existed anywhere within LLMs, the best candidates for such computations would be individual attention heads: each head learns to assign unique patterns of attention weights across tokens, and some such patterns might purely target syntactic relationships between tokens (but not other types of relationships).

Whereas attention heads can, in principle, learn to assign weights between tokens based on any feature of the input, prior work has demonstrated that, in practice, some heads learn to identify human-interpretable features. For instance, some heads encode positional information (e.g., attend most strongly to the previous token), some widely distribute attention over all tokens to create "bag of words" representations, some target specific parts of speech, etc. (Clark et al., 2019; Raganato & Tiedemann, 2018; Vig & Belinkov, 2019; Voita et al., 2019). Critically for our purposes, some attention heads appear "specialized" for identifying specific types of syntactic dependencies between words: one head targets the nominal subject of a verb,

another—the object of a preposition, etc. (Clark et al., 2019). Such heads assign more weight from a dependent to its head (or vice versa) than to any other token in a sentence. However, these attention heads show the pattern of attention described above with less than perfect accuracy (e.g., only in 77.38% of instances, on average, across the eight dependencies that have the most strongly specialized heads; Clark et al., 2019). Perhaps, then, such attention heads are influenced by factors other than syntax alone.

Therefore, we tested whether attention heads that are specialized for particular syntactic dependencies are encapsulated from, or penetrable to, a specific source of non-syntactic information: semantic plausibility. Would the computations of such attention heads be modulated by the plausibility of the sentence they are processing? To this end, we measured the amount of attention allocated by each head to its preferred dependency, across minimal pairs of sentences where that dependency was either semantically plausible (or likely) or implausible (or unlikely). If these heads are encapsulated, the strength of attention to their preferred dependency should not vary across these two types of sentences. If, however, the heads are penetrable to semantic information—as we predicted, based on parallels with the human mind—then attention weights for the preferred dependency would be lower in implausible compared to plausible sentences.

Methods

Model Description

We analyzed two transformer-based LLMs: the "base" sized BERT model (BERT-base-uncased) and the GPT-2 small model. Both models have 12 layers each containing 12 attention heads. The notation (head <layer>-<head_number>) will be used to indicate a particular head, e.g., head 7-9 corresponds to the ninth head in the seventh layer of a given model.

BERT is bidirectional, directing attention from a token to both preceding and following words. It is trained on masked word prediction (e.g., filling in "The galloping <MASK> fell down") and next sentence prediction. GPT-2 is unidirectional (i.e., backward-looking) and trained on next word prediction.

Identifying "Syntax-Specialized" Attention Heads

First, for each of 43 dependency types from the Stanford Dependencies (De Marneffe, MacCartney, & Manning, 2006), we searched for specialized attentions heads that accurately encoded that dependency. Our search procedure is a modified version of the one from Clark et al. (2019), who identified such heads in BERT. We extended the approach to GPT-2, where such heads have not been previously identified.

We annotated the development set (section 22) of the Penn Treebank 2 corpus (Clark et al., 2019; Marcus, Santorini, & Marcinkiewicz, 1993) with Stanford dependencies, using the Stanford CoreNLP toolbox. For each dependency and for each head, specialization was measured in two ways: (1) the percentage of instances of the dependency in the corpus for which the *dependent* token directed more attention to the *head* token than to any other token (dependent→head); and (2) the percentage of instances of the dependency in the corpus for which the *head* token directed more attention to the *dependent* token than to any other token (head→dependent).[2] For the indirect object dependency, we identified the specialized attention head using 203 sentences featuring the *iobj* dependency, which were selected from various folders of the Penn Treebank. This was necessary because there were relatively few instances of the *iobj* dependency in the original corpus used (folder 22). For GPT-2, we excluded instances that required forward-facing attention (e.g., measuring attention from a dependent to a head that comes later in the sentence), because attention in GPT-2 is only backward-looking.

Table 1: Heads specialized for different dependencies

BERT

Relation	Head	Direction	Spec	Base
dobj	7-9	dep->head	86.47	39.78 (-2)
nsubj	7-1	dep->head	60.89	46.46 (1)
pobj	7-10	head->dep	81.95	34.67 (2)
iobj	6-9	dep->head	77.07	46.34 (-1)

GPT-2

Relation	Head	Direction	Spec	Base
dobj	2-8	dep->head	81.6	40.6 (-2)
nsubj	4-3	head->dep	66.87	48.66 (-1)
pobj	2-0	dep->head	76.55	34.62 (-2)

Columns: "spec" = specialization score (%). "base" = baseline score (%). Note that the baseline scores are appreciably lower than the specialization scores.

For each dependency, the maximum specialization score was selected out of a set of 288 scores (12 layers × 12 heads × 2 measures) to identify a single head of interest for that dependency. However, note that some dependencies in the corpus could be identified with some accuracy by merely attending to tokens at a fixed distance (e.g., the object of a preposition is often two tokens to the right of the preposition, as in "up a tree", "on the table", etc.). Therefore, for each dependency type, we found the most common distance

[2] Measure (1) is identical to Clark et al. (2019), but measure (2) differs: for head→dependent, Clark et al. (2019) computed the percentage of instances for which the dependent token *received* more attention weight from the head than *from* any other word in the sentence. We modified this measure to make it analogous to measure (1) (dependent→head).

between tokens in that dependency across the corpus, and measured the percentage of instances exhibiting that distance. This "baseline" score (termed "fixed offset" in Clark et al., 2019) needed to be sufficiently below the maximum specialization score of a head of interest (based on subjective evaluation) for that head to be considered syntax-specialized.

Main Experiment

Materials We chose four dependency types for which we identified specialized heads (one per dependency) and for which we could create minimal sentence pairs for the main experiment, as described below. These dependencies are direct object (*dobj)*, indirect object (*iobj*), nominal subject (*nsubj*), and object of a preposition (*pobj*) (Table 1).

For each dependency, we constructed 50 minimal pairs of sentences such that in one sentence the phrase contained within the dependency was semantically plausible in the broader context of the sentence, and in the other—semantically implausible (Table 2). For example, for *dobj*, a plausible sentence would read "the guide *showed* the visitor a *sculpture*", and its implausible version would read "the guide *showed* the sculpture a *visitor*". All sentences for a given dependency had identical structure, and each word comprised a single token. For the *dobj*, *nsubj*, and *iobj* dependencies, plausibility was manipulated by swapping two words in the sentence with the same part of speech. For the *pobj* dependency, plausibility was manipulated by replacing the prepositional object with another noun.

BERT was tested on all 4 dependencies, whereas GPT-2 was tested on 3 (without *iobj*) due to (1) structural ambiguity at the point of the indirect object, which poses a problem for manipulating semantic plausibility for unidirectional models, and (2) the same head being identified as specialized for *iobj* and *dobj* in GPT-2 (perhaps reflecting this ambiguity). In addition, sentences for the *pobj* dependency differed between BERT and GPT-2: in BERT, the preposition came several tokens after the prepositional object. But because GPT-2 is unidirectional and the specialized head we found attended from the dependent—the object—to the preposition, this preposition had to occur *before* the object. Thus, *pobj* sentences differed in structure across the two LLMs, but they differed minimally in semantic content (BERT: "It was the **shoreline** by the rocks that the waves crashed **against**."; GPT-2: "The waves crashed **against** the strongly built **barrier** on the coast."). For other dependency types, there were slight differences in the stimuli across the two LLMs due to differences in tokenization: if a critical word was treated as a single token by BERT but broken to several tokens in GPT-2, it was replaced for the latter model by a single-token word. Identical sentences were used for the *dobj* and *iobj* dependencies within each model.

Importantly, 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 2). For instance, in the implausible sentence "the guide showed the sculpture a visitor", the word "sculpture" is not the direct object (*dobj*) of the verb in terms of *syntax*, but it is a good *semantic* candidate for such a dependency, because a sculpture is something that would be shown by a guide. We designed the sentences in this way to create an "attractor" that might direct attention away from the syntactic dependency.

Procedure For a given LLM and dependency type, we fed each sentence separately to the model and extracted the attention patterns from the head of interest. We analyzed the attention weights between critical words in our stimuli, contrasting the plausible and implausible sentences. Because attention weights are bounded in [0,1], we logit-transformed them prior to linear, mixed-effects modeling. Data for each dependency type and LLM were analyzed separately, and statistical results were Bonferroni corrected for multiple comparisons for each LLM across dependency types.

In the main analysis, we modeled attention strength between words in the critical dependency (either from the head to the dependent, or from the dependent to the head, depending on the identified attention head; see Table 1). The model included two fixed effects: plausibility (plausible vs. implausible) and, as a covariate, log-frequency for the word in the critical dependency that differed between two sentences in a pair (from the $SUBTLEX_{US}$ corpus; Brysbaert & New, 2009). The model also included a random intercept by sentence pair, unless it did not converge (and was reduced to a fixed-effects model). We predicted that, compared to plausible sentences, implausible sentences would show *lower* attention to the syntactically correct target in the critical dependency.

In a secondary analysis, we analyzed attention strength

Table 2: Example sentence pairs in the main experiment, for BERT

Type	Plausible / likely	Implausible / unlikely
dobj	The guide **showed** the visitor a **sculpture.**	The guide **showed** the sculpture a **visitor.**
nsubj	The **surfer** at the beach **rescued** a swimmer.	The **beach** at the surfer **rescued** a swimmer.
pobj	It was the **ladder** in the tree that the cat climbed **up**.	It was the **simplicity** in the tree that the cat climbed **up**.
iobj	The guide **showed** the **visitor** a sculpture.	The guide **showed** the **sculpture** a visitor.

Words in bold constitute the critical dependency. Attention is directed from the underlined word to the other bolded word. Words in red are "lures".

involving the "lure" word: either from the attention-directing word to the lure (for *pobj* in BERT, and *nsubj* in GPT-2), or from an attention-directing lure to the attention-receiving word in the syntactically correct dependency (for *dobj*, *iobj*, and *nsubj* in BERT; and *dobj* and *pobj* in GPT-2). We predicted that, compared to plausible sentences, implausible sentences would show higher attention to/from the syntactically incorrect lure. We corrected for multiple comparisons in the same manner as for the main analysis.

Results

Consistent with our prediction, attention strength between words in a critical dependency was modulated by whether the sentence was plausible or not, demonstrating an influence of semantics on the computations of syntax-specialized attention heads. In BERT, attention strength between critical words was significantly lower in the implausible condition than in the plausible condition for three out of four dependencies (*dobj*: $t_{(49.60)}$=3.64, p=0.0007; *nsubj*: $t_{(48.92)}$=5.01, $p<10^{-4}$; *pobj*: $t_{(48.42)}$=8.60, $p<10^{-4}$), with the exception of *iobj* ($t_{(49.44)}$=-1.37, p=0.18) (Fig. 1). Attention directed to/from lure words showed the opposite pattern and was significantly higher in the implausible condition than in the plausible condition for *nsubj* ($t_{(97)}$= -9.09, $p<10^{-4}$) and *pobj* ($t_{(49.00)}$=-5.60, $p<10^{-4}$), but not for *dobj* ($t_{(48.31)}$= -0.74, p=0.47) or *iobj* ($t_{(49.22)}$=-1.23, p=0.22). Mixed effects models were used in BERT for all analyses except for the lure analysis of *nsubj* (for which a fixed-effects model was used).

Similarly, in the main analysis for GPT-2, attention strength between critical words was significantly lower in the implausible condition than in the plausible condition for all three dependencies tested (*dobj*: $t_{(49.86)}$=6.35, $p<10^{-4}$; *nsubj*: $t_{(48.75)}$= 3.73, p=0.0005; *pobj*: $t_{(49.10)}$=2.88, p=0.0059). Attention directed to/from lure words was significantly higher in the implausible condition for *dobj* ($t_{(50.06)}$=-4.05, p=0.00018) and *nsubj* ($t_{(48.91)}$=-11.58, $p<10^{-4}$), but not for *pobj* ($t_{(49.00)}$=-1.85, p=0.071). All models were mixed-effects.

Discussion

This study found that syntax-specialized attention heads in BERT and GPT-2 are influenced by semantic plausibility information: even in heads that appear to have a strong preference for a particular syntactic dependency, the attention strength between words in that dependency becomes weaker in implausible (vs. plausible) sentences. This pattern was observed in all cases but one (the *iobj* dependency in BERT). Moreover, in some cases, the implausible sentences show increased attention between words that, syntactically, do not constitute the preferred dependency, but are nonetheless semantically plausible candidates for that dependency. These findings demonstrate that the syntactic computations in these attention heads are not encapsulated: they can be modulated by plausibility information. Such non-encapsulation is functionally similar to the human mind, in which syntactic parsing is similarly penetrable to semantics from the earliest moments of processing. Therefore, our findings lend further support to the use of LLMs as plausible cognitive models.

This study presents a strong test for the encapsulation of syntax in LLMs. First, we tested attention heads that appear to be strongly specialized for specific syntactic dependencies (although they do not identify their preferred dependency in 100% of the cases; Table 1). Given that attention heads link words in a sentence to one another (via attention weights), they are an a-priori likely site for "purely" syntactic computations. Of course, we do not exclude the possibility that syntactic encapsulation might exist elsewhere in LLMs (e.g., in other attention heads, for other dependency types, or in the hidden units). Still, *if* there is syntactic encapsulation anywhere within LLMs, then the attention heads we identified are arguably the best candidates for carrying out such computations. We provide compelling evidence that their computations are instead penetrable.

Second, many of our stimuli contained no temporary syntactic ambiguity: BERT, being bidirectional, has access to all words in a sentence, and all 4 dependencies we tested used

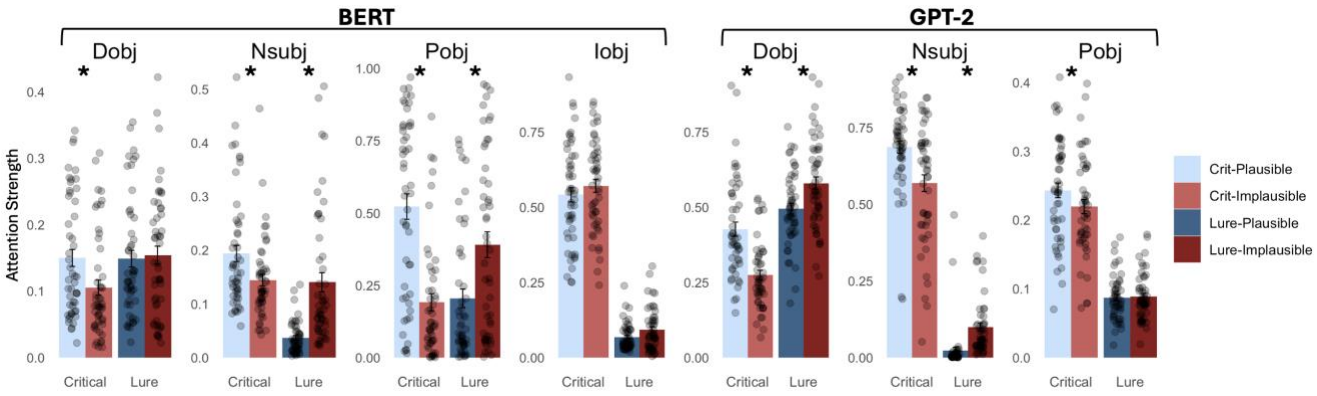

Figure 1: attention strengths between words in a critical dependency, extracted from "syntax-specialized" attention heads. Each dot shows data from one sentence, with bars showing averages and error bars—standard errors. Data are shown for BERT (top) and GPT-2 (bottom), comparing plausible (blue) and implausible (red) sentences. Bright bars show attention for the critical, syntactically correct dependency ("crit"), and dark bars show attention to/from semantic lures.

materials whose overall structure is unambiguous; for GPT-2, which is only left-looking, the dependent in *dobj* sentences was the last word (so the entire structure is available at that point), and in *pobj* sentences the partial context prior to the critical words (e.g., "the cat climbed *up* the *ladder*") is unambiguous.[3] Because there is no need for syntactic disambiguation, there is in principle no need to rely on semantic information to parse the sentence's structure. This contrasts with studies of syntactic encapsulation in humans, where the materials are temporarily ambiguous and thus reliance on non-syntactic information confers an advantage (Trueswell et al., 1994; Tanenhaus et al., 1995; but see Sikos, Duffield, & Kim, 2016). And yet, despite the sufficiency of the syntactic cues for determining sentence structure, the attention heads were still influenced by semantic information. This suggests that penetrability is not optionally triggered when needed, but is rather inherent to the functionality of these attention heads. In some cases, semantic information might even override syntactic information, leading attention heads to incorrectly represent dependencies. For instance, in implausible sentences, *dobj*-specialized heads in both LLMs assign more attention to the lure than to the syntactically correct direct object (compare the red bars in Fig. 1).

In human language processing, the question of syntactic encapsulation concerned timing: not *whether* semantic information can influence syntactic analysis (it does, because different information sources must interact at some point), but rather *how soon*. Namely, non-encapsulated syntactic mechanisms are those that are penetrable immediately, from the very first moments of processing. However, unlike humans, transformers process all words in parallel rather than incrementally. Does this mean that our results do not really address the critical aspect of encapsulation, i.e., timing? We do not think so: the analysis of the unidirectional GPT-2 found plausibility effects at the word that constitutes a critical dependency, a site where any words beyond that dependency are unavailable. This is akin to reading time studies with human participants that report immediate effects at a critical word—rather than later, at a spillover region—as evidence for the early influence of non-syntactic information (e.g., Trueswell et al., 1994).

The results for our analyses are also consistent with noisy channel processing accounts of human language comprehension. Such accounts acknowledge that signals sent over a communication channel are susceptible to corruption due to noise, and therefore the signal perceived by the comprehender is not necessarily identical to the signal intended by the producer. Hence, comprehension involves inferences about intended messages (Gibson, Bergen, & Piantadosi, 2013; Levy et al., 2009). Given an implausible sentence ("The guide showed the sculpture a visitor"), a rational comprehender might "mentally correct" the sentence into a more semantically plausible alternative (e.g., "the guide showed the visitor a sculpture"; Gibson et al., 2013; Poppels & Levy, 2016; Ryskin et al., 2018). The results from our main analysis demonstrate that, in implausible sentences, the attention patterns of "syntax-specialized" attention heads might reflect such a process of searching for a plausible meaning: these heads assign less weight to their preferred dependencies than they do in plausible sentences; and, in some cases, they assign more weight to/from lure words that are semantically plausible (but syntactically incorrect) targets for that dependency.

Whereas we interpret our results in terms of plausibility effects, in our sentence materials plausibility is confounded with predictability: during training, LLMs are likely to see words occupying plausible dependencies (like the critical dependencies in our plausible condition), but might only infrequently be exposed to words occupying implausible dependencies (like in our implausible condition). If LLMs have less experience with, e.g., "showed" and "visitor" being in a direct object dependency compared to "showed" and "sculpture", this difference in exposure could explain our results. This explanation can be tested via an additional condition in which sentences are plausible but surprising due to low frequency but plausible critical words (e.g., "the guide showed the visitor a cyanotype"). If our current results reflect plausibility rather than predictability alone, then attention weights for the critical dependencies should be predicted by plausibility over and above measures of word predictability.

Why might syntax-specialized attention heads in transformer-based LLMs be penetrable to semantic information? One reason is that encapsulated systems are mostly advantageous in domains that are unambiguous, where a system's expertise can operate without external interference. In contrast, informational domains that are pervasively ambiguous—such as syntax—benefit from the penetrability of external information sources to aid analysis on a first pass and resolve indeterminacies without the need for re-processing (Trueswell et al., 1994). Given that (1) LLMs are trained to predict words in language, (2) such prediction can be optimized when using syntactic information, and (3) syntactic information is often ambiguous, LLMs might automatically learn to represent and process syntax in a non-encapsulated manner. In this sense, a fundamental computational property of their attention heads shows functional similarities to the human mind.

References

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In

[3] For *nsubj*, a fragment like "the surfer at the beach rescued…" is temporarily ambiguous between an active transitive construction ("…the swimmer.") and the beginning of a reduced relative clause ("…by the lifeguard was grateful."). For this dependency in GPT-2, our results reflect more closely the kind of materials used with human participants in prior work.

ICLR.

Arehalli, S., Dillon, B., & Linzen, T. (2022). Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, 301–313.

Arps, D., Kallmeyer, L., Samih, Y., & Sajjad, H. (2023). Multilingual Nonce Dependency Treebanks: Understanding how LLMs represent and process syntactic structure. *arXiv preprint arXiv:2311.07497*.

Bencini, G. M., & Goldberg, A. E. (2000). The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language*, *43*(4), 640-651.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, *41*(4), 977-990.

Casenhiser, D., & Goldberg, A. E. (2005). Fast mapping between a phrasal form and meaning. *Developmental science*, *8*(6), 500-508.

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

Chomsky, N. (2014). *Aspects of the Theory of Syntax* (Vol. 11). MIT press.

Chomsky, N. (1995). Language and nature. *Mind*, *104*(413), 1-61.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? an analysis of BERT's attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276-286.

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.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1, 4171–4186 (2019).

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *24*(4), 270-284.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 32–42.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051-8056.

Goldberg, A.E. (1995). Constructions: A construction grammar approach to argument structure. *University of Chicago Press*.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in cognitive sciences*, *7*(5), 219-224.

Goldberg, A. E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Gulordava, K., Bojanowski, P., Grave, É., Linzen, T., & Baroni, M. (2018). Colorless Green Recurrent Networks Dream Hierarchically. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1195–1205.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1725–1744.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Johnson, M. A., & Goldberg, A. E. (2013). Evidence for automatic accessing of constructional meaning: Jabberwocky sentences prime associated verbs. *Language and Cognitive Processes*, *28*(10), 1439-1452.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the national academy of sciences*, *106*(50), 21086-21090.

Li, B., Zhu, Z., Thomas, G., Rudzicz, F., & Xu, Y. (2022). Neural reality of argument structure constructions. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423.

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

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective. *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.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank.

De Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006, May). Generating typed dependency parses from phrase structure parses. In *Lrec* (Vol. 6, pp. 449-454).

Maudslay, R. H., & Cotterell, R. (2021). Do syntactic probes probe syntax? experiments with jabberwocky probing. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 124–131.

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.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283-312.

OpenAI (2022). ChatGPT: Optimizing language models for dialogue. *OpenAI*.

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. T. (2018). Dissecting contextual word embeddings: Architecture and representation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

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

Poppels, T., & Levy, R. (2016). Structure-sensitive Noise Inference: Comprehenders Expect Exchange Errors. *In CogSci*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

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

Raganato, A., & Tiedemann, J. (2018). An analysis of encoder representations in transformer-based machine translation. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. The Association for Computational Linguistics.

Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition*, *181*, 141-150.

van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive science*, *45*(6), e12988.

Sikos, L., Duffield, C. J., & Kim, A. E. (2016). Grammatical predictions reveal influences of semantic attraction in online sentence comprehension: Evidence from speeded forced-choice sentence continuations. *Language, cognition and neuroscience*, *31*(8), 1055-1073.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632-1634.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of memory and language*, *33*(3), 285-318.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 37–42.

Vig, J., & Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 63–76.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5797–5808.

Wu, J. M., Belinkov, Y., Sajjad, H., Durrani, N., Dalvi, F., & Glass, J. (2020). Similarity analysis of contextual word representation models. *arXiv preprint arXiv:2005.01172*.

Wilcox, E. G., Vani, P., & Levy, R. P. (2021). A Targeted Assessment of Incremental Processing in Neural Language Models 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.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.