# Formal Linguistic Competence is not a Monolith: Consequences for LLMs

**Tommi Buder-Gröndahl (tommi.grondahl@helsinki.fi)**[1]
[1]Department of Digital Humanities, Faculty of Humanities, University of Helsinki

## Abstract

The capacities of large language models (LLMs) remain disputed, but one particular hypothesis has recently become widely endorsed: LLMs have a uniform human-level *formal linguistic competence* in grammar and semantics, while diverging in *functional competence* that extends to broader cognitive and social capacities. I raise two challenges with this account. First, the separation between *weak* and *strong generative capacity* should not be sidestepped. LLMs can process and produce grammatical text, but this does not reveal how they represent its structure. Second, the dissociability of the two competences does not eliminate influence between them. Distinct functional competences are not expected to coincide with identical formal competences, since the choice between different weakly equivalent grammars is likely regulated by functional competence. I further evaluate this hypothesis by reviewing experimental work on the behavioral and mechanistic analysis of LLMs, which does not point to a uniform formal competence across models.

**Keywords:** formal linguistic competence; large language models; grammar; syntax; semantics; linguistic representation; weak generative capacity; strong generative capacity

## Introduction

The impressive performance of large language models (LLMs) has led many to conclude that they represent grammatical structure similarly to humans (Kallens et al., 2023; Piantadosi, 2024). At the same time, they continue to be prone to errors and can "hallucinate" material that fails to reflect relevant information (Huang et al., 2025). To clarify the situation, it would be useful to find a generic taxonomy of LLMs' linguistic capacities. To date, the most prominent and well-developed proposal comes from Mahowald et al. (2024), who take LLMs to have mastered *formal* but not *functional linguistic competence*. The former is restricted to grammar and semantics, which the latter extends to abilities required for real-world language use. This account has become widely adopted in recent literature (Azaria et al., 2024; Kibria et al., 2024; Mitchell & Krakauer, 2023; Tuckute et al., 2024).

However, I challenge the taxonomy on two counts. First, *weak* and *strong generative capacity* should be distinguished, where the former concerns which strings are treated as grammatical, while the latter concerns abstract linguistic structures (Chomsky, 1959; Miller, 1999). Since weak equivalence does not guarantee strong equivalence, the grammaticality of an LLM's output does not suffice to reveal its "internal grammar". Out of all possible weakly equivalent grammars, only some are acquired. Since LLMs are architecturally holistic and lack innate language-specific partitions, their *functional* competence is expected to modulate this learning process. Hence, instead of a single uniform formal competence, LLMs are expected to have a range of diverse formal competences.

Second, the dissociability between formal and functional competence does not suffice to discard their mutual impact, and their interaction is still central for guiding grammar acquisition. This is especially clear in usage-based functional linguistics (Croft, 2001; Tomasello, 2003; Van Valin, 2016), although it also holds in generative approaches where the syntax-semantics interface plays a key role (e.g. Hale and Keyser 2002; Harley 2012; Wiltschko 2014). In particular – *contra* Weissweiler et al. (2025) – replacing abstract rule-based grammars with constructionist alternatives does not resolve the issue, but even accentuates it.

Formal competence is thus not a single capacity, but divisible into numerous variants. Divergences across both behavioral findings (Murphy et al., 2025; Qiu et al., 2025) and probing results (Bonial & Madabushi, 2025; Diego-Simón et al., 2025) indeed suggest that LLMs can acquire varying grammars, even if these overlap in weak generative capacity. In light of this, I discuss both theoretical and methodological ramifications of rethinking formal competence in LLMs.

## Large Language Models

State-of-the-art LLMs are predominantly variants of the *transformer* architecture (Vaswani et al., 2017). Operating on vector representations of tokens – known as *embeddings* – they have been trained with massive amounts of text on a generic task, such as masked token prediction for BERT-type models (Devlin et al., 2019), or next token prediction for autoregressive models (such as GPT). Further refinement is possible via reinforcement learning with human feedback (Ouyang et al., 2022) or instruction tuning (S. Zhang et al., 2026).

## Formal Vs. Functional Linguistic Competence

Mahowald et al. (2024) describe formal competence as "knowledge of linguistic rules and patterns" (p. 517), divided between phonology, morphology, lexical semantics, and syntax (pp. 519–520). In line with generative grammar, they treat syntactic phrases as being "captured by a tree-like structure" (p. 519). Within the human brain, formal competence has been localized to a dedicated fronto-temporal network (Fedorenko et al., 2024; Fedorenko & Varley, 2016).



In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* 

Likewise, Mahowald et al. divide functional competence into four categories: formal reasoning (logic, mathematics), world knowledge (memorized facts), situation modeling (tracking discourse participants), and social reasoning (understanding communication contexts). These are not language-specific capacities, but are required for successful real-world language use. Mahowald et al. propose that LLMs have formal competence with natural language, but their functional competence is more varied and falls short of that attained by humans. This taxonomy has quickly become ubiquitously adopted in both theoretical and empirical literature (Azaria et al., 2024; Kibria et al., 2024; Lu et al., 2024; Mitchell & Krakauer, 2023; Momennejad et al., 2023; Piantadosi, 2024; Tuckute et al., 2024; Wu et al., 2024).

## Challenges with the Taxonomy

I raise two theoretical problems with the analysis, in light of which I evaluate notable empirical findings on LLMs.

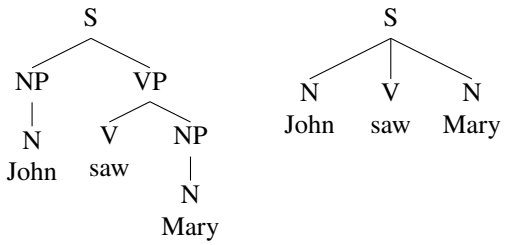
### Linguistic Theory

**Weak and Strong Generative Capacity** Standard formal language theory defines languages extensionally as sets of strings, which grammars generate or classify. Grammars that generate/accept the same strings have the same *weak generative capacity*, and are thereby *weakly equivalent* (Chomsky, 1959). For example, consider context-free grammars $G_1$ and $G_2$ with lexical insertion rules N → *John*|*Mary* and V → *saw*:

($G_1$) S → NP VP
VP → V NP
NP → N

($G_2$) S → N V N

$G_1$ and $G_2$ are weakly equivalent since they accept the same two sentences: *John saw Mary* and *Mary saw John*. Still, they are crucially distinct. As shown below, $G_1$ places the verb and object into a single verb phrase (on the left), whereas $G_2$ gives the sentence a flat structure (on the right):




We can say that the grammars assign different *structural descriptions* to the same strings – that is, they have different *strong generative capacities* (Miller, 1999). Structural descriptions are defined theory-internally, and prominent contenders include phrases, dependency graphs, or constructions. All of these are irreducible to mere linear properties of strings, and thereby allow several weakly equivalent grammars with distinct strong generative capacities.

Consequently, discovering that a cognitive system can process a particular set of strings is insufficient to conclude that it has parsed them via a particular grammar $G_i$ out of all strongly non-equivalent but weakly equivalent grammars $\{G_1, ..., G_n\}$. To assess which grammar (if any) it has employed, further means of evaluation are needed. In the taxonomy of Marr (1982), strong generative capacity belongs to the high *computational* level, realized by lower-level algorithms. In abstract, the goal of explaining formal linguistic competence in terms of strong generative capacity can be construed as mapping algorithmically defined lower-level states to structural descriptions that appropriately capture their high-level computational properties (Buder-Gröndahl, 2024a, 2025; Dupre, 2021).[1]

A focus on strong rather than weak generative capacity characterizes much of contemporary theoretical linguistics (Ott, 2017), and is adopted by Mahowald et al. (2024) as well. As reviewed above, their conception of formal competence goes beyond weak generative capacity by explicitly involving structural descriptions such as hierarchical phrase structure.

**Generative Linguistics** Mahowald et al. (2024) explicitly state that LLMs have internalized syntactic phrase structures familiar from generative linguistics (pp. 517, 522, 527). A *prima facie* challenge here is that generativists are typically committed to a nativist picture of language being largely shaped by innate cognitive capacities (Chomsky, 1965). This "universal grammar" is, of course, a contentious postulate (Dąbrowska, 2015; Evans, 2014). But, whatever its status in human cognition, it is at least clearly absent in LLMs.

However, while this concern has some initial force, it is not decisive, After all, the formal structure of linguistic representations is not the same as their etiology. Perhaps an innate universal grammar is not the only way to obtain phrase-structural representations. If so, what is the alternative?

The most prominent contender is the *interface* between grammar and other cognitive systems. Current mainstream generative linguistics typically treats the syntactic component as very generic, consisting only of a basic combinatory operation (Hauser et al., 2002). This makes syntax itself insufficient to block most ungrammatical constructions, which invites appealing to interface constraints. A central idea here is that phrase structure mirrors thematic argument structure (Baker, 1988; Borer, 2005; Hale & Keyser, 2002; Ramchand, 2008), such as Davidsonian event semantics (Davidson, 1967; Lohndal, 2014; Parsons, 1990; Pietroski, 2018) or situation semantics (Barwise & Perry, 1983; Ramchand, 2018).

As a concrete illustration, Ramchand and Svenonius (2014) posit three syntactic domains linked to dedicated interpreta-

[1] It might be asked if strong generative capacity could be restricted to a purely computational level without the need for *any* algorithmic realization. However, if these levels are fully divorced, weakly equivalent grammars again become assimilated. If computational analysis is construed purely behaviorally without the need of some lower-level algorithm to realize it, structural descriptions that share extensional coverage become equivalent. Multiple realizability entails that no particular realization is necessary; not that they could all be absent. (I thank an anonymous reviewer for raising this issue.)

tions, where the first specifies an event, the second specifies a situation, and the third links the situation to a proposition by anchoring it to the discourse (Giorgi, 2010). In a related account, Wiltschko (2014) proposes four semantically grounded domains with varying language-specific manifestations. The *classification* domain determines an ontological type and thematic structure. The *point-of-view* and *anchoring* domains add further information that relates the entity to the context. Finally, the *linking* domain adds information about the whole speech act, such as illocutionary force.

Mahowald et al. (2024) characterize functional competence as involving "factual and commonsense knowledge about agents, objects, properties, actions, events, and ideas" as well as "the dynamic tracking of objects, agents, and events as a narrative/conversation unfolds over time" (p. 526). The former is clearly required for understanding event structure and thematic roles, while the latter underlies situation semantics along with the representation of discourse (c.f. Kamp 1995). Indeed, Mahowald et al. dub it *situation modelling*.

If syntax and semantics reflect each other, this is expected to guide their co-development during language acquisition – particularly if innate constraints are shunned. We would thus not expect two cognitive systems with markedly different functional competences to develop identical formal competences.

To be clear, not all variants of generative theory stipulate a highly transparent syntax-semantics interface. In particular, the *Simpler Syntax* model (Culicover & Jackendoff, 2005) proposes that syntax does not mirror semantics; instead, it only contains the minimal structure needed for articulation and conceptual interpretation. Nevertheless, this does not make syntax independent of extra-linguistic influence. The minimal structure needed for conceptual interpretation evidently depends on which conceptual representations are present on the other side. Changing them would alter syntax as well.

Overall, then, there are major challenges in trying to fit generative accounts of syntax to the analysis of Mahowald et al. (2024). Even if grammar and conceptual cognition are represented in different parts of the human brain, their links are still crucial for the specific grammatical structures acquired. Interface-driven syntax requires *both sides* of the interface.

But perhaps the generative literature is just the wrong place to look, given that its nativist assumptions stand in clear contrast to LLMs as data-driven learning systems (Piantadosi, 2024). Even though Mahowald et al. explicitly posit phrase structure as the format of syntax in LLMs, their main claim can be reconstructed via alternative frameworks.

**Functionalist Linguistics** Tomasello (2015) summarizes the usage-based approach to linguistics via two principles: (1) meaning is use, and (2) structure emerges from use (p. 69). Language acquisition is traced back to two main cognitive skills: pattern recognition and understanding communicative intentions. This perspective is typically combined with some variant of *construction grammar*, which replaces the traditional lexicon with a unified repository of memorized form-meaning pairs (Goldberg, 2006; Kay & Fillmore, 1999).

The central idea behind usage-based linguistics is that generic socio-cognitive processes are responsible for the emergence of patterns that capture conceptual and communicative functions. These include categorization, associative memory, prototypes, exemplars, analogy, iconicity, metaphor, and frequency effects such as entrenchment (Bybee, 2010; Ibbotson, 2013). All but the last are unequivocally allocated to functional competence in the taxonomy of Mahowald et al. (2024). The same holds even more evidently for *cognitive linguistics* (Evans & Green, 2006; Lakoff, 1987; Langacker, 1987), where the guiding idea is that language arises from pre-linguistic cognitive processes at play in perception, motor control, emotion, and social interaction. Such capacities are plainly allocated to functional competence.

Working within the functionalist picture of language, Goldberg (2025) recognizes that it would be initially surprising if LLMs had human-like linguistic capacities since their inputs are limited to text. Nevertheless, she maintains that they have indeed attained such capacities, and takes this to bolster the constructionist approach to grammar. This initially seems to converge with the proposal of Mahowald et al. (2024).

However, a careful look at Goldberg's argument shows that she is specifically relying on LLMs' increased functional competence. In her assessment, a key leap took place with the adoption of instruction tuned models trained via human feedback (Ouyang et al., 2022). She further provides small case studies by prompting GPT-4 with social inferences, explaining unusual constructions, mathematical problems, and characterizing conceptual metaphors. As these are aspects of functional competence, her account actually ends up *contrasting* Mahowald et al.'s. If the formal competence of LLMs has changed due to their enhanced functional capacities, this illustrates that the competences are linked rather than dissociated.

This point extends to a recent suggestion that usage-based grammar shows LLMs to be more human-like than often thought. Weissweiler et al. (2025) call for rejecting abstract rule-based formulations familiar from formal (especially generative) linguistics, and replacing them with a network of memorized constructions with varying levels of specificity (c.f. Croft 2001; Goldberg 2006). Unlike abstract static rules, constrctions are flexible, can be only partially productive, and display sensitivity to similarity and frequency effects underlying statistical learning (Kallens et al., 2023). Weissweiler et al. also emphasize the need to move beyond binary grammaticality judgements toward graded assessments (Juzek, 2024) and naturalistic data with a better inclusion of context effects.

Indeed, critiques of LLMs have often stemmed from generative perspectives (e.g. Dupre 2021; Katzir 2023), which can facilitate a general affinity between linguistic functionalism and a favorable attitude toward the capacities of LLMs. My present focus, however, is not this debate but the separation between formal and functional competence: is the first uniform but not the second? As Boleda (2025) observes, a strict division between grammar and broader conceptual capacities is particularly dubious within functionalist frameworks:

> (...) as pointed out in functional approaches like cognitive linguistics, interactions between grammar and conceptual aspects of meaning are pervasive in language; and there is no clear dividing line between semantic properties that are relevant vs. irrelevant for grammar (Langacker 1987; Fillmore et al. 1988; Goldberg 1995). Thus, the strict division betweeen linguistic-semantics and other-semantics is questionable.
> (Boleda, 2025, pp. 9368–9369)

This is hardly surprising – after all, a strict formal-functional distinction would be anathema to the foundational notion that both meaning and grammatical structure emerge from language use. Whatever the general merits of linguistic functionalism may be, it does not reflect Mahowald et al.'s taxonomy.

**Summary** Mahowald et al. (2024) maintain that formal competence is uniformly shared between LLMs (and humans) even though their functional competence exhibits divergence. Linguistic theory provides two major reasons to doubt this taxonomy. Contemporary generative perspectives retain a distinction between grammar and conceptual cognition, but nevertheless treat their interface as explanatorily crucial. Functionalist approaches, in turn, shun a clear formal-functional division to begin with. From both standpoints, cognitive systems that differ significantly in functional competence are not expected to develop identical formal competences, especially once innate language-specific constraints are ruled out.

## Formal Linguistic Competence in LLMs

So far, I have argued that convergence in formal linguistic competence across LLMs would be surprising, as long as they diverge in functional competence. But surprising things can happen, and the matter is not resolved from the armchair. Therefore, I now move on to assessing whether experimental literature points towards convergence in formal competence.

**Grammaticality Classification** While the methodological limits of grammaticality judgements have been widely recognized (Juzek & Häussler, 2020; Ott, 2017; Schütze, 1996), they continue to make up much of the data used in linguistic literature. Their main benefit arises from allowing targeted experimentation for evaluating hypotheses concerning the underlying linguistic competence (Gross, 2021). However, results on LLM judgements remain highly mixed, with verdicts ranging from the proposal that LLMs "are better than theoretical linguists at theoretical linguistics" (Ambridge & Blything, 2024, p. 33) to the notion that they "have not yet grasped formal language competence" (Murphy et al., 2025, p. 1).

Accurate judgements are reported by Mahowald (2023) for one English construction on GPT-3, and Ambridge and Blything (2024) for English causatives on text-davinci-002. Cai et al. (2024) show alignment between ChatGPT and humans in 10/12 psycholinguistic tests. Qiu et al. (2025) adopt grammaticality judgement tasks from a prior study on human respondents (Sprouse et al., 2013), and report significant correlations between responses by ChatGPT and humans.

In contrast, Dentella et al. (2023) report negative results for eight syntactic constructions on davinci2/3 (based on GPT-3) and ChatGPT-3.5. Leivada et al. (2023) illustrate Dall·E's insensitivity to grammatical phenomena such as binding principles, thematic roles, negation, and coordination. Dentella et al. (2024) reveal both syntactic and semantic difficulties with Bard, ChatGPT-3.5, ChatGPT-4, Falcon, Gemini, Llama-2, and Mixtral-8. Recently, Murphy et al. (2025) assessed the o3-mini reasoning model across several linguistic tasks, observing failures at detecting ungrammatical sentences, hallucination of inappropriate structures, troubles with generating grammar violations when prompted to do so, and grammaticality assessments that contradict syntactic theory.

One possible reaction to this predicament would be to assign judgements to *functional* rather than formal competence. Song et al. (2025) observe a discrepancy between LLMs' performance in metalinguistic tasks and the string probabilities they assign. For example, even if a model fails at judging which of two sentences is grammatical, it might still assign a higher probability to the grammatical sentence. Based on this, Song et al. argue that judgements only reflect introspection as part of functional competence. This interpretation would initially fit well with Mahowald et al.'s taxonomy, with string probabilities residing in formal competence and grammaticality judgements measuring functional competence.

However, string probabilities alone fail to capture *why* a string might be assigned a low probability. Ungrammaticality is only one possible reason for this, others including semantic or pragmatic oddity, rarity, taboo status, etc. Some of these belong to formal and others to functional competence, but probabilities do not wear such labels on their sleeves. Moreover, given the holistic architecture of LLMs, probabilities are most likely the joint outcome of both formal and functional factors. This hinders their assimilation to formal competence. Empirical evidence also shows that there is no clear cut-off point between probabilities that LLMs assign to grammatical and ungrammatical strings (Leivada et al., 2024).

As a remedy, J. Hu et al. (2026) propose to replace raw string probabilities with probability contrasts across minimal pairs. If $S_1$ and $S_2$ express the same message, $S_1$ is grammatical and $S_2$ is ungrammatical, models should assign a higher probability to $S_1$ than to $S_2$. J. Hu et al. corroborate this hypothesis with both English and Chinese variants of GPT-2 and Llama-3 across several minimal pair datasets. This is indeed in many ways a superior proxy for grammaticality classification than raw probabilities, since the latter do not control for other factors such as lexical rarity or semantic oddity. Nevertheless, without denying its usefulness as a source of evidence for LLMs' weak generative capacity, I raise two considerations that prevent its equivocation with formal competence.

First, not all grammaticality judgements can be replaced by minimal pair comparisons, since many ungrammatical expressions lack a grammatical counterpart for expressing the same message. For example, English does not allow turning a noun head into a *wh*-pronoun while retaining its modifiers:

(1) She bought some fresh carrots

(2) a. *What did she buy some fresh?
b. *What some fresh did she buy?
c. *Some fresh what did she buy?

Even though the intended meaning of (2a–c) is semantically intelligible, no grammatical sentence can express it. Hence, no minimal pair can be formed. Capturing the ungrammaticality of (2a–c) thus requires other means. Given the unreliability of raw probabilities, it is unclear what would be available aside of grammaticality judgements.

Second, minimal pair comparisons do not concern which structural descriptions (if any) the model has internalized. Suppose one model processes transitive sentences in line with grammar $G_1$ and another in line with grammar $G_2$ (see above). Both would produce the same comparisons across grammatical and ungrammatical minimal pairs, but these results could not reveal which grammar each model represents. Minimal pairs cannot get us from weak to strong generative capacity.

Similar limitations apply to benchmarks such as SyntaxGym (Gauthier et al., 2020), which compares model surprisal at different tokens between grammatical and ungrammatical inputs. For example, the model should have a high surprisal for a verb token with faulty subject agreement. This does not reveal the grammatical structure over which the relevant relations (e.g. agreement) are computed, and is thereby compatible with all weakly equivalent analyses of the data.

It is also crucial to distinguish strong generative capacity from model-internal causal pipelines. CausalGym (Arora et al., 2024) modifies SyntaxGym into an intervention-based mechanistic interpretation method (c.f. Geiger et al. 2021; Geiger et al. 2022) that locates circuits for altering the output in a grammatically relevant way, such as changing verb agreement. While this is more informative than mere surprisal, it still does not uncover representations of specific structural descriptions. CausalGym could find the circuit for subject-verb-agreement in transitive clauses; but this still would not tell whether it represented these clauses in line with grammar $G_1$ or $G_2$ (or neither). Even though mechanistic interpretation goes "inside" the model, it does not yet guarantee the assessment of strong rather than weak generative capacity.

That being said, mechanistic analysis has a crucial role for revealing influence between different capacities. Recently, Hanna et al. (2026) compared grammatical and functional tasks across five LLMs (Llama-3, Qwen-2.5, OLMo, Mistral-v0.3, Gemma-2), finding low but non-zero overlap across circuits. In light of the independently discovered semantic influence on attention heads specialized for syntactic dependencies (McGee et al., 2026), these results fit the hypothesis that formal and functional competence are partly dissociable in architectural terms but not fully independent.

It also bears emphasis that my present account is compatible with contemporary LLMs having improved in formal competence when compared to prior ones. It simply predicts that this should correlate with a superior functional competence. That being said, recent research has continued to uncover difficulties that contemporary LLMs face in central functional capacities, including logical reasoning (Ariyani et al., 2025; Beyer & Reed, 2025; Han et al., 2024), the recognition of events (Chen et al., 2024; Z. Hu et al., 2025), social reasoning (Bortoletto et al., 2025; Liu et al., 2026; Wagner et al., 2025), and understanding negation (Kim et al., 2025; Seo et al., 2025; Varshney et al., 2025; Y. Zhang et al., 2024). Especially the last of these highlights a striking problem with drawing a clear formal-functional distinction; negation being both a grammatical marker and a logical operator. Overall, while formal and functional competence have (co-)evolved in recent years, important challenges still remain in both.[2]

**Grammatical Probing** Properly assessing the strong generative capacity of LLMs requires going beyond the strings they produce or classify, toward the structural representations they have incorporated. While significant effort has been put into explaining LLMs' internal mechanisms (Zhao et al., 2024), their computational outlook still remains contested.[3] *Probing* aims to extract information about LLMs' representations and operations. A prominent range of methods uses (linear or non-linear) transformations for mapping embeddings to grammatical labels, or relations between embeddings to syntactic relations (Hewitt & Manning, 2019; Manning et al., 2020; Müller-Eberstein et al., 2022; Tenney et al., 2019; White et al., 2021; Wu et al., 2020). Probing also covers certain behavioral methods such as instruction-based prompting (Bonial & Madabushi, 2025; Dunn & Eida, 2025; Li et al., 2022).

Hewitt and Manning's (2019) probe maps word embeddings to parse distances, which Diego-Simón et al. (2024) appended with directionality and dependency labels. Training these probes on a Universal Dependencies treebank (Silveira et al., 2014), Diego-Simón et al. (2025) assessed which properties are most predictive of their accuracy with three LLMs (Mistral, Llama-2, BERT). They found both probes to be most sensitive to linear distance between words and parse depth.

Kennedy (2025) applied Hewitt and Manning's (2019) probe to four LLMs (BERT, RoBERTa, GPT-2, Qwen2.5) using Subject Raising and Subject Control sentences, which share the same dependency graph but have different phrase structures in standard generative analyses (Rosenbaum, 1967). Despite being trained only on dependencies, the probe assigned different parses to these constructions, in line with the hypothesis that Subject Control sentences have a more

[2]Moreover, smaller models are increasingly employed in domain-specific applications – motivated by resource limitations, efficiency, and adaptability (Wang et al., 2025). Variation in performance across actually deployed models is thus not automatically expected to disappear even as top benchmark performance improves.

[3]Ambridge and Blything (2024, pp. 39–40) maintain that we do actually know how LLMs work, since they are defined via algorithms such as the transformer (Vaswani et al., 2017). However, as Müller (2025, p. 3) notes, this is only true on the low algorithmic level of explanation, which does not yet disclose the higher-level computation enacted within the network (c.f. Dunbar 2019). Analogous concerns arise in neurolinguistics as well (van der Burght et al., 2023).

complex underlying structure. However, the effect was inconsistent across word pairs: it was often missing between the subject and the infinitive marker (*to*) despite being present between the subject and other elements in the embedded clause (the embedded subject or the direct object). As Kennedy (2025, p. 384) notes, such an "arbitrary and alien" structure is not sensible within any known syntactic analysis of English.

To mitigate the challenge of assuming target parses *a priori*, Wu et al. (2020) provide a parameter-free probe for BERT, which constructs dependency graphs or phrase structures from mutual impact measures between embeddings. While their initial results indicated similarity to human parses, subsequent work contrasts this. Looking further into Wu et al.'s method for constructing phrase structures, Niu et al. (2022) concluded that it only partially encodes syntactic information, and barely exceeds a right-branching baseline. Buder-Gröndahl (2024b) compared dependency graphs obtained via the probe to Universal Dependency parses, and found major discrepancies between these especially in embedded clause structure as well as various modifier-head relations.

In construction probing, differences have arisen between models of different size. Li et al. (2022) probed variants of BERT for argument structure constructions via embedding clustering. Smaller models grouped sentences more based on lexical content, while larger models were more sensitive to construction-level information. Bunzeck and Zarrieß (2023) likewise observed variation across smaller and larger RoBERTa models in the influence of part-of-speech information for sentence embeddings.

Constructions obtained via probing can also markedly differ from those proposed in the linguistic literature. Veenboer and Bloem (2023) found that similarities between BERT's embeddings correlate with constructional information, but fails to distinguish between semantically similar constructions that differ in grammatical form. Dunn and Eida (2025) illustrated that both metalinguistic prompting and embedding-based probing of GPT-4 reveals sensitivity to "hallucinated" constructions that are absent from English based on human judgement. Bonial and Madabushi (2025) surveyed experimental literature on construction probing, and extended on it by prompting GPT-3.5 and GPT-4 to find sentences that belong to the same construction. Summarizing both prior work and the novel results, they propose a *Schematicity Hypothesis* concerning constructions in LLMs:

> LLMs have access to the lower levels of the constructional hierarchy and more substantive constructions, but do not have the ability to distinguish higher-order, fully schematic constructions requiring recognition of both the form and associated meaning pole.
> (Bonial & Madabushi, 2025, p. 4571)

This stands in notable contrast to the results of Diego-Simón et al. (2025), who found that structural propertes are relevant for probing syntax whereas statistical effects specific to lexical content are not. In part, it seems that these differences trace back to the linguistic formalisms used as probing targets: dependency probes are more responsive to structural information, while construction probes are more schematic and sensitive to lexical content. While comparison between target formalisms remains severely under-studied, models have indeed shown preferences toward some formalisms over others in both syntactic dependencies (Kulmizev et al., 2020) and semantic argument structure (Kuznetsov & Gurevych, 2020).

**Summary** Experimental results on both grammaticality classification and probing point to significant discrepancies across LLMs, when focus is shifted from weak to strong generative capacity. That being said, Song et al. (2025) are undoubtedly correct that performance in metalinguistic tasks is far from a direct measure of formal competence. Probing results are also influenced by numerous factors, such as the probing algorithm and training scheme. Hence, the results deserve a modest but nevertheless crucial *negative* interpretation: empirical evidence does *not* support the hypothesis of a uniform formal competence across LLMs. The hypothesis is not thereby falsified, but lacks independent support.

## Conclusions

Mahowald et al. (2024) construct a comprehensive taxonomy of linguistic capacities in both humans and LLMs. This is commendable, and thus far no superior alternative has been put forth. That notwithstanding, I contend that LLMs should not be analyzed as having a uniform formal competence. Theoretically, it would be unexpected for identical formal competences to occur with distinct functional competences. This remains compatible with the competences being constitutively distinct (Fedorenko et al., 2024; Hanna et al., 2026); what matters is that their *interaction* shapes grammar. This is in line with empirical findings, which do not point to a uniform strong generative capacity across LLMs. Of course, variation here can be traced to several possible sources, such as model architecture, training data, or tokenization. However, these factors are also likely to influence functional competence, in line with the hypothesis that the competences covary.

I do not deny that LLMs can converge in *weak* generative capacity, which is operationalizable via minimal pair comparisons (albeit only in part). Some recent analyses of LLMs (e.g. Gastaldi and Pellissier 2021; Mickus 2024) have moved toward construing languages purely extensionally, in line with certain variants of classical structuralism (Bloomfield, 1936; Harris, 1951). For this perspective, convergence in formal competence is a far more viable prospect: if nothing is admitted beyond weak generative capacity, the problem certainly becomes more manageable (c.f. Quine 1970). This, however, is not what either Mahowald et al. (2024) or I have been after.

Once strong generative capacity is brought to the forefront, formal linguistic competence should no longer be seen as uniform or independent of functional competence. Far from being resolved, uncovering it in LLMs calls for significant further attention. Despite their shortcomings, grammaticality judgements and probing continue to play crucial roles here, and are not surpassed by string probability comparisons alone.

## Acknowledgments

This research was supported by the Strategic Resource Funding, Research Council of Finland, number 365541. I thank Anna-Mari Wallenberg, Riikka Möttönen, and the anonymous reviewers for valuable feedback. No AI was used for writing. The project was conducted without conflicts of interest, financial or otherwise.

## References

Ambridge, B., & Blything, L. (2024). Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics*, *50*(1–2), 33–48. https://doi.org/https://doi.org/10.1515/tl-2024-2002

Ariyani, N. F., Bouraoui, Z., Booth, R., & Schockaert, S. (2025). There's no such thing as simple reasoning for LLMs. *Findings of the Association for Computational Linguistics: ACL 2025*, 4503–4514. https://doi.org/10.18653/v1/2025.findings-acl.232

Arora, A., Jurafsky, D., & Potts, C. (2024). CausalGym: Benchmarking causal interpretability methods on linguistic tasks. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14638–14663. https://doi.org/10.18653/v1/2024.acl-long.785

Azaria, A., Azoulay, R., & Reches, S. (2024). ChatGPT is a remarkable tool–for experts. *Data Intelligence*, *6*(1), 240–296. https://doi.org/10.1162/dint_a_00235

Baker, M. (1988). *Incorporation: A theory of grammatical function changing*. University of Chicago Press.

Barwise, J., & Perry, J. (1983). *Situations and attitudes*. MIT Press.

Beyer, H., & Reed, C. (2025). Lexical recall or logical reasoning: Probing the limits of reasoning abilities in large language models. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13532–13557. https://doi.org/10.18653/v1/2025.acl-long.664

Bloomfield, L. (1936). Language or ideas. *Language*, *12*(2), 89–95.

Boleda, G. (2025). LLMs as a synthesis between symbolic and distributed approaches to language. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 9365–9379. https://doi.org/10.18653/v1/2025.findings-emnlp.498

Bonial, C., & Madabushi, H. T. (2025). Constructing understanding: On the constructional information encoded in large language models. *Language Resources and Evaluation*, *59*, 4559–4598. https://doi.org/https://doi.org/10.1007/s10579-024-09799-9

Borer, H. (2005). *The nominal course of events: Structuring sense, volume II*. Oxford University Press.

Bortoletto, M., Ruhdorfer, C., & Bulling, A. (2025). ToM-SSI: Evaluating theory of mind in situated social interactions. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 32264–32289. https://doi.org/10.18653/v1/2025.emnlp-main.1642

Buder-Gröndahl, T. (2024a). The ambiguity of BERTology: What do large language models represent? *Synthese*, *203*, 15. https://doi.org/10.1007/s11229-023-04435-5

Buder-Gröndahl, T. (2024b). What does parameter-free probing really uncover? *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 327–336. https://doi.org/10.18653/v1/2024.acl-short.31

Buder-Gröndahl, T. (2025). In defence of mathematical content. *Philosophical Psychology*, 1–33. https://doi.org/10.1080/09515089.2025.2530010

Bunzeck, B., & Zarrieß, S. (2023). Entrenchment matters: Investigating positional and constructional sensitivity in small and large language models. *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, 25–37.

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

Cai, Z., Duan, X., Haslett, D., Wang, S., & M, P. (2024). Do large language models resemble humans in language use? *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 37–56. https://doi.org/10.18653/v1/2024.cmcl-1.4

Chen, M., Ma, Y., Song, K., Cao, Y., Zhang, Y., & Li, D. (2024). Improving large language models in event relation logical prediction. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9451–9478. https://doi.org/10.18653/v1/2024.acl-long.512

Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control*, *2*(2), 137–167. https://doi.org/10.1016/S0019-9958(59)90362-6

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Croft, W. A. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

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

Dąbrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it? *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00852

Davidson, D. (1967). The logical form of action sentences. In N. Resher (Ed.), *The logic of decision and action* (pp. 81–95). University of Pittsburgh Press.

Dentella, V., Günther, F., & Leivada, E. (2023). Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(51), e2309583120. https://doi.org/https://doi.org/10.1073/pnas.2309583120

Dentella, V., Günther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Re-*

*ports*, *14*, 28083. https://doi.org/https://doi.org/10.1038/s41598-024-79531-8

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Diego-Simón, P. J., Chemla, E., King, J.-R., & Lakretz, Y. (2025). Probing syntax in large language models: Successes and remaining challenges. *Second Conference on Language Modeling*.

Diego-Simón, P. J., d'Ascoli, S., Chemla, E., Lakretz, Y., & King, J.-R. (2024). A polar coordinate system represents syntax in large language models. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://doi.org/10.52202/079017-3344

Dunbar, E. (2019). Generative grammar, neural networks, and the implementational mapping problem: Response to Pater. *Language*, *95*(1), e87–e98. https://doi.org/10.1353/lan.2019.0013

Dunn, J., & Eida, M. M. (2025). LLMs learn constructions that humans do not know. *Proceedings of the Second International Workshop on Construction Grammars and NLP*, 13–23.

Dupre, G. (2021). (What) can deep learning contribute to theoretical linguistics? *Minds and Machines*, *31*(7), 617–635. https://doi.org/10.1007/s11023-021-09571-w

Evans, V. (2014). *The language myth: Why language is not an instinct*. Cambridge University Press.

Evans, V., & Green, M. (2006). *Cognitive linguistics: An introduction*. Edinburgh University Press.

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, *25*, 289–312. https://doi.org/10.1038/s41583-024-00802-4

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

Fillmore, C., Kay, P., & O'Connor, C. (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language*, *64*(3), 501–538. https://doi.org/10.2307/414531

Gastaldi, J. L., & Pellissier, L. (2021). The calculus of language: Explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*, *46*(4), 569–590. https://doi.org/10.1080/03080188.2021.1890484

Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. (2020). SyntaxGym: An online platform for targeted evaluation of language models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 70–76. https://doi.org/10.18653/v1/2020.acl-demos.10

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. *Proceedings of the 35th International Conference on Neural Information Processing Systems*.

Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., & Potts, C. (2022). Inducing causal structure for interpretable neural networks. *Proceedings of the 39th International Conference on Machine Learning*, 7324–7338.

Giorgi, A. (2010). *About the speaker: Towards a syntax of indexicality*. Oxford University Press.

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

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Goldberg, A. E. (2025). Usage-based constructionist approaches and large language models. *Constructions and Frames*, *16*(2), 220–254. https://doi.org/10.1075/cf.23017.gol

Gross, S. (2021). Linguistic judgements as evidence. In N. Allott, T. Lohndal, & G. Rey (Eds.), *A companion to Chomsky* (pp. 544–556). Wiley Blackwell.

Hale, K., & Keyser, S. J. (2002). *Prolegomen to a theory of argument structure*. MIT Press.

Han, S., Schoelkopf, H., Zhao, Y., Qi, Z., Riddell, M., Zhou, W., Coady, J., Peng, D., Qiao, Y., Benson, L., Sun, L., Wardle-Solano, A., Szabó, H., Zubova, E., Burtell, M., Fan, J., Liu, Y., Wong, B., Sailor, M., ... Radev, D. (2024). FOLIO: Natural language reasoning with first-order logic. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 22017–22031. https://doi.org/10.18653/v1/2024.emnlp-main.1229

Hanna, M., Belinkov, Y., & Pezzelle, S. (2026). Are formal and functional linguistic mechanisms dissociated in language models? *Computational Linguistics*, *52*(1), 43–83. https://doi.org/10.1162/COLI.a.24

Harley, H. (2012). Semantics in Distributed Morphology. In K. von Heusinger & P. Portner (Eds.), *Semantics. an international handbook of natural language meaning vol.3.* (pp. 2151–2171). Mouton de Gruyter.

Harris, Z. S. (1951). *Methods in structural linguistics*. The University of Chicago Press.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. *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)*, 4129–4138. https://doi.org/10.18653/v1/N19-1419

Hu, J., Wilcox, E. G., Song, S., Mahowald, K., & Levy, R. P. (2026). What can string probability tell us about grammaticality? *Transactions of the Association for Computational Linguistics*, *14*, 124–146. https://doi.org/10.1162/tacl.a.611

Hu, Z., Li, Z., Jin, X., Bai, L., Guo, J., & Cheng, X. (2025). Large language model-based event relation extraction with rationales. *Proceedings of the 31st International Conference on Computational Linguistics*, 7484–7496.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/https://doi.org/10.1145/3703155

Ibbotson, P. (2013). The scope of usage-based theory. *Frontiers in Psychology*, *4*, 255. https://doi.org/https://doi.org/10.3389/fpsyg.2013.00255

Juzek, T. S. (2024). The syntactic acceptability dataset (preview): A resource for machine learning and linguistic analysis of English. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 16113–16120.

Juzek, T. S., & Häussler, J. (2020). Data convergence in syntactic theory and the role of sentence pairs. *Zeitschrift für Sprachwissenschaft*, *39*(2), 109–147. https://doi.org/10.1515/zfs-2020-2008

Kallens, P. C., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, *47*(3), e13256. https://doi.org/10.1111/cogs.13256

Kamp, H. (1995). Discourse representation theory. In J.-O. Ö. J. Verschueren & J. Blommaert (Eds.), *Handbook of Pragmatics* (pp. 253–257). John Benjamins.

Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics*, *17*, e13153. https://doi.org/https://doi.org/10.5964/bioling.13153

Kay, P., & Fillmore, C. (1999). Grammatical constructions and linguistic generalizations: The what's X doing Y? construction. *Language*, *75*(1), 1–33. https://doi.org/10.2307/417472

Kennedy, M. (2025). Evidence of generative syntax in LLMs. *Proceedings of the 29th Conference on Computational Natural Language Learning*, 377–396. https://doi.org/10.18653/v1/2025.conll-1.25

Kibria, R., Dipta, S. I. U., & Adnan, M. A. (2024). On functional competence of LLMs for linguistic disambiguation. *Proceedings of the 28th Conference on Computational Natural Language Learning*, 143–160. https://doi.org/10.18653/v1/2024.conll-1.12

Kim, J., Koo, S., & Lim, H. (2025). Semantic inversion, identical replies: Revisiting negation blindness in large language models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 21434–21471. https://doi.org/10.18653/v1/2025.emnlp-main.1088

Kulmizev, A., Ravishankar, V., Abdou, M., & Nivre, J. (2020). Do neural language models show preferences for syntactic formalisms? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4077–4091. https://doi.org/10.18653/v1/2020.acl-main.375

Kuznetsov, I., & Gurevych, I. (2020). A matter of framing: The impact of linguistic formalism on probing results. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 171–182. https://doi.org/10.18653/v1/2020.emnlp-main.13

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago University Press.

Langacker, R. W. (1987). *Foundations of cognitive grammar, volume 1, theoretical prerequisites*. Stanford University Press.

Leivada, E., Günther, F., & Dentella, V. (2024). Reply to Hu et al.: Applying different evaluation standards to humans vs. large language models overestimates AI performance. *Proceedings of the National Academy of Sciences*, *121*(36), e2406752121. https://doi.org/10.1073/pnas.2406752121

Leivada, E., Murphy, E., & Marcus, G. (2023). Dall·e 2 fails to reliably capture common syntactic processes. *Social Sciences & Humanities Open*, *8*(1), 100648. https://doi.org/https://doi.org/10.1016/j.ssaho.2023.100648

Li, B., Zhu, Z., Thomas, G., Rudzicz, F., & Xu, Y. (2022). Neural reality of argument structure constructions. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 7410–7423). https://doi.org/10.18653/v1/2022.acl-long.512

Liu, Z., Gong, Z., Ai, L., Hui, Z., Chen, R., Leach, C. W., Greene, M. R., & Hirschberg, J. (2026). A review of incorporating psychological theories in LLMs. *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7459–7495. https://doi.org/10.18653/v1/2026.eacl-long.350

Lohndal, T. (2014). *Phrase structure and argument structure: A case study in the syntax-semantics interface*. Oxford University Press.

Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., & Gurevych, I. (2024). Are emergent abilities in large language models just in-context learning? *arXiv preprint arXiv:2309.01809*. https://arxiv.org/pdf/2309.01809

Mahowald, K. (2023). A discerning several thousand judgments: GPT-3 rates the article + adjective + numeral + noun construction. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 265–273. https://doi.org/10.18653/v1/2023.eacl-main.20

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540. https://doi.org/https://doi.org/10.1016/j.tics.2024.01.011

Manning, C. D., Clark, K., & Hewitt, J. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117

Marr, D. (1982). *Vision*. W.H. Freeman; Company.

McGee, T. A., Zhang, Y., & Blank, I. A. (2026). Evidence against syntactic encapsulation in large language models. *Cognitive Science*, *50*(3), e70187. https://doi.org/https://doi.org/10.1111/cogs.70187

Mickus, T. (2024). Language models and the paradigmatic axis. *Proceedings of the Society for Computation in Linguistics 2024*, 63–74.

Miller, P. H. (1999). *Strong generative capacity. the semantics of linguistic formalism*. CSLI publications.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Momennejad, I., Hasanbeig, H., Vieira Frujeri, F., Sharma, H., Jojic, N., Palangi, H., Ness, R., & Larson, J. (2023). Evaluating cognitive maps and planning in large language models with CogEval. *Advances in Neural Information Processing Systems*, *36*, 69736–69751.

Müller, S. (2025). Large language models: The best linguistic theory, a wrong linguistic theory, or no theory at all? *Zeitschrift für Sprachwissenschaft*, *44*(1). https://doi.org/https://doi.org/10.18148/zs/2025-2001

Müller-Eberstein, M., van der Goot, R., & Plank, B. (2022). Probing for labeled dependency trees. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7711–7726. https://doi.org/10.18653/v1/2022.acl-long.532

Murphy, E., Leivada, E., Dentella, V., Montero, R., Fritz Günther, F., & Marcus, G. (2025). Fundamental principles of linguistic structure are not represented by ChatGPT. *Biolinguistics*, *19*, e19021. https://doi.org/10.5964/bioling.19021

Niu, J., Lu, W., Corlett, E., & Penn, G. (2022). Using Roark-Hollingshead distance to probe BERT's syntactic competence. *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 325–334. https://doi.org/10.18653/v1/2022.blackboxnlp-1.27

Ott, D. (2017). Strong generative capacity and the empirical base of linguistic theory. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01617

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 27730–27744.

Parsons, T. (1990). *Events in the semantics of English*. MIT Press.

Piantadosi, S. T. (2024). Modern language models refute Chomsky's approach to language. In E. Gibson & M. Poliak (Eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett* (pp. 353–414). Language Science Press.

Pietroski, P. (2018). *Conjoining meanings: Semantics without truth-values*. Oxford University Press.

Qiu, Z., Duan, X., & Cai, Z. G. (2025). Grammaticality representation in ChatGPT as compared to linguists and laypeople. *Humanities and Social Sciences Communications*, *12*, 617. https://doi.org/doi.org/10.1057/s41599-025-04907-8

Quine, W. V. O. (1970). Methodological reflections on current linguistic theory. *Synthese*, *21*, 386–398.

Ramchand, G. (2008). *Verb meaning and the lexicon: A first phase syntax*. Cambridge University Press.

Ramchand, G. (2018). *Situations and syntactic structures: Rethinking auxiliaries and order in english*. The MIT Press.

Ramchand, G., & Svenonius, P. (2014). Deriving the functional hierarchy. *Language Sciences*, *46*(B), 152–174. https://doi.org/10.1016/j.langsci.2014.06.013

Rosenbaum, P. (1967). *The grammar of English predicate complement constructions*. The MIT Press.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. University of Chicago Press.

Seo, J., Moon, H., & Lim, H. (2025). The impact of negated text on hallucination with large language models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 13554–13572. https://doi.org/10.18653/v1/2025.emnlp-main.684

Silveira, N., Dozat, T., de Marneffe, M.-C., Samuel, Connor, M., Bauer, J., & Manning, C. D. (2014). A gold standard dependency corpus for English. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2897–2904.

Song, S., Hu, J., & Mahowald, K. (2025). Language models fail to introspect about their knowledge of language. *arXiv preprint arXiv:2503.07513*. https://arxiv.org/abs/2503.07513

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, *134*, 219–248. https://doi.org/https://doi.org/10.1016/j.lingua.2013.07.002

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601. https://doi.org/10.18653/v1/P19-1452

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tomasello, M. (2015). In E. Bavin & L. Naigles (Eds.), *The cambridge handbook of child language (2nd edition)* (pp. 69–87). Cambridge University Press.

Tuckute, G., Kanwisher, N., & Fedorenko, E. (2024). Language in brains, minds, and machines. *Annu Rev Neurosci*, *47*(1), 277–301. https://doi.org/10.1146/annurev-neuro-120623-101142

van der Burght, C. L., Friederici, A. D., Maran, M., Papitto, G., Pyatigorskaya, E., Schroën, J. A., Trettenbrein, P. C., & Zaccarella, E. (2023). Cleaning up the brickyard: How theory and methodology shape experiments in cognitive neuroscience of language. *Journal of Cognitive Neuroscience*, *35*(12), 2067–2088. https://doi.org/10.1162/jocn_a_02058

Van Valin, R. D. (2016). Functional linguistics. In M. Aronoff & J. Rees-Miller (Eds.), *The handbook of linguistics* (pp. 141–157). Wiley.

Varshney, N., Raj, S., Mishra, V., Chatterjee, A., Saeidi, A., Sarkar, R., & Baral, C. (2025). Investigating and addressing hallucinations of LLMs in tasks involving negation. *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, 580–598. https://doi.org/10.18653/v1/2025.trustnlp-main.37

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhins, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing*, 6000–6010.

Veenboer, T., & Bloem, J. (2023). Using collostructional analysis to evaluate BERT's representation of linguistic constructions. *Findings of the Association for Computational Linguistics: ACL 2023*, 12937–12951. https://doi.org/10.18653/v1/2023.findings-acl.819

Wagner, E., Alon, N., Barnby, J. M., & Abend, O. (2025). Mind your theory: Theory of mind goes deeper than reasoning. *Findings of the Association for Computational Linguistics: ACL 2025*, 26658–26668. https://doi.org/10.18653/v1/2025.findings-acl.1368

Wang, F., Zhang, Z., Zhang, X., Wu, Z., Mo, T., Lu, Q., Wang, W., Li, R., Xu, J., Tang, X., He, Q., Ma, Y., Huang, M., & Wang, S. (2025). A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, *16*(6). https://doi.org/10.1145/3768165

Weissweiler, L., Mahowald, K., & Goldberg, A. E. (2025). Linguistic generalizations are not rules: Impacts on evaluation of LMs. *Proceedings of the Second International Workshop on Construction Grammars and NLP*, 61–74.

White, J. C., Pimentel, T., Saphra, N., & Cotterell, R. (2021). A non-linear structural probe. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 132–138". https://doi.org/10.18653/v1/2021.naacl-main.12

Wiltschko, M. (2014). *The universal structure of categories: Towards a formal typology*. Cambridge University Press.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chenã, B., Wang, B., Kim, N., Andreas, J., & Kim, Y. (2024). Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1819–1862. https://doi.org/10.18653/v1/2024.naacl-long.102

Wu, Z., Chen, Y., Kao, B., & Liu, Q. (2020). Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4166–4176. https://doi.org/10.18653/v1/2020.acl-main.383

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wang, G., & Wu, F. (2026). Instruction tuning for large language models: A survey. *58*(7). https://doi.org/10.1145/3777411

Zhang, Y., Singh, S., Sengupta, S., Shalyminov, I., Su, H., Song, H., & Mansour, S. (2024). Can your model tell a negation from an implicature? Unravelling challenges with intent encoders. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 552–567. https://doi.org/10.18653/v1/2024.acl-long.33

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), 1–38. https://doi.org/10.1145/3639372