

Rethinking Linguistic Tests for Large Language Models: Airplanes Aren't Birds

Elena Callegari, University of Iceland

Annika Simonsen, University of Iceland

Abstract This article challenges the growing trend of using traditional linguistic tests to evaluate Large Language Models (LLMs), arguing that this approach fundamentally misunderstands both the purpose of these tests and the nature of LLMs. As a result, such evaluations often yield misleading conclusions, answering no real questions and failing to provide meaningful insights. At the core of our argument is the distinction between linguistic competence and performance: traditional linguistic tests were designed to probe human linguistic competence, yet LLMs do not acquire language through the same mechanisms and do not possess competence in the human sense. Using the analogy of birds and airplanes -both achieving flight through fundamentally different means- we illustrate why similar linguistic outputs do not imply shared underlying processes. This article calls for a change in how language-based tests are used in LLM evaluation: tests that operationalize performance can provide meaningful measures of LLM behavior, whereas tests designed to probe human linguistic competence should not be used to draw analogous competence-level conclusions about LLMs.

1 Introduction

Large Language Models (LLMs) have attracted considerable attention in recent years: LLMs' ability to generate human-like text has sparked both excitement and debate within the scientific community. LLMs demonstrate remarkable capabilities across a wide array of tasks such as text classification, machine translation, question answering, summarization and conversational interaction. Many current systems also operate across multiple modalities, processing combinations of text, images, audio and video. These models leverage Deep Neural Network (DNN) architectures inspired by the human brain's neural structure, consisting of interconnected nodes that process and transmit information as numerical signals. The dominant architecture underlying contemporary LLMs is the *transformer* architecture, which uses a mechanism called *attention* to process input sequences (Vaswani et al., 2017); this allows the model to weigh the importance of different words in a sentence when predicting the next word or understanding the context. Building on this architectural foundation, LLMs undergo an extensive training process using vast datasets sourced from the internet, often comprising trillions of tokens. Using this data, the model learns to predict sequences of words using probability.

While LLMs are now capable of generating text that is essentially indistinguishable from text generated by

humans, particularly in standard written English, it is important to note that LLMs differ fundamentally from human language speakers in several ways, including, but not limited to:

Input volume: LLMs are trained on enormous amounts of data. For example, GPT-3 (Brown et al., 2020) was trained on approximately 300 billion tokens, while GPT-4 has been estimated to have been trained on approximately 14 trillion tokens (Hickman et al., 2024). More recent openly documented models have reached still larger training scales: for example, Llama 4 Scout was pretrained on approximately 40 trillion tokens (Meta, 2025). Exact training-set sizes are not publicly disclosed for many contemporary proprietary models, including recent models in the GPT-5 series, but available information indicates that frontier models are routinely developed using datasets containing trillions of tokens. In contrast, human children are exposed to less than 100 million words by age 13 (Gilkerson et al., 2017). Although tokens and words are not directly equivalent units, the difference in scale is nevertheless several orders of magnitude. To give an intuitive sense of scale, the entire *The Lord of the Rings* trilogy contains approximately 480,000 words. A training corpus containing tens of trillions of tokens therefore corresponds, very roughly, to linguistic material on the order of tens of millions of copies of the trilogy.

Lack of grounded experience: unlike humans, LLMs

do not interact with the physical world, resulting in a lack of grounded experience and world knowledge (Hossain et al., 2019; Bisk et al., 2020).

Different forms of social learning: human language is acquired through sustained and reciprocal social interaction with caregivers and other speakers, which plays an important role in language development (Snow et al., 1976; Kuhl, 2007). Although contemporary LLMs may receive human feedback during training and post-training, this differs fundamentally from the situated and reciprocal social environment in which children acquire language.

Furthermore, while LLMs can generate grammatically and contextually appropriate language, their responses are based on patterns and probabilities derived from training data, without true awareness or intent. Therefore, despite their impressive performance, LLMs have limitations and potential issues:

- *Lack of true understanding:* while LLMs can generate coherent and contextually appropriate text, they do not possess human-like comprehension of the content they produce (Bender and Koller, 2020; Bender and Hanna, 2025).
- *Bias:* LLMs can reflect and amplify biases present in their training data (Mehrabian et al., 2021; Ferrara, 2023; Hofmann et al., 2024; Weissburg et al., 2025).
- *Hallucination:* LLMs can generate plausible-sounding but factually incorrect information (Ji et al., 2023; Xu et al., 2024; Kalai et al., 2025).

These limitations of LLMs have led to significant attention from researchers and linguists, who have felt compelled to explicitly articulate the boundaries of what these models can and cannot do. Their discussions have ranged from philosophical arguments about the nature of understanding, to practical concerns about the limitations of current AI approaches. Notably, Bender and Koller (Bender and Koller, 2020) argue that LLMs do not truly understand language, as they rely solely on form rather than meaning, lacking joint attention and inter-subjectivity, crucial to human language acquisition. Chomsky et al. (Chomsky et al., 2023) contend that systems like ChatGPT are merely “statistical engines for pattern matching”, incapable of engaging in true causal reasoning. Their critique goes beyond concerns about plagiarism or creativity, questioning the very foundation of current AI approaches in relation to human cognition and ethics. Lastly, the implications of these critiques extend to the field of theoretical linguistics itself. Recent studies suggest that LLMs struggle to acquire certain complex linguistic phenomena from input alone. Some researchers interpret this as evidence for innate linguistic knowledge, arguing that human language acquisition may rely on

pre-existing cognitive biases rather than exposure alone (Lan et al., 2024). However, others contend that these findings instead highlight the limitations of current AI architectures rather than proving the necessity of innateness. Moreover, LLMs fail to exhibit knowledge of fundamental linguistic principles that children easily acquire and lack the biases necessary to explain linguistic universals (Katzir, 2023). These critiques underscore the ongoing debate about the effectiveness and future of LLMs in understanding and generating human language, highlighting the importance of carefully considering how we evaluate and interpret LLM performance on linguistic tasks. This leads us to examine a recent trend within Natural Language Processing (NLP) which is to use traditional linguistic tests as a means of evaluating LLMs’ language capabilities.

In this article, we argue that the way traditional linguistic tests have been used in several recent papers to assess LLMs reflects a misunderstanding of both the nature of these tests and the distinction between linguistic *competence* and linguistic *performance*. While LLMs can produce language that appears human-like, their linguistic abilities are driven by statistical pattern recognition rather than rule-based grammatical knowledge. As such, traditional linguistic tests should not be used to assess LLMs’ linguistic *competence* in the same way they are used for humans. Performance-oriented tests can be used to evaluate LLM *performance*; however, competence-oriented tests cannot straightforwardly be used to test LLM *competence*, since the constructs those tests target in humans do not carry over to LLMs.

The central methodological problem can be stated simply: similarity in observable behavior does not entail similarity in the mechanisms that produce that behavior. This motivates the birds-and-airplanes analogy developed in this paper: birds and airplanes can both satisfy a behavioral criterion such as successful flight, yet a diagnostic designed to reveal the biological mechanisms underlying bird flight would not thereby become a valid diagnostic of the mechanisms underlying an airplane. The same distinction matters when linguistic tests developed to support inferences about human cognitive mechanisms are transferred to LLMs. Our objection is therefore not to comparison itself, but to carrying over the inferential interpretation of a test without establishing that the construct it was designed to measure transfers to the new system.

This article is structured as follows: in Section 2, we examine traditional linguistic tests, discussing their original purpose as tools for probing human linguistic competence. We outline several well-known examples, such as the Wug Test and garden path sentences, and highlight how these tests were designed to uncover implicit linguistic knowledge. In Section 3, we explore the

distinction between competence and performance, central to the Generative Grammar tradition, and demonstrate why this distinction is crucial when assessing LLMs. We use the analogy of birds and airplanes to illustrate how systems that achieve similar outcomes can rely on fundamentally different mechanisms. In Section 4, we examine both the appropriate and inappropriate ways to use language-based tests to assess LLMs, emphasizing that while testing competence in LLMs is possible, it must be done using the correct methods. In Section 5, we synthesize our argument into a set of practical recommendations for model-appropriate LLM evaluation. Section 6 is a concluding section.

2 Linguistic Tests

Linguistic tests are part and parcel of a linguist's skill set: they are, and have been, crucial in exploring different dimensions of language acquisition, processing and cognitive functions. These tests are typically designed to elicit specific linguistic responses, providing valuable insights into how language is comprehended, produced and learned. A particularly famous example of a linguistic test is the *Wug* Test; this is a seminal experiment in the study of morphological acquisition in children that was conducted by Jean Berko Gleason (Gleason, 1958). Its primary purpose is to determine whether children can apply grammatical rules to novel words and thus demonstrate morphological abilities that go beyond mere memorization. In this test, children are presented with pictures of imaginary creatures or objects and are introduced to them with a novel word, such as "This is a *wug*". They are then shown a picture of two of these creatures and prompted to apply a pluralization rule, for example, by asking, "Now there are two of them. There are two?". Most children will correctly respond with "*wugs*", showing their ability to pluralize a word they have never encountered before.

Other notable types of linguistic tests include the following:

In Syntax: *Garden Path Sentences* are used to study syntactic processing and how individuals interpret complex sentence structures (Bever, 1970; Christianson et al., 2006; Sanz et al., 2013). *Jabberwocky* sentences are used to investigate how the brain processes syntactic structures independent of semantic content (Moro et al., 2001; Röder et al., 2002; Yamada and Neville, 2007; Johnson and Goldberg, 2013).

In Pragmatics: *Scalar Implicature* tests examine how individuals (and especially children) derive meaning from context and use implicatures in communication (Smith, 1980; Reinhart, 2006; Noveck, 2001; Foppolo et al., 2012; Horowitz et al., 2018).

In Semantics: *Semantic Priming* tests (Schvaneveldt

et al., 1976) are used to study semantic processing and the influence of semantic information on cognitive tasks.

These kinds of traditional linguistic tests were not designed to test the explicit knowledge we have of language. Rather, they were (and are) designed and conceived as a way of probing the implicit rules and hidden cognitive procedures that underpin how we process and produce language. For instance, the *Wug* Test provides compelling evidence that language acquisition involves the internalization of abstract grammatical principles (e.g. "add the phoneme /s/, /z/, or /ɪz/ to the stem of a noun") rather than the simple memorization of word forms.

Many, if not all, of these tests also rely on specific assumptions concerning the structure, capacity and inner workings of the human language faculty. Take the famous "Colorless green ideas" sentence: this was designed by Noam Chomsky in 1957 to illustrate how a sentence can be grammatically correct while being semantically nonsensical (Chomsky, 1957). This demonstrates that grammatical rules can generate syntactically correct sentences even when they lack coherent meaning, which, according to Chomsky, highlights the distinction between syntax and semantics, with each being governed by separate cognitive modules.

In the past few years, we have seen numerous papers attempting to test the language capabilities of LLMs using traditional linguistic tests. For example, Weissweiler et al. (2023) conducted a study to examine the morphological capabilities of LLMs. Specifically, they tested ChatGPT's morphological capabilities using the *Wug* Test across four different languages: English, German, Tamil, and Turkish. Anh et al. (2024) also conducted a similar *Wug* Test study, but with several models, including GPT-4, and looking at Spanish, German, Vietnamese, Portuguese, French, and Romanian. Another example is Piantadosi (2024), who tested whether ChatGPT can produce sentences similar to Chomsky's "Colorless green ideas" sentence if asked. Other examples of papers using linguistic tests to test LLMs are Wilcox et al. (2023), who looked at English filler-gap dependencies and island constraints; Dentella et al. (2023), who looked at plural attraction, anaphora, center embedding, comparatives, intrusive resumption, NPI, the order of adjectives, the order of adverbs, and Mortensen et al. (2024), who look at conversion (or zero-derivation) in English.

3 LLMs: Airplanes, not Birds

We believe that it is essential to carefully distinguish between cases where it is appropriate to use language-based tests to evaluate LLMs, and areas where it is not. This distinction needs to be clearly articulated, as it is

currently insufficiently defined; in particular, one finds significant confusion in the literature, with many papers conflating notions and approaches that should not be conflated, leading to inconsistent and unclear assessments of LLMs.

A key distinction to consider when determining whether it is appropriate to test LLMs using linguistic methods is the difference between *competence* and *performance*. This distinction, most clearly articulated within the Generative Grammar tradition –originating with Chomsky’s work in the 1960s– refers to the difference between a speaker’s underlying linguistic knowledge (competence) and the actual use of language in real-time situations (performance). While these terms are rooted in generative theory, the underlying insight (that internal linguistic representations may differ from observable behavior) transcends any single theoretical framework. It applies broadly to linguistic inquiry, whether one adopts a formalist, usage-based, or cognitive perspective. Linguistic *competence* refers to the knowledge that a speaker has of their language, including an internalization of grammatical, syntactic, semantic and phonological rules that allow for the comprehension and production of an infinite number of sentences. According to Chomsky, competence is an abstract, mental representation of linguistic knowledge. It is what enables speakers to intuitively judge the grammaticality of sentences in their language even if they have never encountered that sentence before. Within the generative tradition, this account of linguistic competence is closely associated with Universal Grammar (UG), the hypothesis that the ability to acquire language is hard-wired into the brain, and that all human languages share a common underlying structure (Chomsky, 1965). In contrast, linguistic performance is concerned with the actual use of language in concrete situations. As such, performance also includes all the factors that come into play when people produce or comprehend utterances, such as memory limitations, attention, errors, hesitations, and other cognitive and social factors. Performance is variable and often imperfect, reflecting the complexities and constraints of real-world communication. According to Chomsky, performance does not provide a direct window into linguistic competence (Chomsky, 1965). The goal of linguistic theory is then to characterize the implicit knowledge that speakers have of their language, which can be obscured by the variability of performance.

Throughout this article, we use the term *linguistic test* to refer to traditional diagnostics developed in linguistics (most prominently in the generative grammar tradition, but not exclusively) whose primary purpose is to probe human linguistic competence. We use the broader term *language-based test* to refer to any evaluation that uses language as the medium of assess-

ment, including tests that target either performance or competence-like properties in artificial systems. As mentioned above, many of the tests developed within linguistics, such as tests based on acceptability judgments, novel-word generalization, or controlled minimal pairs, were developed as tools for probing human *competence*: the implicit grammatical knowledge posited to underlie speakers’ linguistic intuitions. Although these tests are behaviorally instantiated and can be influenced by performance factors, the data they elicit is typically interpreted as evidence about underlying representations rather than as measures of real-world communicative ability.

In this paper, we argue that testing LLMs using linguistic tests that were designed to test human linguistic *competence* represents a misapplication of methodologies: tests are being used that were designed for specific conditions and contexts on objects that lack the essential qualities for which these tests were originally intended. By applying these tests outside their proper context, the findings derived from such practice answer no real questions and fail to provide meaningful insights. As Guest and Martin (2023) point out, drawing inferences about shared competence based on shared behavioral outputs risks committing a metatheoretical fallacy: it conflates surface similarity with mechanistic equivalence. To avoid such errors, researchers must be explicit about what they hope to learn by applying language-based tests to LLMs, and whether the inferences they draw are licensed by the underlying assumptions of the test.

Even within “human” linguistics, the mere availability of a test does not always justify its application: testing must be appropriate to the context and the goals of one’s inquiry. For instance, one would not typically test whether a native speaker of Icelandic who is learning Danish as a second language can also pass the “colorless green ideas” test when processing Danish sentences. While this is technically possible, it would be pointless, as the test is designed to address a specific claim about the human language faculty, one that is not expected to vary across different languages or to impact competence across linguistic boundaries. Therefore, discovering that the speaker can also separate syntax from semantics in Danish would offer little value. The point is that testing, even when technically feasible, must still align with meaningful goals.

3.1 How Birds Migrate and Airplanes Land on Time

To fully appreciate the strengths and weaknesses of an argument, it can be valuable to apply it to a completely different context or domain. We will do that by comparing the differences between natural language vs. LLM-

generated language to the differences between birds and airplanes.

Many bird species undertake long-distance migrations between breeding and wintering grounds. To migrate, birds use a combination of genetic programming, environmental cues (changes in day length, position of the sun and stars, earth's magnetic field, visual landmarks), and an internal biological clock to time their migrations (Gupta et al., 2017; Wiltchko and Wiltchko, 2023; Kappeler, 2022; Newton, 2023). This combination of genetic endowment and environmental information allows birds to fly to the correct part of the world, at the correct time of the year. Similarly, through the sophisticated integration of navigation technology, airline scheduling and efficient air traffic control systems, modern airplanes typically land punctually and at the intended destination. While the functional outcome (flight) is the same, birds and airplanes achieve it through fundamentally different structures and mechanisms. Birds generate and control aerodynamic forces using flexible, actively deformable wings powered by muscles. Airplanes, by contrast, generate lift using comparatively rigid wings and obtain thrust from engines, with their motion controlled through aerodynamic control surfaces and mechanical systems. And the navigation skills of birds, which rely on innate instincts, astronomical cues and Earth's magnetic field, have little in common with the engineered precision navigation achieved in the modern airline industry: airplanes use inertial navigation systems that employ accelerometers and gyroscopes to track the aircraft's position, velocity and heading; GPS to provide precise position information; radio-navigation systems such as the Instrument Landing System (ILS) to guide approaches when required. Finally, airplanes also benefit from surveillance technologies such as radar and ADS-B, together with air traffic control, to coordinate safe and orderly arrivals.

Birds' ability to exhibit such complex navigation skills, traversing hundreds of kilometers and landing in the right location at the same time each year, has attracted the attention of numerous researchers. These scientists have developed various tests to better understand the genetic underpinnings of this remarkable behavior. This is remarkably similar to how linguists study human language, delving into the genetic and cognitive foundations that enable such intricate and precise abilities. Consider birds' internal timing mechanisms. Birds have an internal circadian clock that helps regulate daily activities, including migratory behavior. This clock is influenced by environmental cues like light and darkness, but maintains its rhythm even in altered ecological conditions; for example, birds kept under consistent 12-hour or 18-hour light cycles continue to show seasonal migratory restlessness (a phe-

nomenon called *Zugunruhe*), indicating the presence of an internal circannual rhythm that regulates migratory activity independent of external cues (Gwinner, 1972, 2012). Applying the same tests used to study birds' genetic ability to migrate to evaluate airplanes' ability to navigate the globe would be absurd: if we were to test whether airplanes can still depart on time even if kept under constant 18-hour light cycles, the answer would be an obvious "yes, they can", providing exactly zero value. This experiment would neither better explain the mechanisms of bird migration nor enhance our understanding of how airplanes operate. This is an example of a competence test designed for one domain being applied to assess competence in another domain, with nonsensical end results because the two systems operate on fundamentally different principles.

This does not mean that comparing objects with different architectures but identical outcomes is necessarily scientifically meaningless. On the contrary, such comparisons can be insightful, highlighting how two distinct systems achieve the same outcome through different mechanisms. For instance, studying birds has inspired innovations in aviation technology (a phenomenon known as *biomimicry*), leading to new ideas for more efficient, flexible, and sustainable flight technologies (Hoidis et al., 2017). One well-documented example is the development of winglets: the small, up-turned extensions at the ends of airplane wings. These were inspired by how birds curl their wing feathers upward to reduce drag and improve aerodynamic efficiency during flight. By applying this principle to aircraft design, engineers significantly reduced vortex drag, leading to improved fuel efficiency and enhanced stability in modern airplanes (Smith, 1975; NASA, 2005). Today, winglets are a standard feature on commercial aircraft.

While it would be nonsensical to use a competence test designed for birds to assess an airplane's capabilities, applying *performance* tests designed for one system to the other can be illuminating. Consider a performance test designed to evaluate how efficiently birds fly over long distances, such as migratory birds traveling thousands of kilometers using minimal energy. Evaluating airplanes using a comparable test, such as measuring their fuel efficiency over long-distance flights, can provide valuable insights. In this scenario, the test focuses on energy consumption during sustained flight. Migratory birds rely on highly efficient use of energy, specialized wing shapes, and flight formations like V-shaped patterns to reduce energy expenditure. Airplanes, on the other hand, consume vast amounts of fuel to achieve similar distances. While airplanes can certainly travel longer distances more quickly, birds are far more energy-efficient. This performance comparison can be valuable because it high-

lights where technology has yet to match nature's efficiency and performance.

4 The Right Way and the Wrong Way

Since LLM training and human language acquisition rely on fundamentally different mechanisms, language-based tests should be used for LLM evaluation only insofar as they operationalize *performance*: the model's ability to generate well-formed and context-appropriate text under specified conditions. Competence-oriented linguistic tests, as traditionally understood in theoretical linguistics, presuppose a theory of human competence, and applying them to LLMs as if they tested the same construct amounts to a category error. A model's success on such tests does not warrant inferences about human-like implicit grammatical knowledge. When the goal is to probe something closer to model-internal competence (the representational and mechanistic organization that supports model behavior), one must use methods that are appropriate to LLMs, rather than relying on human competence diagnostics alone.

For performance-oriented evaluation, second-language proficiency frameworks provide one natural source of language-based assessments because they define success in terms of observable language use under specified conditions, without presupposing any particular underlying linguistic mechanism. For example, one can ask whether a model's written production in Faroese meets CEFR-aligned B1/B2 descriptors across controlled tasks, or whether its performance reaches the B2 standard on relevant components of Norskprøven, Norway's official Norwegian-language proficiency test. These types of tests assess performance rather than competence because they focus on the model's ability to generate appropriate, coherent and grammatically correct responses in specific, practical scenarios. In other words, these tests measure how well the LLM can perform a task, producing language that fits certain criteria, like fluency and accuracy, without necessarily probing which internal representation of language rules the model has.

Examples of papers that use language-based tests on LLMs in a more principled way include Tan et al. (2024), who introduce DevBench, a multimodal benchmark consisting of seven evaluation tasks spanning lexical, syntactic and semantic domains, with behavioral data from both children and adults. Tan et al. compare vision-language models to human behavioral patterns on paradigms drawn from experimental psycholinguistics. For instance, one task presents an image of a ball and an image of an apple side by side, followed by the

word *ball*, and measures how long the subject looks at each image. We view this as a more appropriate and informative use of linguistic methodology for LLM evaluation: the goal is not to infer human-like linguistic competence from test outcomes, but to compare behavioral patterns under a shared experimental protocol. In this sense, the benchmark is used as a performance-oriented measure of how models behave in controlled settings, rather than as a competence diagnostic in the traditional linguistic sense. Another relevant example is Hu and Levy (2023), who compare two ways of evaluating linguistic generalizations in LLMs: asking the model to provide an explicit metalinguistic judgment through prompting, and directly measuring the probabilities it assigns to alternative words or sentences. Their prompts include questions such as which word is most likely to complete a sentence, whether one continuation is more semantically plausible than another, and which of two sentences is preferable in a syntactic comparison. They find that direct probability measures are generally more reliable than prompted judgments. Their results show that a model's response to a human-style linguistic question need not provide the best measure of its linguistic behavior; measures tied more closely to the model's own next-token prediction mechanism can be more informative. Ultimately, neither paper aims to use the linguistic tests to see if LLMs have become models of human language, but rather apply the tests in a way that is meant to improve the development of LLMs.

While we highlight certain papers as moving in the right direction, we do not necessarily endorse all aspects of their methodologies.

4.1 Testing LLMs' Competence

While LLMs do not possess linguistic competence in the human sense, they do exhibit systematic patterns in how they generate language. These patterns can be studied *as a form of* machine competence. However, this requires evaluation methods specifically designed for LLMs, rather than the application of linguistic tests originally developed for human cognition. Researchers should aim to design LLM-specific tests that reveal how models internally represent linguistic knowledge without assuming that these representations correspond to human language competence. One such example are tests designed to shed light on the inner workings of LLMs (often referred to as demystifying the LLM "black box").

One of the significant challenges with LLMs is the opacity of their internal processes, commonly referred to as the "black box" problem. This term highlights the difficulty in understanding how these models arrive at their outputs, because the high-dimensional and distributed computations that drive their behavior are dif-

difficult to interpret, even when model parameters and activations are technically accessible. This lack of transparency poses challenges for researchers and practitioners trying to interpret, explain, or control the behavior of LLMs (Luo and Specia, 2024; Zhao et al., 2024; Palikhe et al., 2025). To address these challenges, researchers have developed methods for formulating and testing hypotheses about the internal representations and computations that support LLM behavior.

Probing methods test whether information about a specified property or task-relevant variable can be decoded from a model's hidden representations, often at different layers of the model (Belinkov, 2022; Pimentel et al., 2020). Such results can establish that information about a property is available in a representation, but they do not by themselves show that the model uses that information in generating the relevant behavior. This distinction is crucial when the aim is to characterize a model's machine competence rather than merely to describe correlations in its internal states. Causal-intervention methods address this limitation by altering internal activations or model components and measuring the resulting change in behavior. Elazar et al. (2021), for example, remove information identified by probing from a model's representations in order to test whether it is functionally relevant to model behavior. More generally, activation-patching approaches make it possible to test counterfactual hypotheses about the contribution of particular internal states to a model's outputs (Ghandeharioun et al., 2024). These methods provide stronger evidence than decodability alone when the question is whether a representation or component plays a functional role in producing a particular behavior.

Mechanistic interpretability provides a related approach by seeking to identify circuits: sets of attention heads, MLPs, features and connections that jointly implement a task-relevant computation. Wang et al. (2023), for example, identify and causally validate a circuit supporting indirect-object identification in GPT-2 Small. Extending circuit analysis across checkpoints and model sizes, Tigges et al. (2024) find substantial consistency in the functional organization of circuits associated with particular tasks, even where the individual components instantiating those functions differ. Together, probing, causal intervention and circuit analysis offer model-appropriate ways of studying the representational and computational mechanisms that support LLM linguistic behavior, without presupposing that those mechanisms correspond to human linguistic representations.

Studies attempting to uncover the internal workings of LLMs can be considered analogous to linguistic tests used to probe human language competence; in both cases, the objective is to uncover the under-

lying mechanisms that produce the observed behavior. For human language, linguistic tests like grammaticality judgments or the Wug test probe a person's competence: the abstract knowledge they have of linguistic rules, independent of real-time performance constraints. Similarly, in LLMs, interpretability methods can be used to investigate what may be called machine competence: the internal representations and computations that support a model's behavior.

Exactly like resorting to traditional linguistic tests to assess a language model's *competence* represents a fundamentally misguided approach (as we have argued), applying methodologies developed to understand the inner workings of LLMs to humans would be equally nonsensical. This is because methods designed to probe the workings of these statistical models are built on assumptions about how these models function, and these assumptions simply do not translate to the biological, cognitive processes underlying human language understanding. For example, input-attribution techniques, such as Integrated Gradients, attempt to trace a model's prediction back to specific input tokens, identifying their contribution to a particular output (Sundararajan et al., 2017; Sanyal and Ren, 2021). Such methods are meaningful for LLMs because their inputs are explicitly tokenized and their computations can be analyzed mathematically. Human language processing also involves probabilistic prediction (Kuperberg and Jaeger, 2016), but applying Integrated Gradients directly to a human language user would be nonsensical: the biological and multimodal processes supporting human language processing are not accessible as a differentiable computational system to which gradient-based analysis can be applied. The method is meaningful for LLMs precisely because it is tailored to the kind of system an LLM is.

4.2 Using LLMs as Theories of the Human Language Capacity

There are thus both appropriate and inappropriate ways to use language-based tests to test LLMs, and the key distinction lies in whether these tests are used to evaluate the model's performance or its competence. It is crucial to clearly differentiate between the two uses, as many papers tend to blur the lines, conflating concepts that belong to both domains. This confusion is one of the primary reasons we felt the need to open a discussion on this matter and write this article, hoping to clarify these distinctions. For example, consider Dentella et al. (2023). Their paper evaluates the language abilities of three large language models (GPT-3/text-davinci-002, GPT-3/text-davinci-003, and GPT-3.5) using systematic grammaticality judgment tasks. The study focuses on determining whether these models

can accurately and consistently recognize grammatical versus ungrammatical sentences, as well as whether their performance improves with repetition. However, the authors consistently blur the distinction between performance and competence. For instance, in one part of the paper, they state that “the LM inability to recognize [grammatical prompts as such] is a sign of LMs lacking *competence* over structural relations” (our italics), but just a few lines later, they refer to their results as “an index of a *performance* mismatch between humans and LMs” (our italics).

The authors also state that their work “aims to inform the conversation on whether LMs can act as theories of natural language”. However, this claim seems inconsistent with the type of experiments they carry out: grammaticality judgment tests alone are insufficient to determine whether a system like an LLM can serve as a theory of natural language. While such tests can assess a model’s ability to produce or recognize grammatically correct sentences, they do not provide insight into whether the model possesses an underlying language competence comparable to that of humans. In terms of our analogy, the fact that an airplane can successfully navigate to a destination does not establish that its navigation system constitutes a theory of bird navigation: GPS, inertial navigation and air-traffic-control infrastructure may produce comparable outcomes without modeling the mechanisms by which birds navigate. A comprehensive theory of natural language must account for deeper cognitive processes and the mechanisms that drive language acquisition and use, which go beyond surface-level grammaticality judgments.

The article concludes with the statement “Such results [the authors’ results] are ultimately incompatible with the view that the tested LMs can serve as cognitive models of the human language capacity at their current stage of development”. The way the term “capacity” is used here is problematic. If it is used to refer to literal processing ability, the comparison is flawed since humans cannot process trillions of tokens the way LLMs can, rendering the authors’ claim trivial. Conversely, if “capacity” is intended to mean “human language competence”, the comparison remains inaccurate, as discussed earlier in this article.

Consider also this passage from that same article: “If LMs truly learn language, this should in principle entail the ability to learn and apply rules of grammar consistently and reliably”. We know, however, that LLMs do not learn language in the same way humans do, so their conditional proposition is based on a faulty assumption. Moreover, their claim that “human-like formal competence in LMs presupposes that these models have succeeded in extracting rules of natural language grammar from their training data” is misleading. Once again, similarity in outcome does not imply

similarity in the underlying processes that bring about those outcomes, as our birds versus airplanes analogy has demonstrated.

Similarly, we do not endorse the notion of using LLMs as mental models for human language acquisition or cognition, as the architectures and underlying mechanisms of human and machine language processing are inherently distinct (see also [Dupre \(2021\)](#)). Suggesting that the success of LLMs completely refutes generative theories of grammar, as Piantadosi does in his paper ([Piantadosi, 2024](#)), is misguided. The two systems are simply not comparable enough. The fact that both LLMs and humans can produce similarly high-quality language does not imply that they rely on similar mechanisms. They do not, as we have discussed. Piantadosi notes in particular that “they [LLMs] are the only models in existence that do a good job of capturing the basic dynamics of human language”. However, this statement is problematic. By this logic, a well-trained parrot’s ability to mimic human speech could also be considered a reasonable model of human language. If the goal is to understand the process behind language, and it is evident that two systems achieve similar results through different methods, studying one as a model for the other would be misguided.

A natural response to this critique could be to ask whether LLMs might become more informative as models of human language acquisition if they are trained under more human-like constraints. Work in this direction includes [Huebner et al. \(2021\)](#), who introduced BabyBERTa by fine-tuning a masked language model (RoBERTa-base) on only 5 million tokens of child-directed speech, intended to approximate the input children receive between ages 1 and 6. In a similar spirit, the CoNLL/CMCL BabyLM Challenge ([Warstadt et al., 2023](#)) was launched in 2023, constraining participants to train language models on only 100 million words before evaluating them on acceptability and related benchmarks. We view these efforts as valuable primarily as contributions to resource-efficient language modeling and controlled comparative evaluation. Restricting data and compute can yield practical insights into which training regimes improve generalization under constraints, and it can help disentangle the effects of scale from the effects of objective and data composition. However, we do not believe that training LMs on smaller datasets (even datasets that resemble child-directed speech) turns them into faithful models of human language acquisition. The learning setup and representational format still remain fundamentally different from those of human learners. For this reason, such models may be useful experimental artifacts for improving LLM training, but they should not be treated as theories of the human language faculty.

As a final example, let us consider Weissweiler et

al.'s (Weissweiler et al., 2023) paper, where the Wug test is applied to ChatGPT. The paper opens with the following question: "Do large language models (LLMs) possess human-like linguistic capabilities?". This initial framing seems to suggest that the authors are making use of linguistic tests for what could be considered the "right" goal: seeking to assess LLM *performance* and measure how it diverges from human linguistic abilities. As the reader delves deeper into their paper, however, it becomes clear that the basic rationale behind the tests carried out in Weissweiler et al. (Weissweiler et al., 2023) is in fact to evaluate whether LLMs have internalized generalizable morphological rules similar to those of human speakers. Has ChatGPT also arrived at a rule like "Add the suffix '-s', '-es', or '-ies' to the stem of a noun" to pluralize English nouns, similarly to the way English-speaking children have? The authors are thus testing LLMs' competence using human-designed competence tests. We are now back to the methodological fallacy of testing airplanes like we would be testing birds. From the fact that an airplane produces thrust, we would not infer that it does so by flapping its wings simply because this is how birds generate thrust during powered flight. Likewise, from the fact that an LLM successfully pluralizes a novel noun, we cannot infer from that behavioral result alone that it implements the same morphological rule or representation as a human speaker.

Not surprisingly, the authors' response to the question of whether LLMs have internalized that same morphological rules as humans have is negative. In particular, they write: "We think ChatGPT is not applying morphological rules to a stem, which would be in line with item-and-process accounts of morphology (Hockett, 1954). Rather, it seems to linguistically decode the point in its representational space defined by the semantic constraints in the prompt. In cases where this point (and its immediate neighborhood) is unoccupied, it generates a form based on the nonce word, but in cases where there is a form of a real word close to the point (e.g., because of superficial orthographic similarity), it generates this form instead".

5 Towards System-Appropriate Evaluation: Practical Guidelines for Testing LLMs

If LLMs exhibit a form of statistical competence distinct from human linguistic competence, then it follows that the methods we use to evaluate them must be aligned with that distinction. Evaluating LLMs as if they are flawed or incomplete versions of human minds risks conflating functional similarity with mechanistic equivalence. To avoid such missteps, we advocate for

a model-appropriate approach to LLM evaluation, one that treats these systems on their own terms and avoids the temptation to overinterpret surface-level behavioral overlap. Below we outline four core guidelines to support more rigorous and insightful evaluations of LLMs.

5.1 Be Explicit About the Target of Evaluation

Researchers should state explicitly what an evaluation is designed to test. Is the goal to measure *performance* on a task (i.e., what outputs are being produced under a given prompt and scoring procedure), or to probe something closer to *competence* (i.e., the internal organization and representations that support those outputs)? Without making this distinction explicit, evaluations often drift: a study is set up as a behavioral benchmark, but the results are then interpreted as evidence about competence, leading to conclusions the methodology could not possibly warrant.

Researchers should then keep the competence/performance distinction in view throughout. Performance tests measure how a system functions under specific conditions, focusing on immediate outcomes. Competence refers to the inherent capabilities that consistently enable a system to achieve its goals.

We therefore recommend that authors (i) state early on which of these targets they are evaluating and why their method is appropriate for that target, and (ii) keep terminology consistent with that choice. For example, do not describe a behavioral benchmark as evidence of "competence", and do not draw competence-level conclusions (e.g., about rules, constraints, or human-like grammatical knowledge) from performance-level results unless the paper provides an explicit justification for that inference.

5.2 Avoid Category Imposition Without Justification

Many LLM probing studies rely on human-defined linguistic categories such as part-of-speech tags, syntactic constituents or agreement (Starace et al., 2023; Finlayson et al., 2021). These categories originate from theoretical models of human grammar, and there is no guarantee that they map cleanly onto the distributed representations that LLMs develop through self-supervised training.

We fully realize that it might be impossible to avoid human linguistic categories entirely, since they are the primary tools we use to describe and evaluate language. However, researchers should be careful not to treat these categories as if they were automatically "real" inside the model. In many cases, it may be preferable to complement category-based probing with more

model-native or less theory-loaded analyses, for example, unsupervised clustering of contextual embeddings, geometry-based analyses of representation spaces¹, or other methods that do not presuppose the relevant linguistic distinctions in advance.

If human linguistic categories *are* imposed on LLM internals, researchers should be explicit about which categories are being assumed, and should justify why those categories are expected to be relevant to the model's representational organization. The burden of proof here lies with the analyst: if human categories (POS, constituents, case, etc.) are imposed on model internals, one risks getting structure because the method forces it, not because the model genuinely organizes information that way.

In short, researchers should not assume that categories transfer from humans to LLMs simply because they are useful in human linguistics. Category-based analysis can be valuable, but it must be treated as an interpretive choice that requires justification, not as a default mapping.

5.3 Avoid Unqualified Claims About “Modeling Language”

LLMs are often described as “models of human language”. We believe this to be a dangerously underspecified claim that easily invites stronger interpretations than the evidence supports. It easily encourages a slide from the fact that a model produces fluent, human-like text to the conclusion that it is therefore a model of human language in a cognitively meaningful sense.

As [Guest and Martin \(2023\)](#) warn, inferences from behavioral similarity to shared underlying mechanisms often rest on a classic logical error (affirming the consequent): if a system exhibits behavior similar to humans, it is incorrectly inferred to possess the same underlying competence. Unqualified claims that LLMs “model human language” invite exactly this kind of inference, perhaps even when it is not the author's intention.

We therefore recommend that researchers avoid blanket statements of the form “LLMs are models of language” unless they are accompanied by a clear qualification of what is meant and, crucially, what is *not* being claimed. In most cases, it is more accurate to describe LLMs as systems that approximate certain observable regularities of linguistic behavior, without implying that they model the mechanisms that underlie human linguistic competence, acquisition or cognition.

¹Such methods still involve methodological and interpretive choices, but they can reduce reliance on a single predefined linguistic category.

5.4 Favor Comparative Over Absolute Metrics

It is often more informative to ask how changes in model architecture, training data, or scale affect performance on a linguistic task than to ask whether a model “has” a particular linguistic capacity. Many current evaluations are framed as binary questions (e.g., “Can the model pluralize novel words?”; “Does the model know island constraints?”). This framing also more easily invites competence-style conclusions from what are, in most cases, performance-level measurements.

The same distinction is familiar in engineering. Knowing that an airplane flies tells us relatively little about the factors responsible for its performance; to understand the system, engineers vary or compare features such as wing design, propulsion and control systems, and examine how these changes affect outcomes such as efficiency or maneuverability. The analogous questions for LLMs concern how linguistic behavior changes as properties of the model or its training regime change. Rather than asking whether an LLM “has” morphological competence, for example, it is more informative to ask how morphological generalization changes when training data is limited to human-scale quantities, as in the BabyLM Challenge ([Warstadt et al., 2023](#)). These comparative questions, whether larger models better handle island constraints, whether instruction tuning improves reflexive interpretation, yield actionable insights that a simple pass/fail framing cannot provide, and they avoid the trap of drawing categorical conclusions about language knowledge from one-off behavioral probes.

In sum, we call for a shift in how language-based evaluations of LLMs are conceptualized and interpreted. Rather than treating LLMs as partial replications of human minds, they should be evaluated as distinct systems whose capabilities emerge from fundamentally different principles. This shift entails a move away from cognitive theory-testing toward the behavioral analysis of artificial systems, and, importantly, toward an explicit metatheoretical awareness of the inferences we license when we test these models.

6 Conclusion

In this article, we have explored the limitations of applying traditional linguistic tests, such as the Wug test, to Large Language Models (LLMs). While these tests have proven invaluable in assessing human linguistic competence, their direct application to LLMs is problematic. LLMs, trained on vast amounts of data, operate on statistical patterns without the type of human linguistic competence that such tests are designed to measure. This point is captured by the comparison be-

tween airplanes and birds. Both can fly, but evidence that an airplane reaches its destination does not show that it navigates through the biological mechanisms that support bird migration. Likewise, an LLM's successful pluralization of a novel word or sensitivity to a syntactic contrast does not establish that it has acquired a human-like morphological rule or grammatical representation.

The critical distinction between competence and performance highlights why traditional linguistic tests fail to capture the true nature of LLMs' capabilities. LLMs should not be evaluated as if they possess human-like linguistic competence but rather on their ability to perform tasks relevant to real-world language use. Going forward, we call for researchers working on assessing LLMs' capabilities to clearly define what dimension of inquiry they are testing, and to avoid conflating or confusing performance with competence. Evaluations should be aligned with the specific architecture and purpose of LLMs, ensuring that the tests used are suitable for assessing either performance or competence. While it is possible to test competence, this must be done using methods designed specifically for LLMs, rather than relying on tests intended for human linguistic competence.

References

- Anh, Dang, Limor Raviv, and Lukas Galke. 2024. Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.
- Belinkov, Yonatan. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.
- Bender, E.M. and A. Hanna. 2025. *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. HarperCollins.
- Bender, Emily M and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.
- Bever, TG. 1970. The cognitive basis for linguistic structures. *Cognition and the Development of language/Wiley*.
- Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Chomsky, Noam. 1957. *Syntactic Structures*, volume 4 of *Janua Linguarum, Series Minor*. Mouton, The Hague.
- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Massachusetts: The MIT Press, Cambridge.
- Chomsky, Noam, Ian Roberts, and Jeffrey Watumull. 2023. The false promise of ChatGPT. *The New York Times*. Accessed on July 12, 2024.
- Christianson, Kiel, Carrick C Williams, Rose T Zacks, and Fernanda Ferreira. 2006. Younger and older adults' "good-enough" interpretations of garden-path sentences. *Discourse processes*, 42(2):205–238.
- Dentella, Vittoria, Fritz Günther, and Evelina Leivada. 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*, 120(51):e2309583120.
- Dupre, Gabe. 2021. (What) can deep learning contribute to theoretical linguistics? *Minds and Machines*, 31(4):617–635.
- Elazar, Yanai, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.
- Ferrara, Emilio. 2023. Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*, 28(11).
- Finlayson, Matthew, Aaron Mueller, Sebastian Gehrmann, Stuart M Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843.

- Foppolo, Francesca, Maria Teresa Guasti, and Gennaro Chierchia. 2012. Scalar implicatures in child language: Give children a chance. *Language learning and development*, 8(4):365–394.
- Ghandeharioun, Asma, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: a unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Gilkerson, Jill, Jeffrey A Richards, Steven F Warren, Judith K Montgomery, Charles R Greenwood, D Kimbrough Oller, John HL Hansen, and Terrance D Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, 26(2):248–265.
- Gleason, Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.
- Guest, Olivia and Andrea E Martin. 2023. On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, 6(2):213–227.
- Gupta, Neelu Jain, Vatsala Dwivedi, Bhanu P Singh, and Sanjay K Bhardwaj. 2017. Orientation in migrating animals: Role of biological clocks. *Biological Timekeeping: Clocks, Rhythms and Behaviour*, pages 643–658.
- Gwinner, Eberhard. 2012. *Circannual rhythms: endogenous annual clocks in the organization of seasonal processes*, volume 18. Springer Science & Business Media.
- Gwinner, Eberhard G. 1972. Endogenous timing factors in bird migration. *NASA, Washington Animal Orientation and Navigation*.
- Hickman, Louis, Patrick D Dunlop, and Jasper Leo Wolf. 2024. The performance of large language models on quantitative and verbal ability tests: Initial evidence and implications for unproctored high-stakes testing. *International Journal of Selection and Assessment*, 32(4):499–511.
- Hofmann, Valentin, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts AI decisions about people's character, employability, and criminality.
- Hoidis, Jana, Steffen Opitz, and Srinivas Vasista. 2017. Of birds and aircrafts; how biological examples are optimising technology in aviation. *DLR magazine*, (153):20–23.
- Horowitz, Alexandra C, Rose M Schneider, and Michael C Frank. 2018. The trouble with quantifiers: Exploring children's deficits in scalar implicature. *Child development*, 89(6):e572–e593.
- Hossain, MD Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36.
- Hu, Jennifer and Roger P. Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Huebner, Philip A, Elinor Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning*, pages 624–646.
- Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Johnson, Matt A and Adele E Goldberg. 2013. Evidence for automatic accessing of constructional meaning: Jabberwocky sentences prime associated verbs. *Language and Cognitive Processes*, 28(10):1439–1452.
- Kalai, Adam Tauman, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate.
- Kappeler, Peter M. 2022. Orientation in time and space. In *Animal Behaviour: An Evolutionary Perspective*, pages 69–91. Springer.
- Katzir, Roni. 2023. Why large language models are poor theories of human linguistic cognition. a reply to piantadosi (2023). *Manuscript. Tel Aviv University*. url: <https://lingbuzz.net/lingbuzz/007190>.
- Kuhl, Patricia K. 2007. Is speech learning 'gated' by the social brain? *Developmental science*, 10(1):110–120.
- Kuperberg, Gina R and T Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1):32–59.
- Lan, Nur, Emmanuel Chemla, and Roni Katzir. 2024. Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*, pages 1–56.
- Luo, Haoyan and Lucia Specia. 2024. From understanding to utilization: A survey on explainability for large language models. *arXiv preprint arXiv:2401.12874*.

- Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.
- Meta. 2025. Llama 4 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md. Accessed 2026-08-11.
- Moro, Andrea, Marco Tettamanti, Daniela Perani, Caterina Donati, Stefano F Cappa, and Ferruccio Fazio. 2001. Syntax and the brain: disentangling grammar by selective anomalies. *Neuroimage*, 13(1):110–118.
- Mortensen, David R., Valentina Izrailevitch, Yunze Xiao, Hinrich Schütze, and Leonie Weissweiler. 2024. Verbing weirds language (models): Evaluation of English zero-derivation in five LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17359–17364, Torino, Italia. ELRA and ICCL.
- NASA. 2005. Winglet technology: Lessons from nature applied to aviation. Technical report, NASA Technical Reports.
- Newton, Ian. 2023. *The migration ecology of birds*. Elsevier.
- Noveck, Ira A. 2001. When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2):165–188.
- Palikhe, Avash, Zhenyu Yu, Zichong Wang, and Wenbin Zhang. 2025. Towards transparent AI: A survey on explainable large language models. *arXiv preprint arXiv:2506.21812*.
- Piantadosi, Steven T. 2024. Modern language models refute Chomsky’s approach to language. In Edward Gibson and Moshe Poliak, editors, *From fieldwork to linguistic theory: A tribute to Dan Everett*, pages 353–414. Language Science Press.
- Pimentel, Tiago, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.
- Reinhart, Tanya. 2006. Interface strategies. Optimal and costly derivations.
- Röder, Brigitte, Oliver Stock, Helen Neville, Siegfried Bien, and Frank Rösler. 2002. Brain activation modulated by the comprehension of normal and pseudo-word sentences of different processing demands: a functional magnetic resonance imaging study. *Neuroimage*, 15(4):1003–1014.
- Sanyal, Soumya and Xiang Ren. 2021. Discretized integrated gradients for explaining language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10285–10299.
- Sanz, Montserrat, Itziar Laka, and Michael K Tanenhaus. 2013. *Language down the garden path: The cognitive and biological basis for linguistic structures*. OUP Oxford.
- Schvaneveldt, Roger W, David E Meyer, and Curtis A Becker. 1976. Lexical ambiguity, semantic context, and visual word recognition. *Journal of experimental psychology: human perception and performance*, 2(2):243.
- Smith, Carol L. 1980. Quantifiers and question answering in young children. *Journal of experimental child psychology*, 30(2):191–205.
- Smith, H. 1975. Aeroelasticity and vortex drag reduction using winglets. *Journal of Aircraft*, 12(8):501–507.
- Snow, Catherine E, Anjo Arlman-Rupp, Yvonne Hasing, Jan Jobse, Jan Joosten, and Jan Vorster. 1976. Mothers’ speech in three social classes. *Journal of Psycholinguistic Research*, 5:1–20.
- Starace, Giulio, Konstantinos Papakostas, Rochelle Choenni, Apostolos Panagiotopoulos, Matteo Rosati, Alina Leiding, and Ekaterina Shutova. 2023. Probing LLMs for joint encoding of linguistic categories. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7158–7179.
- Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.
- Tan, Alvin W. M., Sunny Yu, Bria Long, Wanjing Anya, Tonya Murray, Rebecca D. Silverman, Jason D. Yeatman, and Michael C. Frank. 2024. Devbench: A multi-modal developmental benchmark for language learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 77445–77467. Curran Associates, Inc.
- Tigges, Curt, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. LLM circuit analyses are consistent

- across training and scale. In *Advances in Neural Information Processing Systems*, volume 37, pages 40699–40731. Curran Associates, Inc.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wang, Kevin Ro, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*.
- Warstadt, Alex, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors. 2023. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Singapore.
- Weissburg, Iain, Sathvika Anand, Sharon Levy, and Haewon Jeong. 2025. LLMs are biased teachers: Evaluating LLM bias in personalized education. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5650–5698, Albuquerque, New Mexico. Association for Computational Linguistics.
- Weissweiler, Leonie, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Hao-fei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.
- Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2023. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44.
- Wiltschko, Roswitha and Wolfgang Wiltschko. 2023. Animal navigation: how animals use environmental factors to find their way. *The European Physical Journal Special Topics*, 232(2):237–252.
- Xu, Ziwei, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.
- Yamada, Yoshiko and Helen J Neville. 2007. An erp study of syntactic processing in english and nonsense sentences. *Brain research*, 1130:167–180.
- Zhao, Haiyan, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.