Check for updates

TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 17 (2025) 948–961
© 2024 The Authors. *Topics in Cognitive Science* published by Wiley Periodicals LLC on behalf of Cognitive Science Society.
ISSN: 1756-8765 online
DOI: 10.1111/tops.12733

This article is part of the topic "Parallelism in the Architecture of Language," Giosuè Baggio, Neil Cohn, and Eva Wittenberg (Topic Editors).

Neural Generative Models and the Parallel Architecture of Language: A Critical Review and Outlook

Giulia Rambelli,[a] Emmanuele Chersoni,[b] Davide Testa,[c] Philippe Blache,[d] Alessandro Lenci[e]

[a]*Department of Modern Languages, Literatures, and Cultures, University of Bologna*
[b]*Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University*
[c]*Fondazione Bruno Kessler, Trento*
[d]*Laboratoire Parole et Langage, CNRS*
[e]*Department of Philology, Literature, and Linguistics, University of Pisa*

Received 31 August 2023; received in revised form 15 March 2024; accepted 21 March 2024

Abstract

According to the parallel architecture, syntactic and semantic information processing are two separate streams that interact selectively during language comprehension. While considerable effort is put into psycho- and neurolinguistics to understand the interchange of processing mechanisms in human comprehension, the nature of this interaction in recent neural Large Language Models remains elusive. In this article, we revisit influential linguistic and behavioral experiments and evaluate the ability of a large language model, GPT-3, to perform these tasks. The model can solve semantic tasks autonomously from syntactic realization in a manner that resembles human behavior. However, the

Correspondence should be sent to Emmanuele Chersoni, The Hong Kong Polytechnic University, 11 Yuk Choi Kowloon, Hong Kong. E-mail: emmanuele.chersoni@polyu.edu.hk

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

outcomes present a complex and variegated picture, leaving open the question of how Language Models could learn structured conceptual representations.

Keywords: Neural large language models; Statistical learning; Parallel architecture; Syntax-semantics interface; GPT-3 prompting; Enriched composition; Semantic composition

1. Introduction

The last few years have seen an incredible advance in Natural Language Processing and artificial intelligence thanks to the rise of neural *Large Language Models* (LLMs), also referred to as *foundation models* (Bommasani et al., 2021). These models consist of deep neural networks (LeCun, Bengio, & Hinton, 2015) trained on unannotated textual data with the general objective of predicting upcoming linguistic material. Their complex architecture based on Transformers with attention mechanisms (Vaswani et al., 2017) and the possibility of being trained on massive datasets have allowed these models to achieve unprecedented performances in generating human-like texts. Moreover, LLMs like GPT-3 reveal "emergent abilities" to carry out various linguistic tasks (e.g., translating, question-answering, etc.) without any task-specific training (Brown et al., 2020; Wei et al., 2022). Therefore, the "core knowledge" of foundation models is not only a "knowing that" of structures of languages but also a "knowing how" to use language itself (Lenci, 2023, p. 5). Their emergent abilities have also opened the way to a new unsupervised method for solving tasks called *in-context learning* (Brown et al., 2020), which consists of feeding the model with a natural language instruction, namely, a *prompt*, that contains a description of the task.

The outstanding success of LLMs has also been hailed in cognitive science as empirical proof against previous theoretical models of the architecture of the language faculty. Piantadosi (2023) explicitly regards LLMs as a refutation of the Chomskyan view of language. In fact, LLMs lack some of the major properties that Chomsky assumes to be essential for natural language learning: (i) they do not use symbolic representations but continuous vectors (also known as *contextual embeddings*); (ii) they do not include principles based on hierarchical structures as "innate" learning biases; (iii) they do not assume any separation between syntax and semantics. These points can be already found in the connectionist debate in the 1990s (Elman et al., 1996); however, they are now strengthened by the existence of computational models that adhere to such architectural principles and that for the first time ever show—at least *prima facie*—an almost human-analogue generative capacity. Goldstein et al. (2022) claim that autoregressive language models like GPT-3 share with the brain important computational principles, as the brain is also constantly involved in next-word prediction as it processes natural language. Since they acquire linguistic knowledge through statistical learning from large text samples, Goldstein et al. (2022) argue for a strong connection between LLMs and the usage-based, constructionist paradigm (Bybee, 2010; Goldberg, 2019), which states that the complexity of language emerges not as a result of a language-specific device, but through the interaction of cognition and use.

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In this paper, we investigate whether and to what extent the linguistic abilities of LLMs provide support to parallel architectures of human language processing (Baggio, 2018; Baggio, 2021; Jackendoff, 2007), according to which syntactic and semantic information processing can interact selectively during language comprehension. We focus on two main assumptions of parallel models: (i) semantics drives meaning composition in synergy with, but also autonomously from, syntactic structures (*autonomous semantics*; see Baggio, 2018, Jackendoff, 1997); (ii) there is no principled distinction between semantics and pragmatics, and mechanisms of context-sensitive "pragmatic enrichment" are inherent parts of the standard process to construct meaning, as revealed by phenomena like coercion and metonymy. These assumptions make meaning construction in parallel models depart radically from "classical" Fregean compositionality (Culicover & Jackendoff, 2006).

The relationship between parallel models and LLMs is not straightforward, as they present some essential commonalities but also crucial differences. While parallel models assume a *dual-stream architecture* in which syntax and semantics act as autonomous components that produce independent yet linked representations, LLMs consist of a *single-stream architecture* that simultaneously learns and processes syntactic, semantic, and pragmatic information encoded together in the internal embeddings. This configuration departs from the principle of autonomous semantics, which argues for the combinatorial independence of both syntax and semantics, rather than the absence of any distinction between the two. On the other hand, like in the parallel architecture, any principled separation between semantic and pragmatic information is lacking in LLMs by design.

The complexity of LLMs (especially of the largest ones) and their "black box" nature makes any purely architectural analysis speculative at most, as a direct inspection of information flow is often impossible or extremely hard to determine. A more effective method is to infer how models represent different kinds of linguistic information from their abilities in addressing tasks *specifically designed* to single out that information. An increasing amount of research has been deployed to dissect the linguistic abilities of LLMs (for a synthesis of recent results, see Chang & Bergen, 2024), going beyond the merely "observational" analysis of the human-like nature of the texts they generate. Several works have shown that LLMs possess a nontrivial extent of syntactic knowledge (Y. Goldberg, 2019; Hewitt & Manning, 2019; Lin, Yi, & Frank, 2019; Linzen & Baroni, 2021; Liu et al., 2019; Tenney et al., 2019), though it is still an open debate about how much use they make of it in language processing (Glavaš & Vulić, 2021; Prange et al., 2022; Warstadt et al., 2020). Besides, these models also capture a broad range of lexical and semantic phenomena, such as argument structure constructions (Li et al., 2022), idioms and multiword expressions (Dankers, Lucas, & Titov, 2022; Nedumpozhimana & Kelleher, 2021; Rambelli, Chersoni, Senaldi, Blache, & Lenci, 2023), compounding (Buijtelaar & Pezzelle, 2023; Miletić & im Walde, 2023; Ormerod et al., 2024), and lexical semantics (Bommasani, Davis, & Cardie, 2020; Vulić, Ponti, Litschko, Glavaš, & Korhonen, 2020). Pedinotti et al. (2021) and Kauf et al. (2023) indicate that foundation models have knowledge about events and their plausible participants, while Hu, Floyd, Jouravlev, Fedorenko, and Gibson (2023) systematically investigate the pragmatic abilities of foundation models and find that they solve some of them with an accuracy close to human performance.

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1
Overview of the experiments introduced in our study

Experiment	Targeted ability	Task	Dataset
1 (Section 2**; SM B.2)**	*Logical metonymy interpretation*	Retrieve the omitted verb.	Rambelli, Chersoni, Lenci, Blache, & Huang (2020)
2 (Section 2**; SM B.3)**	*Elliptical sentence resolution*	Retrieve the omitted semantic material from the antecedent clause.	Testa et al. (2023)
3 (Section 3**; SM B.4)**	*Semantic attraction resolution* (syntactic and semantic violation)	Identify whether the model prefers a meaning- or a syntax-driven prediction of the Agent in sentences with semantic attraction.	Same subset of Kim & Osterhout (2005) and Michalon & Baggio (2019)
4 (Section 3**; SM B.5)**	*Scrambled sentences interpretation* (syntactic violation)	Interpret a sentence whose syntactic structure has been perturbed.	Mollica et al. (2020)

Nevertheless, other studies point out the challenges faced by LLMs in dealing with pragmatic phenomena (Liu et al., 2023; Ruis et al., 2022), often necessitating a grounding in real conversational or communicative contexts (Andreas, 2022; Pezzelle, 2023; Schlangen, 2022), and their general being prone to as errors deriving from overgeneralizations or undergeneralizations of learned patterns in the training data (Chang & Bergen, 2024).

In this work, we investigate the connection between LLMs and the parallel architecture with four case studies that exemplify syntax-semantics mismatches: *logical metonymy*, *ellipsis*, *semantic attraction*, and *semantic composition in scrambled sentences* (Table 1 and Table 2 provide an overview of the implemented tasks). The common feature of these phenomena is the lack of a perfect alignment between syntax and semantics, which motivates the fact that they have been taken to support parallel models of language. We selected GPT-3 *davinci*-002 (Brown et al., 2020), one of the largest LLMs (this model was selected among other members of the GPT family because of its lack of fine-tuning with reinforcement learning from human feedback), to replicate linguistic and psycholinguistic experiments designed to investigate the autonomy between syntactic and semantic processing streams, as well as its interplay with pragmatic information, progressively increasing the task difficulty. We prompted the model with linguistic questions that resemble those submitted to human participants and evaluated its answers. Since the model was trained only on sentence completion, we formulated the input prompt in two ways: (i) a *zero-shot setting* and (ii) a *few-shot setting*, where some examples (two Q-A pairs) of the experimental task are provided (see Fig. 1). The final aim is to gain insights into the interplay between syntax and semantics in LLMs. Due to space limitations, we only discuss the experiments' main findings; details are reported in the Supplementary Material (SM).

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2
Example prompts for all Experiments

Experiment	Subtask	Prompt (+ illustrative answer)
1	–	Q: The writer finished the novel. What did the writer finish doing? **A: The writer finished writing the novel**.
2	–	Q: Kate played the piano, and Carl the guitar. What did Carl do? **A: Carl played the guitar**.
3	*Agent retrieval*	Q: The murder had been witnessing by the three bystanders. What is the agent of the sentence? **A: The murder**. (or **A: The three bystanders**.)
	Grammatical subject retrieval	Q: The murder had been witnessing by the three bystanders. What is the grammatical subject of the sentence? **A: The murder**.
4	*Agent identification*	Q: The detective looked from one man to another and flexed his moustaches. What is the agent of the sentence? **A: The detective**.
	Event resolution	Q: The detective looked from one man to another and flexed his moustaches. What did the detective do? **A: The detective looked from one man to another and flexed his moustaches**.

Note. Examples of good answers are marked in **boldface**.

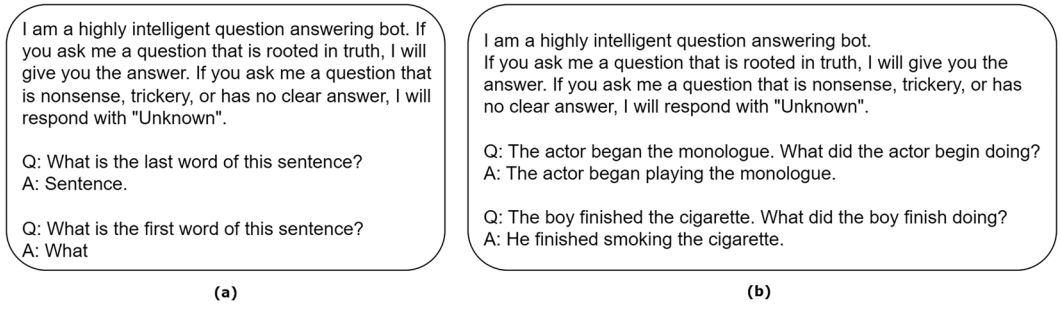

Fig. 1. Examples of zero-shot (left) and few-shot (right) prompts. The text of the preamble is inspired by OpenAI's QA example (to explicitly tell the model how it should behave or respond), followed by two examples of a question-answer alternation, which are unrelated (a) or related (b) to the target task. This solution helps the model correctly answer the question starting with "A:", as GPT-3 *davinci-002* was not fine-tuned for this task.

2. How do neural LLMs recover covert lexical material?

The first two phenomena under investigation—*logical metonymy* (cf. SM Section B.2) and *ellipsis* (cf. SM Section B.3)—are constructions in which some semantic elements are not overtly realized. *Logical metonymy* (Pustejovsky, 1995) is defined as a type clash between an event-selecting metonymic verb (*begin*) and an entity-denoting direct object (*symphony*) that triggers the recovery of a covert event (*playing*):

(1) The pianist begins the symphony.

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3
Experiments' results

Experiment	Subtask	Metric	Results	
			zeroShot	fewShot
1	*Logical metonymy interpretation*	P@1	0.37	0.69
		P@3	0.43	0.85
2	*Elliptical sentence resolution*	Event retrieval	0.57	0.80
		Verb match	0.71	0.88
3	*Agent retrieval (AV sentences)*	Agent = First NP	0.21	0.18
		Agent = Second NP	0.70	0.82
	Grammatical subject retrieval (AV sentences)	Accuracy	0.79	1
4	*Agent identification*	Accuracy	0.74	0.74
	Event reconstruction	Accuracy	0.42	0.70

Note. Exp. 1: Accuracy as the number of times GPT-3's outputs correspond to the 1st (*P@1*) or is among the 3 (*P@3*) most produced verbs. Exp. 2: Accuracy as *Event Retrieval* (how many times the model correctly retrieved the omitted verbal phrase) or *Match Verb* (how many times the model correctly retrieved the verb). Exp 3: How many times the Agent is the first or second noun phrase, and the accuracy of correctly identifying the grammatical subject. Exp. 4: Accuracy of correctly answering with the same agent (*Agent identification*) or event (*Event reconstruction*) as in the original nonswapped sentence.

In (1), what the pianist begins doing is not lexically explicit, but most speakers would interpret it as *the pianist begins playing the symphony*. To master logical metonymy, LLMs must have information about the typical events associated with nouns and integrate this prior knowledge to generate a context-sensitive representation of sentence meaning (Rambelli et al., 2020).

Overall, we observe that LLMs are mostly able to recover the implicit content. The high accuracies (see Table 3) in the logical metonymy interpretation task reveal that GPT-3 is quite good at recovering the missing pieces of meaning when an example of the task is given (few-shot setting). Nevertheless, the answers it produces are strongly driven by previous expectations, that is, the most likely item previously seen in that context. Error analysis revealed that frequency plays a crucial role in the model's interpretation: If the direct object strongly co-occurs within a particular event, the corresponding verb is recalled regardless of the context. This tendency is relatively controlled in the logical metonymy task, where the model sometimes fails when expectations are based only on the direct object without considering the subject (e.g., *The publisher began writing the novel* instead of *publishing the novel;* more examples in SM B.2). The results show that the GPT-3 has a good amount of semantic knowledge about events and their typical participants, allowing it to perform well in recovering the covert event, at least in the few-shot setting. However, sometimes, it does not consider the overall context (the subject, in this case). In general, high-frequency verb-object associations override contextual information, which plays an essential role in human interpretation and pertains to pragmatic competence.

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Ellipsis is also a phenomenon of interpretation based on covert material, since it consists of the omission of a word or phrase that is expected to occupy a place in the syntactic structure of a sentence (McShane, 2005):

(2) The researcher completed the project, *but the student didn't*.

According to Culicover and Jackendoff (2005), the interpretation of an elliptical construction is derived from the semantic representation (Conceptual Structure in their model) of its antecedent or its context.

For elliptical sentence resolution, LLMs must fill the gap by relying on semantic material "copied" by the previous clause. Even if GPT-3 is trained to look at the context to produce a probability distribution, the model seems to make predictions in a top-down way, preferring highly likely events rather than truly interpreting the elliptic gap. Again, the model only reaches a reasonably high accuracy in the few-shot setting (cf. Table 3). Admittedly, this tendency could be an artifact of the parameter choice. Indeed, we set a low temperature to make the model more deterministic and force it to produce the most likely output. However, this constraint should not impact the ability of the model to look at the previous information. Moreover, if that were the case, it should consistently fail, while accuracy varies depending on the thematic fit of the omitted verb with its arguments. This evidence seems to confirm Warstadt et al. (2020)'s observation: while LLMs show high performance on tasks related to syntax (and, more generally, the sentence structure and some metalinguistic components), they often prefer to rely on highly localized mutual expectations. Even in the less challenging few-shot setting, there are still cases in which GPT-3 fails by copying the sequence in the antecedent clause without adapting it to the content of the elliptic gap. In these examples, it retrieves the direct object from the antecedent relying on frequent lexical co-occurrences (e.g., *Q: The student will hear the lecture, and the nanny will the baby. What will the nanny do? A: The nanny will hear the lecture*).

An intriguing aspect concerns the elliptic sentences containing a selectional preference violation:

(3) * The photographer used the camera, and the piano did too.

A key aspect of elliptic constructions is that the reconstructed material must preserve the semantic constraints of its overt "copy": We can judge (3) to be anomalous precisely because we are able to interpret the missing verb phrase as being identical to the one in the antecedent. However, the violation of selectional preferences always determines a significant performance drop in GPT-3. This outcome confirms that the model has not really managed to solve the elliptical construction. Like in other cases, its behavior seems to be guided more by lexical cues (e.g., highly frequent events) rather than by genuine mastery of linguistic structure.

Overall, logical metonymy and ellipsis interpretation tasks confirm the influence of the prototypicality of the event participants in interpreting such linguistic structures and a lack of pragmatic ability to adapt common event knowledge given a broader context.

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3. How do neural LLMs reconstruct semantic structures in syntactically anomalous sentences?

As the parallel architecture proposes, syntactic and semantic information processing can interact selectively during language comprehension. This hypothesis means that, in specific conditions, the processing system can rely on the semantic relationships between words to produce an interpretation, even if that contradicts syntactic cues. We investigated this assumption with two types of syntactic anomaly: sentences with *semantic attraction* (cf. SM Section B.4) and sentences with *scrambled word order* (cf. SM Section B.5).

Semantic attraction is a sentence-processing phenomenon in which a given argument violates the selectional requirements of a verb, but comprehenders do not perceive this violation due to its attraction to another noun in the same sentence, which is syntactically unrelated but semantically sound (Cong, Chersoni, Hsu, & Lenci, 2023). A classic example from Kim and Osterhout (2005) is the following, in which the verb *devouring* is perceived as either syntactically or semantically anomalous:

(4) The hearty meal was devouring the kids.

When asked to identify the Agent of the sentence, we observe that GPT-3 is driven mainly by the underlying semantic expectations rather than the syntactic constraints in the two settings (the interpretation of the Agent as the *by*-phrase is 70% in the zero-shot and 82% in the few-shot setting). Specifically, even if the grammatical subject is the Agent of the sentence from a structural point of view, the model seems to suppress the structural information and it chooses the prototypical Agent for the target verb, even if it is in the wrong syntactic position. This tendency is indirect evidence that the model relies on its previous knowledge of events (Kauf et al., 2023) and, in cases of ambiguities, the answer is grounded on what it expects the most. Nevertheless, this preference for prior collocation knowledge could sometimes produce nonsensical or out-of-context utterances. However, the behavioral data regarding semantic attraction consist of measurements of online processing (recording brain activity), while prompting strategies are more analogous to offline judgments. A more robust comparison should investigate the correlation between human measurements and the model surprisal scores (Hale, 2001; Levy, 2008), considered the best predictor of N400 (Michaelov & Bergen, 2022).

A parallel question of interest is whether semantic composition can take place even in sentences with *scrambled word order*, whose syntactic structure is not licensed by the language's grammar. In this experiment, we tested GPT-3 on the stimuli used by Mollica et al. (2020) in their neurolinguistic investigation about how composition is robust to syntactic violations. They gradually degraded the syntactic structure of a sentence by increasing the number of local word swaps while preserving local dependency relationships (so that combinable words remained close to each other). We adjusted the original experimental paradigm to display what GPT-3 can reconstruct of the meaning of a scrambled sentence. We presented two prompts to the model, corresponding to an *Agent identification task* (like in the semantic attraction experiment) and an *Event reconstruction task*. To correctly associate the interpretation of a sentence whose syntactic structure has been perturbed, LLMs should be able to (i) detect semantic and

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

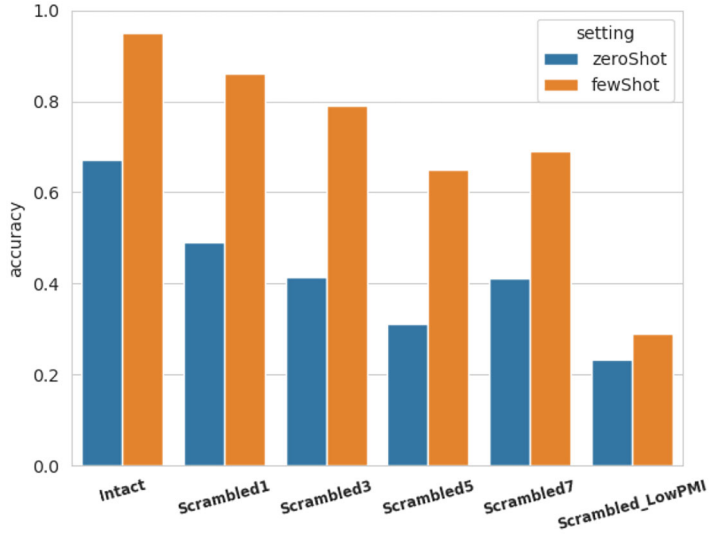

Fig. 2. Accuracy of the Event reconstruction subtask for all conditions, including the original (Intact) sentence as a baseline.

syntactic local dependencies among nearby words and (ii) reconstruct the global sentence meaning from such dependencies.

In language processing, the brain tolerates a certain degree of structural anomaly, as long as local combinatoriality is preserved at the lexical level (Mollica et al., 2020). The analysis reveals that GPT-3 can reconstruct some essential aspects of the evoked event; nonetheless, it has problems in reconstructing the whole scenario after only two swaps in the zero-shot setting (performing below 0.5 in all conditions). However, the model is more comparable to human behavioral patterns when provided with some examples. GPT-3 exhibits a significant but smooth drop in accuracy with the increase of word-order swaps, which is above the chance level except for the *ScrLowPMI* condition (Fig. 2). This trend is not entirely comparable to neural responses (cf. fig. 2 in Mollica et al., 2020), but better aligns with the reconstruction abilities of humans (cf. fig. 4b in Mollica et al., 2020), who also faced increasing difficulty in reconstructing the original sentences as the number of swaps increases. Overall, these observations suggest that the model is highly dependent on the surface structure in its default version: The more the syntactic order is violated, the lower the model performance, which starts to output random texts unrelated to the input. However, like humans, the syntactic irregularities are less problematic when limited and localized for a model prompted with few examples.

It is worth bearing in mind that, as already stated in Section 1, LLMs integrate syntax and semantics in a continuous and indistinct way (Piantadosi, 2023). The internal representations of words in the hidden layers retain both aspects of semantic and syntactic word-order constraints, and each prediction is the result of the interaction of several statistical distributions

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

in complex ways. Consequently, distinguishing the effects of one module over the other is nontrivial.

However, this struggle is the same experienced in psycho- and neurolinguistics, as it is hard to fully dissociate the two streams of processing in the human brain. Developing unique paradigms that could be applied to both humans and machines could provide novel evidence about how the predictive faculty is prominent in cognitive and computational processing.

To summarize, the behavior of GPT-3 reveals that *neural LLMs capture some metalinguistic categories*, such as the abstract notions of Agent and grammatical subject, and *they are relatively robust to limited and localized word-order violations*. However, the ability to successfully interpret these anomalous sentences does not entirely match human performance.

4. Can neural LLMs be regarded as computational models of the parallel architecture?

This question does not have a straightforward answer and the experiments we have presented compose a complex and variegated picture. The parallel architecture rests on two fundamental principles: (i) semantics is a combinatorial stream independent from syntax (*autonomous semantics*), and (ii) pragmatic, contextual information is an inherent component of phrasal meaning construction (*enriched composition*). The knowledge that LLMs like GPT-3 use to process natural language consists of an intertwined bundle of syntactic, semantic, and pragmatic information derived from textual data through distributional learning and encoded into their internal embeddings (Lenci & Sahlgren, 2023). Therefore, while meaning construction in LLMs can be regarded as a case of enriched composition, their architecture departs from the principle of autonomous semantics. Besides, the latter is not *stricto sensu* autonomous from syntax, as no distinction exists between the two information flows. In this paper, we analyzed the behavior of GPT-3 in several test cases of misalignment between syntax and semantics, and the experimental results show that, despite the lack of such distinction at the architectural level, the model *can solve semantic tasks autonomously from syntactic realization* in a manner that resembles human behavior, especially when provided with few examples of the task (few-shot setting). So, *even if current LLMs cannot be explicitly regarded as models of the parallel architecture, their behavior shows some important consonance with key predictions of the latter*, particularly with respect to the relationship between syntax and semantics processing. However, there are also crucial limits, as shown for instance in the reconstruction of elliptic gaps, and differences, as in the case of sentences with scrambled word order.

One possible reason of such discrepancies might be the lack of a full-fledged autonomous semantics in LLMs. Their representations encode distributional information derived from texts and related to several linguistic dimensions (morphology, syntax, semantics, world knowledge, etc.). Indeed, the range of phenomena encoded in language that can be recovered from distributional statistics is far greater than we could have ever imagined in the past. These linguistic dimensions are, however, bundled together in vectors that look like Hegels's night, in which "all cows are black" (Hegel, 1979): A lot of linguistic information is there,

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

but LLMs do not manipulate structured world representations (Lenci, 2023; McCoy, Yao, Friedman, Hardy, & Griffiths, 2023). LLMs mainly identify highly sophisticated associative links between linguistic expressions but do not yet have anything close to the conceptual structures that truly support semantic representations. Mahowald et al. (2023) claim that LLMs have sophisticated *formal linguistic competence*, that is, knowledge of linguistic rules and patterns, while they fall short of *functional competence*, as the ability to understand and use language in the world, which is in turn strongly related to the semantic stream in the parallel architecture. Functional competence requires complex "theories" of the world and mind and mechanisms that allow their use to drive linguistic behavior. This type of information is still qualitatively different from the one that foundation models seem to be able to acquire from linguistic data and represent in their continuous embeddings, as shown by their differences from human semantic processing (Chang & Bergen, 2024; Kauf et al., 2024; Kauf et al., 2023; Pedinotti et al., 2021). For semantics to become a truly autonomous stream governing meaning construction, conceptual structures supporting functional competence might be required. The extent to which such structured information can be learned and represented in LLMs is still an open issue.

Author contributions

This article is the result of the collaboration between the authors. In particular, AL, PB, and EC formulated the research goals and conceived the study. GR implemented Experiments 1, 3, and 4 (Sections 2 and 3). DT implemented Experiment 2 (Section 2) and was responsible for the verification of the reproducibility of results (Sections 2 and 3). The article was written by GR and it was edited by AL and EC before submission. GR performed the final edits on the complete draft and all revisions. For the specific concerns of the Italian academic attribution system, GR is responsible for Sections 1, 2, and 3; AL is responsible for Section 4.

Acknowledgments

We thank the three anonymous reviewers for their useful comments and suggestions. EC was supported by a GRF grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 15612222) and by a grant from the PROCORE France/Hong Kong Joint Research Scheme (Project No. F-PolyU501/21). This research was also partly funded by PNRR—M4C2—Investimento 1.3, Partenariato Esteso PE00000013—"FAIR—Future Artificial Intelligence Research"—Spoke 1 "Human-centered AI," funded by the European Commission under the NextGeneration EU programme. This work was carried out while DT was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Fondazione Bruno Kessler.

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Open Research

Data, scripts, and detailed results are hosted on the Open Science Framework https://osf.io/c7f4u/?view_only=201206d3d8574e9b84429419be4587e7.

References

Andreas, J. (2022). Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 5769–5779). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Baggio, G. (2018). *Meaning in the brain*. MIT Press.

Baggio, G. (2021). Compositionality in a parallel architecture for language processing. *Cognitive Science*, *45*(5), e12949.

Bommasani, R., Davis, K., & Cardie, C. (2020). Interpreting pretrained contextulalized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4758–4781).

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … & Percy, L. (2021). On the opportunities and risks of foundation models. *ArXiv: 2108.07258*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., 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., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901).

Buijtelaar, L., & Pezzelle, S. (2023). A psycholinguistic analysis of BERT's representations of compounds. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2222–2233).

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Chang, T. A., & Bergen, B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics*, *50*, 1–58.

Cong, Y., Chersoni, E., Hsu, Y., & Lenci, A. (2023). Are language models sensitive to semantic attraction? A study on surprisal. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics* (pp. 141–148).

Culicover, P. W., & Jackendoff, R. (2005). *Simpler syntax*. Oxford University Press.

Culicover, P. W., & Jackendoff, R. (2006). The simpler syntax hypothesis. *Trends in Cognitive Sciences*, *10*(9), 413–418.

Dankers, V., Lucas, C., & Titov, I. (2022). Can transformer be too compositional? Analysing idiom processing in neural machine translation. In *Proceedings of ACL* (pp. 3608–3626).

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness. A connectionist perspective on development*. MIT Press.

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

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. *ArXiv: 1901.05287*.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Reichart, R., Devore, S., Flinker, A., Hasenfratz, L., Levy, O., Hassidim, A., Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., & Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380.

Glavaš, G., & Vulić, I. (2021). Is supervised syntactic parsing beneficial for language understanding tasks? An empirical investigation. In *Proceedings of EACL* (pp. 3090–3104)

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*.
Hegel, G. W. F. (1979). *Phenomenology of spirit* (A. V. Miller, Trans.). Oxford University Press.
Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings NAACL-HLT* (pp. 4129–4138).
Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2023). A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of ACL* (pp. 4194–4213).
Jackendoff, R. (2007). A parallel architecture perspective on language processing. *Brain Research*, *1146*, 2–22.
Jackendoff, R. (1997). *The architecture of the language faculty*. MIT Press.
Kauf, C., Chersoni, E., Lenci, A., Fedorenko, E., & Ivanova, A. A. (2024). Comparing Plausibility Estimates in Base and Instruction-Tuned Large Language Models. *arXiv preprint arXiv:2403.14859*.
Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., Fedorenko, E., & Lenci, A. (2023). Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science*, *47*(11), e13386.
Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, *52*(2), 205–225.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.
Lenci, A. (2023). Understanding natural language understanding systems. A critical analysis. *ArXiv: 2303.04229*.
Lenci, A., & Sahlgren, M. (2023). *Distributional semantics*. Cambridge University Press.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.
Li, B., Zhu, Z., Thomas, G., Rudzicz, F., & Xu, Y. (2022). Neural reality of argument structure constructions. In *Proceedings of ACL* (pp. 7410–7423).
Lin, Y., Yi, C. T., & Frank, R. (2019). Open Sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (pp. 241–253).
Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.
Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., Swayamdipta, S., Smith, N. A., & Choi, Y. (2023). We're afraid language models aren't modeling ambiguity. In *Proceedings of EMNLP 2023*.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*: *1907.11692*.
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. *ArXiv*: *2301.06627*.
McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. *ArXiv: 2309.13638*.
McShane, M. J. (2005). *A theory of ellipsis*. Oxford University Press.
Michaelov, J., & Bergen, B. (2022). The more human-like the language model, the more surprisal is the best predictor of N400 amplitude. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*.
Michalon, O., & Baggio, G. (2019). Meaning-driven syntactic predictions in a parallel processing architecture: Theory and algorithmic modeling of ERP effects. *Neuropsychologia*, *131*, 171–183.
Miletić, F., & im Walde, S. S. (2023). A systematic search for compound semantics in pretrained BERT architectures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 1499–1512).
Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R., Keanm, H., Qian, P., & Fedorenko, E. (2020). Composition is the core driver of the language-selective network. *Neurobiology of Language*, *1*(1), 104–134.
Nedumpozhimana, V., & Kelleher, J. (2021). Finding BERT's idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)* (pp. 57–62).
Ormerod, M., Martínez del Rincón, J., & Devereux, B. (2024). How is a "kitchen chair" like a "farm horse"? Exploring the representation of noun-noun compound semantics in transformer-based language models. *Computational Linguistics*, 1–33.

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Pedinotti, P., Rambelli, G., Chersoni, E., Santus, E., Lenci, A., & Blache, P. (2021). Did the cat drink the coffee? Challenging transformers with generalized event knowledge. In *Proceedings *SEM 2021* (pp. 1–11).

Pezzelle, S. (2023). Dealing with semantic underspecification in multimodal NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 12098–12112). Toronto, Canada: Association for Computational Linguistics.

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

Prange, J., Schneider, N. & Kong, L. (2022). Linguistic Frameworks Go Toe-to-Toe at Neuro-Symbolic Language Modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4375–4391). Seattle, United States: Association for Computational Linguistics.

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.

Rambelli, G., Chersoni, E., Lenci, A., Blache, P., & Huang, C. R. (2020). Comparing probabilistic, distributional and transformer-based models on logical metonymy interpretation. In *Proceedings of AACL-IJCNLP* (pp. 224–234).

Rambelli, G., Chersoni, E., Senaldi, M. S. G., Blache, P, & Lenci, A. (2023). Are frequent phrases directly retrieved like idioms? An investigation with self-paced reading and language models. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)* (pp. 87–98).

Ruis, L. E., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2022). Large language models are not zero-shot communicators.

Schlangen, D. (2022). Norm participation grounds language. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment* (pp. 62–69). Gothenburg, Sweden: Association for Computational Linguistics.

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. In *Proceedings of ICLR 2019*.

Testa, D., Chersoni, E., & Lenci, A. (2023). We Understand Elliptical Sentences, and Language Models should Too: A New Dataset for Studying Ellipsis and its Interaction with Thematic Fit. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3340–3353). Toronto, Canada: Association for Computational Linguistics.

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

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., & Korhonen, A. (2020). Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7222–7240).

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S. F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *ArXiv: 2206.07682*.

17568765, 2025, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tops.12733 by Area Sistemi Dipart & Doc, Wiley Online Library on [13/07/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License