MASTER IN COGNITIVE SCIENCE AND LANGUAGE
MASTER THESIS
September 2026

Assessing the Epistemic Value of LLMs vs. Human Language

by *Laura López Zambrano*

Under the supervision of:
Evelina Leivada
Universitat Autònoma de Barcelona

Abstract: This thesis investigates the epistemological and cognitive validity of Large Language Models (LLMs) as models of human language, addressing whether the high-level linguistic performance of disembodied, transformer-based architectures justifies inferences regarding shared internal human-LLM cognitive mechanisms. Situated at the intersection of philosophy of language, generative linguistics, and 4E (Embodied, Embedded, Enacted, Extended) cognitive science, the central inquiry evaluates whether transformer self-attention mechanisms are structurally comparable to human embodied cognition and double-scope conceptual blending, or if their disembodied, gradient-optimized nature creates an impassable architectural boundary. Using a comparative framework, this study synthesizes neuroimaging evidence on Universal Grammar constraints, computational probing of recursive syntax, Fauconnier and Turner's Conceptual Blending Theory, and empirical assessments of LLM conceptual stability. The analysis demonstrates that while LLMs achieve impressive functional simulation of human text generation, their underlying computational primitives fundamentally diverge from those involved in human cognitive processing. Whereas human syntax is biologically channeled in specialized cortical networks and semantics is anchored in sensorimotor grounding, LLM outputs stem from statistical optimization over discrete sub-word tokens within ungrounded vector spaces. Consequently, transformers operate as meta-representational engines lacking referential truth-conditions, intrinsic intentionality, and stable conceptual cores. By bringing to the fore the performance-versus-cognition fallacy, this work contributes a refined theoretical framework separating formal linguistic fluency from functional cognitive competence, concluding that genuine semantic integration cannot emerge from disembodied distributional reweighting, but requires an embodied, biologically constrained, and environment-enacted mind.

Keywords: Philosophy of Language, 4E Cognition, Conceptual Blending Theory, Large Language Models, Formal vs. Functional Competence

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Evelina, for her exceptional guidance, insightful feedback, and continuous support throughout the writing of this thesis. Her sharp expertise and dedication were essential in shaping the intellectual trajectory of this work.

I am endlessly grateful to my parents, Merche and Javi, and my brother, Sergio, for supporting me not just through this project, but at every step of my life. To my partner, Issi, thank you for being there whenever doubts crept in. Her continuous encouragement helped me find clarity, and her patience in enduring endless discussions about language and machines was invaluable. I am also incredibly lucky to have friends who listen and support me no matter the distance between us. I love you all.

Finally, I want to express my gratitude to the generations of researchers, philosophers, and scientists who paved the way before me. This endeavor would not be possible without the collective scientific community working to make sense of our world, one breakthrough at a time.

List of abbreviations

4E: Embodied, Embedded, Enacted, Extended
AI: Artificial Intelligence
BPE: Byte-pair encoding
CBT: Conceptual Blending Theory
FLB: Faculty of Language in the Broad sense
FLN: Faculty of Language in the Narrow sense
fMRI: Functional magnetic resonance imaging
GRU: Gated Recurrent Unit
LM: Language Model
LLM: Large Language Model
RNN: Recurrent Neural Networks
UG: Universal Grammar
WS: World Scopes

List of figures

Index

"The question of whether a computer can think is no more interesting than the question of whether a submarine can swim."
—**Edsger W. Dijkstra**

1 Introduction

The rapid rise of transformer-based Large Language Models (LLMs) has fundamentally changed contemporary Artificial Intelligence (AI) and rekindled long-standing debates within the philosophy of language, linguistics, and cognitive science regarding the uniqueness of language in humans. Modern LLMs generate syntactically rich, context-sensitive, and pragmatically plausible text, achieving surface-level fluency that mimics human verbal output across a wide array of domain tasks (Binz et al., 2025). This phenomenon presents an urgent epistemological challenge: Does the high-level linguistic performance of disembodied, gradient-optimized transformer architectures justify inferences regarding shared internal cognitive mechanisms, or does it mask a fundamental architectural divergence between statistical token manipulation and human mental processing?

Within modern cognitive science, two primary paradigms offer different frameworks for understanding language and thought. The traditional computational-formalist paradigm, historically rooted in generative grammar, treats language as an autonomous, rule-governed symbolic capacity constrained by biological universals (Berwick & Chomsky, 2016). Under this view, linguistic competence is best captured through abstract structural legitimacy and recursive compositionality. On the contrary, the embodied, grounding paradigm, supported by 4E (Embodied, Embedded, Enacted, Extended) cognition, and Conceptual Blending Theory (CBT); argues that natural language semantics is inextricably bound to sensorimotor grounding, physical environment interaction, and dynamic mental space mappings (Lakoff & Johnson, 1999; Fauconnier & Turner 2003). As LLMs achieve great results on formal linguistic benchmarks, supporters of computational formalism frequently suggest that transformer self-attention mechanisms uncover implementation-independent universals of language (Piantadosi, 2023). However, this inference often commits a classic categorical mistake. By equating surface behavior with shared cognitive architecture, contemporary claims regarding LLM linguistic "competence" compare statistical sequence compression to genuine semantic comprehension, intentionality, and cognitive structure.

1.1 Contextualization and goal

The central question addressed in this thesis is whether the statistical attention mechanisms inherent to transformer architectures are structurally and mechanistically comparable to human embodied cognition and conceptual blending, or if their disembodied, text-only pretraining creates an impassable boundary that limits their validity as models of human language and mind. To address this question, this research consists of 3 core objectives: (i) to evaluate whether an LLM's behavioral preference for natural over counterfactual grammars reflects biological constraints or vector space optimization; (ii) to contrast human hierarchical, recursive syntax processing with linear sub-word tokenization; and (iii) to assess whether attention-driven vector blending structurally mirrors human double-scope conceptual integration. In a nutshell, this thesis defends the hypothesis that while transformer architectures achieve impressive functional emulation of human linguistic output through high-dimensional statistical pattern-

matching, they are structurally and mechanistically disanalogous to human language processing.

1.2 Structure of the work

To defend this hypothesis, the argument progresses through 3 interconnected thematic pillars. First, I examine the biological reality of Universal Grammar (UG) in human neurophysiology, reviewing fMRI evidence showing that the human brain selectively recruits dedicated networks (largely corresponding to Broca's area) exclusively for UG-compliant languages. I then delve into how inferring shared cognitive mechanisms from statistical surface similarities constitutes an ungrounded epistemological projection. Next, I explore the representational architecture of syntax, contrasting the human brain's hierarchical, tree-structured recursive processing with the transformer's reliance on Byte-Pair Encoding (BPE) sub-word tokens and soft self-attention weights. I show that LLM parameters lack referential truth-conditions, relying instead on human interlocutors to project intentional meaning onto generated text. Last, I evaluate semantic integration by contrasting CBT and 4E sensorimotor grounding with the scaled dot-product attention equation. Through empirical studies on LLM conceptual instability and World Scopes framework, I demonstrate that while attention mechanisms route and reweight context-dependent vectors, they lack the grounded conceptual cores, dynamic mental simulations, and active environmental feedback loops essential to human double-scope integration.

2 State of the art

To evaluate whether the linguistic capabilities of LLMs reflect genuine computational parallels to human cognition, we must systematically examine the empirical and theoretical frameworks that define modern language processing. This section provides a critical synthesis of current literature across generative linguistics, neuroimaging, computational modeling, and 4E cognitive science. By tracing the tension between computational formalism and embodied grounding paradigms, we analyze how structural constraints operate in human neurophysiology versus artificial neural architectures, exposing the theoretical and methodological boundaries that separate statistical performance from cognitive competence.

2.1 Universal Grammar and the Neural Activation of Structural Constraints

To assess the claim that LLMs conform to human-like linguistic universals, we must first establish how structural constraints operate within human cognitive physiology. A foundational pillar of generative linguistics is the notion of UG: the proposal that human language acquisition is constrained by innate biological boundaries that determine which grammatical systems are learnable and which are impossible (Chomsky, 1957). Under this framework, language acquisition is not merely the outcome of domain-general statistical learning, but the execution of a biologically channeled cognitive capacity.

The biological reality of these constraints has been empirically tested through investigating whether the human brain selectively responds to grammar that adhere to UG, while rejecting artificial systems that violate it. A neuroimaging study by Musso et al. (2003) aimed to determine whether language processing engages specialized neural structures or relies exclusively on domain-general learning mechanisms. To isolate structural legitimacy from simple familiarity, the authors designed "unreal" grammars: artificial systems that explicitly manipulated and violated core structural rules of natural

languages (such as linear precedence rules overriding hierarchical structure), rather than presenting random noise. Musso et al. (2003) argued that if language processing was mediated solely by domain-general statistical mechanisms, real and unreal grammars should activate similar neural networks and yield comparable learning trajectories. But, as the authors state: "When postulating an epigenetic role of a brain system in language learning, the acquisition of linguistic competence should involve this brain system only when the new language [...] is based on the principles of UG" (Musso et al., 2003, p. 774).

The functional neuroimaging data confirmed this specialized engagement. Learning UG-compliant rules selectively recruited Broca's area (specifically the left inferior frontal gyrus), whereas learning UG-violating rules failed to engage this dedicated network, relying instead on non-specific, domain-general neural systems. This selective activation held true across diverse natural language families: "Our results show that Broca's area has a key role in the acquisition of 'real' rules of language, independent of the linguistic family to which the language belongs" (Musso et al., 2003, p. 778).

The authors concluded that this provides explicit neurophysiological evidence that adult language acquisition "involves a brain system that is different from that involved in learning grammar rules that violate UG" (Musso et al., 2003, pp. 777-778). In humans, structural legitimacy is not an abstract mathematical property learned through unbounded optimization; it is a physiological constraint enforced by specialized neural architecture.

2.2 "Possible" vs. "Impossible" Languages in Computational Architectures

The empirical demonstration of biological constraints in human speakers inspired researchers in computational linguistics to apply the distinction between "possible" (UG-compliant) and "impossible" (UG-violating) languages to LLMs. The theoretical premise behind this translation is straightforward: if non-biological, gradient-based artificial neural networks also display a systematic preference for possible languages over impossible ones, then these linguistic constraints might reflect fundamental computational properties of language itself (thus learnable from data alone), rather than specific aspects of human biology. In a direct attempt to evaluate this hypothesis, Kallini et al. (2024) tested autoregressive transformer models across a continuum of counterfactual and syntactically impossible languages. Responding to prior claims that deep learning models learn arbitrary non-natural patterns with equal ease, Kallini et al. (2024) reported that autoregressive models exhibit marked performance degradation when trained on impossible grammars: "Our core finding is that GPT-2 struggles to learn impossible languages when compared to English as a control" (Kallini et al., 2024, p. 1).

Acknowledging the theoretical difficulty of defining the exact boundary of impossibility, the authors explicitly noted: "We do not feel positioned ourselves to assert such a definition, so we instead offer some examples of impossible languages on a continuum of intuitive complexity" (Kallini et al., 2024, p. 2). Despite this conceptual fluidity, the statistical disparity remained clear: transformers learn natural, hierarchically organized languages significantly faster and with lower cross-entropy loss than counterfactual systems based on arbitrary counting or inverted positional rules.

On the surface, this behavioral parity between humans and LLMs can lead to a tricky deduction. The implicit reasoning suggests that because humans learn possible, natural languages more efficiently than impossible ones through the mediation of UG, and because LLMs exhibit an identical performance advantage when processing possible over

impossible languages, the computational behavior of LLMs must therefore support the necessity of innate structural constraints operating under shared cognitive principles. Supporters of this view suggest that because "LLMs have been shown to learn the complex structures of human language and have a preference for learning such structures over unnatural counterfactuals, it follows that they are clearly relevant to investigations and claims about the necessary innate priors for language learning" (Kallini et al., 2024, p. 9).

However, this theoretical deduction contains a fundamental formal fallacy: affirming the consequent. Showing that a model exhibits behavioral preferences aligned with UG predictions does not demonstrate that the model achieves this behavior through UG-like mechanisms or constraints. The same behavioral output can be generated by entirely distinct underlying computational operations. Furthermore, Kallini et al. (2024) themselves highlight architectural facts that undermine any claim of shared mechanical principles. They emphasize that "models and humans exhibit fundamental differences" driven directly by "specific architectural decisions" (p. 2), such as static positional encodings, BPE tokenization, and soft causal attention masks. For example, "[m]odels can perform operations that involve counting tokens because LLMs rely on tokens as basic units" (Kallini et al., 2024, p. 9).

Where human cognitive processing operates on morphosyntactically grounded units and hierarchical constituent structures (Bolhuis et al., 2014), transformers calculate statistical dependencies over discrete, sub-word token indices. The model's difficulty with "impossible" languages often stems not from a violation of biological linguistic universals, but from a mismatch between the counterfactual grammar and the transformer's specific mathematical primitives, such as positional encoding limits or vector space optimization trajectories. Thus, a theoretical tension emerges. While surface behavioral similarities across possible and impossible languages are cited as evidence for shared cognitive constraints, the underlying processing units (BPE tokens vs. morphemes) and computational primitives (matrix multiplication over positional vectors vs. biologically constrained neural circuits) diverge in ways that undermine claims that take for granted system equivalence.

2.3 The Epistemological Fallacy: From Performance to Cognition

The tendency to project human cognitive structures onto artificial systems that demonstrate high-level behavioral competence is not a novel phenomenon unique to transformer architecture. It represents a recurring epistemological confusion in the history of cognitive science. In his critique of computational paradigms, Edwin Hutchins (1995) provided a clear exemplification of this tendency. Hutchins observed that when an artificial system successfully mimics human performance on a complex task, researchers routinely make an inferential leap, assuming the system employs the same internal mechanisms as human agents. Hutchins warned that failing to distinguish between system boundaries leads directly to structural misattribution: "Failing to recognize the cultural nature of cognitive processes can lead to a misidentification of the boundaries of the system that produced the evidence of intelligence" (Hutchins, 1995, p. 356). When system boundaries are improperly defined, the theoretical framework used to explain the behavior becomes distorted. As Hutchins (1995) further notes, researchers risk attributing "the right properties to the wrong system or (worse) invent the wrong properties and attribute them to the wrong system" (p. 356).

This historical diagnosis applies directly to modern claims regarding LLM linguistic performance. Because language is traditionally viewed as the primary marker of higher human cognition, the production of fluent, syntactically rich text by LLMs allowed some researchers to infer cognitive life into statistical surface patterns. This projection is further amplified by classical computationalism, which states that the brain is essentially a digital information-processing machine, an assumption Hutchins (1995) notes "turned out to be wrong" (p. 357). Historically, by "embodying the growing knowledge of human information processing psychology in computer programs" early cognitive scientists built models that successfully reproduced key aspects of human task behavior (Hutchins, 1995, p. 359). This historical success fostered the belief that artificial implementations of formal symbolic operations could directly reveal the nature of human thought. AI and psychology were treated as mutually confirming disciplines: one studying biological systems as computational engines, the other studying abstract intelligence across implementation-independent substrates.

However, as Hutchins (1995) highlights, this assumption is an extrapolation grounded in past heuristic successes rather than a demonstrated principle. High performance on a language task does not prove functional equivalence at the cognitive or mechanistic level. In contemporary LLMs, language generation is achieved via next-token prediction over vast multidimensional vector spaces. Inferring that this process mirrors human semantic comprehension or syntactic processing mistakes functional simulation for architectural identity. Without rigorous attention to physical implementation, architectural boundaries, and computational primitives, moving from *what* an LLM can produce to claims about *what language is* constitutes an ungrounded theoretical projection.

2.4 Methodological Pathology and Empirical Divergences

The theoretical confusion identified by Hutchins is further intensified by methodological inconsistencies in contemporary empirical literature comparing humans and LLMs. While certain studies highlight instances where LLMs match human sentence processing benchmarks, a broader look at the empirical landscape reveals conflicting, brittle, and difficult-to-interpret results. In a comprehensive methodological review, Leivada et al. (2024) critique the growing practice of attributing internalized grammatical competence to LLMs based on selective behavioral evaluations. They note that current evaluation strategies fail to establish whether a model possesses structural competence, asking "whether this truly is a superior method for approaching the internalized grammatical abilities of LLMs" (Leivada et al., 2024, p. 11).

Leivada et al. (2024) critique the current literature by identifying 3 pervasive methodological fallacies that distort evaluations of artificial models. First, researchers frequently make unlicensed generalizations by observing an LLM's success on a narrow syntactic probe and improperly inferring a systematic, domain-general mastery of the underlying rule. Unlike human cognition, where performance typically signals an integrated underlying competence, an LLM's task success is often highly fragile and prone to collapsing under slight prompt variations. Second, studies stumble into the human-like paradox by asserting that an LLM exhibits human-like behavior while simultaneously conceding that its massive data scale, training regimes, and underlying architecture share no plausible common ground with human biological learning. Finally, the literature relies on double standards by assessing humans and machines through asymmetric criteria, such as allowing LLMs thousands of hyperparameter adjustments or

zero-shot iterations while testing human subjects on single-exposure tasks, which falsely inflates the model's performance and creates an illusion of functional equivalence.

These methodological issues reveal a core problem: surface performance metrics alone cannot uncover the underlying representational nature of LLMs. Because transformer architectures operate as black boxes, their output can often be explained by multiple, mutually incompatible computational hypotheses. A model may pass a complex syntactic test not by constructing hierarchical trees, but by exploiting high-dimensional *n*-gram correlations present in its massive training dataset. Consequently, using sporadic performance overlaps with human subjects to claim that LLMs and humans share linguistic or cognitive mechanisms is theoretically ungrounded. Surface convergence may mask profound architectural divergence.

2.5 Grounding the Theoretical Gap

Synthesizing these insights reveals the central theoretical gap that motivates our broader research framework: the disconnect between distributional syntax reweighting and grounded semantic integration. The formalist tradition, regardless of whether it is instantiated in terms of a Chomskyan UG or simulated within transformer attention matrices, treats language as a closed, self-referential system of structural transformations. Under this view, if a system successfully maps inputs to outputs according to valid syntactic constraints, it is deemed to possess linguistic competence.

The empirical evidence reviewed across these domains highlights the fundamental limitations of assuming that formal syntactic processing implies shared cognitive architecture. In human biological systems, language operates not as an abstract computational property, but as an epigenetically channeled constraint realized within specialized neural networks that interact dynamically with broader sensorimotor and cognitive systems (Musso et al., 2003). Also, while artificial neural networks may exhibit surface preferences for natural linguistic patterns, these outputs arise from token-based statistical optimization across discrete vector spaces, entirely lacking human physiological constraints and basic representational units like morphemes and words (Kallini et al., 2024; Leivada et al., 2024). From an epistemological standpoint, ascribing human-like cognitive mechanisms to an artificial model based purely on its surface task performance fundamentally misidentifies system boundaries, resulting in an ungrounded theoretical projection of biological cognition onto what are ultimately disembodied, statistical mechanisms (Hutchins, 1995).

Crucially, this literature highlights what formal computational models lack. As mentioned above, human language processing does not occur in a disembodied statistical vacuum. Human semantics relies fundamentally on embodied (4E) cognition, where linguistic meaning is grounded in sensorimotor interaction with the physical environment; and double-scope conceptual blending, which dynamically integrates different mental spaces to construct novel conceptual structures. While LLM self-attention layers provide powerful mechanisms for dynamic vector reweighting across context windows, they operate entirely within an ungrounded, disembodied statistical space. They lack sensorimotor grounding, real-world intentionality, and the capacity for true conceptual integration.

Confusing the structural capabilities of LLMs with human cognitive processing misinterprets both artificial neural networks and biological minds. By establishing that

surface structural parity does not *ipso facto* imply shared cognitive architecture, the way is paved for the next aims of this thesis: demonstrating how true semantic coherence requires sensorimotor grounding, attention-based cognitive routing, and dynamic conceptual blending; that is, capacities that remain entirely outside the scope of current transformer architectures.

3 Syntax and Representation: Hierarchical Constraints vs. Distributional Regularities

At the intersection of computational linguistics and cognitive science lies the question of whether the internal mechanisms of transformer architectures are structurally comparable to those underlying human linguistic competence. While LLMs display impressive performance when generating complex, grammatically well-formed sentences, a fundamental epistemological challenge remains: does this surface fluency reflect an internal representation of hierarchical, tree-structured grammar, or is it merely an optimization of high-dimensional, linear distributional regularities? To resolve this, we must examine the formal divide between human computational mechanisms and statistical models, evaluating how structural biases, inductive constraints, and distributional data shape representational capacity.

3.1 Hierarchical Representation of Grammar Structure in Humans

Human linguistic competence relies on a domain-specific capability to construct, manipulate, and interpret abstract, hierarchical representations. In their foundational framework, Hauser, Chomsky, and Fitch (2002) distinguish between the Faculty of Language in the Broad sense (FLB) and the Faculty of Language in the Narrow sense (FLN). The FLB has three parts working together: the sensory-motor system, the conceptual-intentional system, and computational mechanisms; and FLN is hypothesized to consist strictly of the computational core dedicated to recursion. Recursion provides the mind with the capacity to generate an infinite range of structured expressions from a finite set of discrete lexical elements (Hauser et al., 2002) This theoretical frame positions human syntactic competence not as an unconstrained surface-predictive engine, but as an internal system optimized to build structured, recursive representations that interact directly with intentional and perceptual states.

From a neurocomputational perspective, the human brain processes temporal sequences by mapping sequential auditory or visual inputs onto nested, abstract tree structures. Tracking sequence processing across cognitive architecture reveals that predictive mechanisms operate along different levels of representational abstraction (Dehaene et al., 2015). At a low level, sequence learning depends on tracking statistical transition probabilities between adjacent items, an unconscious, temporally detailed, and item-specific mechanism that records localized evens: “sequences are stored by keeping a record of the transitions between events and their approximate timing. At this level, the sequence code is unconscious, shallow, item specific, and temporally detailed. [...] This mechanism seems to be duplicated in many cortical and subcortical circuits.” (Dehaene et al., 2015, p. 5).

While non-human species can learn simple patterns by linking items that appear together, this mechanism is not sufficient to match the complexity of human language. Human grammar is built on structural relationships between words that can be separated by long-

distance dependences, which means that one cannot figure out how a sentence works just by looking at which words appear adjacent to each other. These long-distance dependencies connect words, even when separated by other words in a sentence. For this reason, simple pattern-detection, word order, or probability models that only look at adjacent words fail to capture grammar. Going beyond adjacency, the human mind processes and produces language by using nested, tree-like mental structures that link related words (both lexical and functional) across various syntactic positions and configurations (Dehaene et al., 2015). Neuroimaging studies confirm that the human brain constructs these hierarchies in real time. For example, cortical tracking of hierarchical linguistic structures shows that distinct neural processing timescales track words, phrases, and sentences concurrently, building abstract structural units independently of acoustic or surface statistical cues (Ding et al., 2016). This internal architecture operates through an innate, inductive bias for structural hierarchy causing human language processing to form abstract structural relations.

To evaluate whether neural networks mirror these human mechanisms, cognitive neuroscience systematically separates linguistic execution into two functional domains:

> "To mitigate the language-thought conflation fallacies, we propose to systematically distinguish between two kinds of linguistic competence: formal linguistic competence—the knowledge of rules and statistical regularities of language—and functional linguistic competence—the ability to use language in real-world situations. [...] Both formal and functional linguistic competence are essential components of human language use: an effective communicator needs to both generate grammatical, meaningful utterances and strategically use those utterances to achieve diverse, context-dependent goals." (Mahowald et al., 2024, p. 3)

In the human brain, these two dimensions are robustly dissociable. Formal competence relies on a dedicated, domain-specific frontoparietal language network that operates selectively on structural and syntactic relations (Ding et al., 2016; Mahowald et al., 2024). On the other hand, functional competence requires the integration of distributed neural circuits supporting social cognition, executive function, world knowledge storage, and action planning. This neurobiological dissociation demonstrates that formal syntactic mastery is not an all-purpose cognitive engine. Instead, it is an architectural interface designed to express, format, and structure thought for real-world application.

3.2 Representation in Large Language Models

To evaluate whether transformer-based architectures achieve a structural parallel to human formal competence, we must analyze the mechanisms by which LLMs process, encode, and predict text. Unlike human cognitive systems, which map discrete symbols onto mental models of physical and social reality, transformers operate strictly on tokenized textual sequences (Vaswani et al., 2017). Tokenization breaks continuous text into sub-word units based on corpus-wide statistical distributions, a process that frequently diverges from etymological, morphological, or syntactic boundaries (Ruyant, 2026). As a result, the foundational units processed by an LLM are not invariant linguistic symbols or semantic primitives, but *arbitrary* statistical fragments deployed for sequence compression.

The underlying transformer architecture relies on scaled dot-product self-attention mechanisms to calculate dynamic relationships between tokens across an input context window (Vaswani et al., 2017). Through multi-layer attention operations and feed-

forward networks, models build internal dense representations that capture complex contextual relationships, surface collocations, and stylistic features. Ruyant (2026) summarizes the common conceptual interpretation of these layers:

"It can be intuitive to view embedding tables as encoding the meaning of tokens in a semantic space, attention heads as encoding conceptual relations between them as well as grammatical and near-pragmatic rules governing their linguistic use, and feed forward network as storing the factual knowledge needed to predict the next word of a text." (Ruyant, 2026, p. 7)

However, mapping these engineered mechanisms onto human cognitive representations introduces serious theoretical misunderstandings. The internal parameter values of a transformer are not structured intentionally by human modelers, nor are they selected to mirror biological cognitive constraints. They are derived purely through gradient descent algorithms designed to minimize cross-entropy loss over a next-token prediction objective across massive, unstructured web-scale corpora (Bender et al., 2021; Ruyant, 2026). One might argue that this optimization extracts a linguistically-informed ordering of syntactic, semantic, or morphological primitives. However, such vector clusters reflect high-dimensional surface co-occurrences and linear heuristics rather than discrete, rule-governed representations of linguistic structure (McCoy et al., 2020). Consequently, interpreting transformer outputs as direct linguistic representations of external reality mischaracterizes the model's actual functional capacity.

Because tokens do not inherently map onto physical or intentional objects, the internal parameters do not maintain referential truth-conditions about the external world. Instead, as Ruyant (2026) demonstrates, LLMs are fundamentally meta-representational:

"A first reason to think that LLMs do not represent the world directly is that, as we have seen, tokens do not systematically correspond to meaningful linguistic units. [...] LLMs represent some linguistic phenomena rather than the world directly. [...] If this is correct, then some users misuse LLMs, and this poses a problem for intended-interpretation-views, since they cannot account for misuse." (Ruyant, 2026, pp. 2, 14–15)

This meta-representational status explains why LLMs generate fluent text alongside showing marked structural failures, such as systemic factual confabulations ("hallucinations") and sensitivity to irrelevant distractor tokens (Mahowald et al., 2024). The optimization target of an LLM is the statistical plausibility of a text string, completely disconnected from referential correctness, truth-tracking, or ground-truth verification. When a human user interacts with an LLM, the communicative intentionality, referential meaning, and semantic depth do not reside within the machine's internal parameters. Rather, they are projected onto the output by the human interlocutor; surging from the human tendency to anthropomorphize:

"In most if not all cases, the text represented by the user is fictional rather than actual. [...] When discussing with a chatbot, we should consider that we are engaged in a fiction: we are asked to imagine that a certain kind of conversation takes place, and the LLM is used as a tool to make inferences about how this imaginary conversation would typically unfold." (Ruyant, 2026, p. 17)

"Text generated by an LM is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind. [...] [O]ur perception of natural language text, regardless of how it was generated, is mediated by our own linguistic competence and our predisposition to

interpret communicative acts as conveying coherent meaning and intent, whether or not they do." (Bender et al., 2021, p. 616).

To sum up, meaning is not generated by the statistical model itself. It emerges only when an intentional subject engages with the model's output, reading subjective relevance into a sequence of statistically calculated tokens.

3.2.1 Architectural Constraints, Statistical Distribution, and Training Data

The functional rift between human formal competence and LLM output becomes most apparent when evaluating how each system processes complex, recursive syntax. Human syntactic competence is characterized by a strong inductive bias toward hierarchical tree structures, enabling humans to process nested dependencies easily while maintaining systematicity across novel constructions (Hauser et al., 2002; McCoy et al., 2020). Standard transformer architectures, by contrast, lack an explicit pushdown stack or dedicated recursive mechanism. They rely instead on high-dimensional soft-attention weights to approximate linear and hierarchical dependencies implicitly (Murty et al., 2023).

When tested on recursive structures that diverge even slightly from their training distribution, standard transformers exhibit severe performance drops. Lakretz et al. (2022) demonstrate that both causal and masked[1] transformer LLMs fail consistently when evaluating complex recursive nested constructions, specifically long-distance subject-verb agreement across embedded relative clauses. The fact that both causal and masked transformers fail on the same inner-dependency constructions suggests their processing strategies may mirror older recurrent neural networks (RNNs) models. According to Lakretz et al. (2022), these below-chance results show a need to re-examine how transformers achieve their reported syntactic successes and why they struggle to generalize across subtle grammatical differences. Overall, these failures highlight a critical architectural constraint: transformer models do not acquire a generalizable, rule-based hierarchical bias merely from exposure to vast linear sequences.

In probing whether hierarchical generalization can emerge purely from sequence processing, McCoy et al. (2020) demonstrated that sequential neural models, when presented with training data ambiguous between linear rules and hierarchical rules, consistently adopt brittle linear heuristics (such as *n*-gram[2] co-occurrence or sequence position) rather than robust hierarchical generalizations:

"The lack of hierarchical bias in the sequential GRU [Gated Recurrent Unit] with bracketed input indicates that simply providing parse information in the input and target output is insufficient to induce a model to favor hierarchical generalization; it appears that such parse information must be integrated into the model's structure to be effective [...] these findings suggest that the hierarchical preference displayed by humans when acquiring English requires making explicit reference to hierarchical structure, and cannot be argued to emerge from more general biases applied to input containing cues to hierarchical structure." (McCoy et al., 2020, pp. 133, 135)

[1] Causal transformers predict the next token by looking only at preceding text, in a left-to-right manner; whereas masked transformers predict tokens by attending to the entire surrounding context simultaneously, both left and right.

[2] An *n*-gram is a sequence of *n* consecutive elements from a text. This is used in computational linguistics to predict language based on local statistical patterns of *n* length.

To bridge this generalization gap, computational researchers often augment transformers with explicit structural mechanisms. For example, pushdown transformers incorporate structured pushdown layers directly into the architecture, enabling the model to explicitly operate on stack-based, recursive constituency parses (Murty et al., 2023). The necessity of such structural interventions confirms that standard transformer self-attention mechanisms do not naturally induce human-like syntactic processing. Instead, standard transformers rely on high-capacity pattern-matching over surface distributional cues. This structural limitation is further compounded by the nature of the training data itself. Ungrounded text corpora contain only surface form, lacking the perceptual, prosodic, and physical context that surrounds human language acquisition (McCoy et al., 2020; Bender et al., 2021). Human language learning does not occur in a vacuum of isolated text strings; it is embedded within multimodal, active perceptual interactions with the environment. As Bender et al. (2021) observe:

"In accepting large amounts of web text as 'representative' of 'all' of humanity we risk perpetuating dominant viewpoints... languages are systems of signs, i.e. pairings of form and meaning. But the training data for LMs is only form; they do not have access to meaning. Therefore, claims about model abilities must be carefully characterized." (Bender et al., 2021, pp. 614–615)

Because large-scale pre-training data consists exclusively of ungrounded forms, transformers are fundamentally constrained to modeling surface statistical correlations. I argue that the apparent syntactic competence of LLMs is therefore a *high-dimensional illusion of scale*: an advanced capacity to approximate hierarchical surface structures without the underlying structural constraints, recursive stack architectures, or embodied grounding that define human linguistic competence.

3.3 Implications for Cognitive Modeling

The structural differences between human syntactic processing and transformer sequence prediction reveal an essential epistemological boundary in cognitive science. As argued before, humans construct linguistic representations using a domain-specific, hierarchical system (FLN) that interfaces with perceptual and intentional modules to support goal-directed action (Hauser et al., 2002; Dehaene et al., 2015; Mahowald et al., 2024). Transformers, on the other hand, process ungrounded sub-word tokens via self-attention matrices to generate plausible sentence completions based on distribution-wide co-occurrences (Vaswani et al., 2017; Bender et al., 2021; Ruyant, 2026).

Mixing up surface linguistic performance with internal cognitive structure leads to the performance vs. cognition fallacy: high accuracy on formal linguistic tasks does not imply that a model shares structural, neural, or computational principles with the human mind. While transformers can be valuable tools for modeling linguistic surface patterns and statistical properties of text corpora, treating them as valid, direct cognitive models of human language acquisition or representation confuses functional analogy with structural equivalence. Recognizing this boundary is crucial for advancing both AI engineering and cognitive neuroscience: human language is not a matter of linear token prediction, but an embodied, hierarchically constrained capacity deployed for the expression of thought.

4 Semantics and Embodiment: Conceptual Blending vs. Attention-Based Integration

A central issue in contemporary philosophy of language and cognitive science is whether LLMs function merely as sophisticated statistical mimics or whether they represent legitimate cognitive models of human language. This debate hinges on the relationship between linguistic form, semantic architecture, and cognitive embodiment. Human language processing is not a disembodied manipulation of abstract symbols; rather, it is anchored in sensorimotor systems, cultural practices, and dynamic cognitive processes. Among these, double-scope conceptual blending stands out as a core mechanism of human creativity and linguistic flexibility.

By examining the structural mechanisms of human linguistic competence, with a special focus in Conceptual Blending Theory and the 4E cognition paradigm, against the inner workings of Transformer-based attention architectures, we can evaluate their comparability. While multi-head self-attention mechanisms can combine information across different representation spaces, fundamental structural disanalogies remain. The absence of embodied, active environmental interaction in language models results in fragile conceptual structures, highlighting the computational limits of disembodied statistical learning.

4.1 Conceptual Blending Theories: How Humans Integrate Concepts

Human linguistic competence relies on a complex architecture that integrates different conceptual domains into novel, emergent mental spaces. In CBT, Gilles Fauconnier and Mark Turner (2003) argue that the capacity for complex, "double-scope" conceptual integration is the defining feature of human thought and language. In their words, mental spaces are not exhaustive encyclopedic entries; they are "small conceptual packets constructed as we think and talk, for purposes of local understanding and action—they are very partial assemblies containing elements, structured by frames and cognitive models" (Fauconnier & Turner, 2003, p. 58). Language does not explicitly encode human thought. Instead, grammar serves as a culturally entrenched prompt for complex conceptual operations that the mind executes automatically. As Fauconnier and Turner (2003) argue:

"A language is a powerful culturally developed means of creating and transmitting blending schemes. The capacity for language depends intricately on the capacity for blending and compression. The patterns we find in a language are the surface manifestation of blending schemes that have emerged within a culture and that have wide applicability [...]. The language itself does not have to carry such operations as compression or pattern completion because human brains supply those operations at no linguistic cost." (pp. 69, 76)

In its foundational form, a conceptual integration network consists of 4 connected mental spaces: two or more partially matched input spaces, a generic space reflecting structure shared by the inputs, and a blended space.

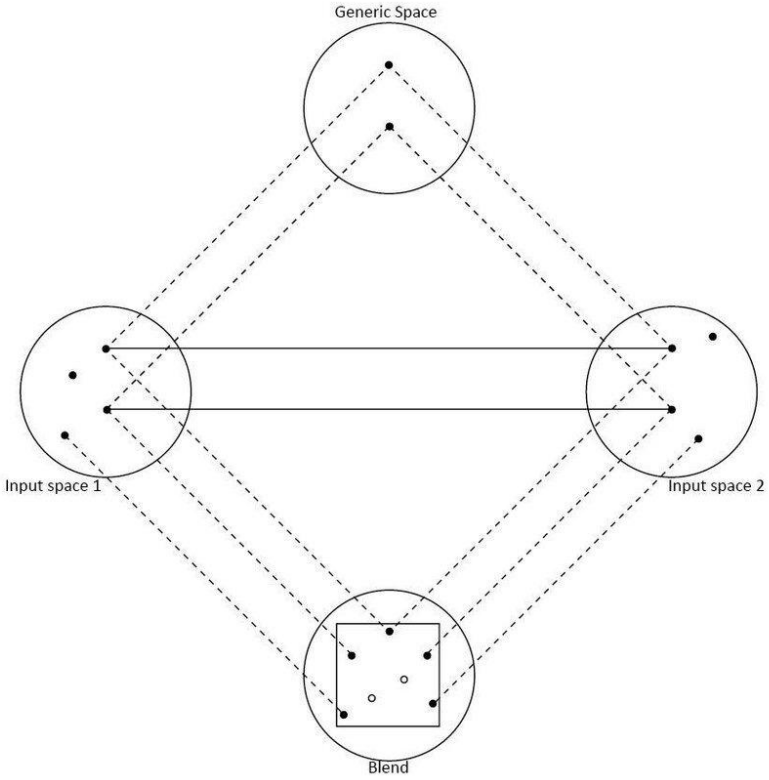

Figure 1: Diagramatic representation of conceptual blending integration. Reprinted from: (Fauconnier & Turner, 2003).

The blended space inherits selective projections from the inputs, resulting in an emergent structure through three operations: Composition, the juxtaposition of elements from distinct input spaces that do not co-exist outside the blend; Pattern Completion, the unconscious recruitment of background knowledge, frames, and cultural models to complete the newly composed structure; and Elaboration, the dynamic "running" of the blend as a cognitive simulation to test imaginative outcomes.

Neurobiologically, elements in mental spaces correspond to activated neural assemblies, while cross-space mappings reflect neurobiological binding and co-activation in working memory. The interface between syntax and conceptual structure is mutually constitutive: "Far from being an independently specified set of forms, grammar is an aspect of conceptual structure and its evolution" (Fauconnier & Turner, 2003, p. 86). Syntax prompts the construction of complex integration networks, providing a compressed form for diffuse conceptual relations. For instance, in caused-motion constructions (e.g., "she sneezed the napkin off the table"), a syntactic template compresses an agent-action event with a spatial trajectory, relying on an established generic space to integrate these distinct domains.

This conceptual architecture is rooted in human physical and perceptual existence. Cognitive linguistics and 4E cognition (Embodied, Embedded, Enacted, Extended) demonstrate that abstract human concepts derive from sensorimotor experience (Lakoff & Johnson, 1999; Carney, 2020). Image schemas (such as *CONTAINER*, *SOURCE-PATH-GOAL*, and *BALANCE*) emerge from bodily interactions with the physical environment and organize abstract domain representations (Johnson, 1987). Neuroimaging studies confirm this grounding: comprehending motor-related metaphors (e.g., "grasping an idea") recruits premotor and primary motor visual areas (Boulenger et

al., 2009; Desai et al., 2011), while spatial motion metaphors activate motion-sensitive visual areas like MT/V5 in the visual cortex (Saygin et al., 2010). Human conceptual blending operates over mental spaces that are grounded in sensorimotor systems and continually calibrated through active environmental engagement.

4.2 LLM Integration: Mechanisms, Coherence, and Constraints

To evaluate whether transformer architectures possess structurally comparable mechanisms, we must examine how modern LLMs integrate distributed tokens. Early connectionist networks mapped words to fixed vectors or discrete symbolic structures. In contrast, current LLMs use multi-head self-attention mechanisms to construct contextualized representations (Vaswani et al., 2017). Mathematically, a transformer computes attention over input sequences by projecting token embeddings into Query (Q), Key (K), and Value (V) matrices. The interaction between tokens is governed by scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Figure 2: Attention function responsible for calculating the relevance scores between all tokens in a sentence. Reprinted from: (Vaswani et al., 2017).

This operation dynamically reweighs token-level representations based on surrounding context. Multi-head attention projects input into distinct representation subspaces, allowing the model to attend to information from different token positions simultaneously. From an algorithmic perspective, this mechanism performs dynamic semantic integration (Sato, 2025). The input spaces correspond to token clusters prompted within a sequence; the generic space reflects background statistical distributions learned from pretraining data; and the generated output acts as a blended space displaying novel composition. Multi-head attention modules function as computational "integrators", dynamically routing, compressing, and recontextualizing representations across disparate sequence domains.

However, serious structural disanalogies emerge when comparing attention-based integration with human double-scope conceptual blending. As Sato (2025) notes, while transformer outputs can mimic conceptual blending at the computational level, this integration occurs entirely within a closed text-to-text vector space, unconstrained by physical reality or sensorimotor feedback. The fundamental limitation of transformer-based semantic integration lies in the disembodied nature of its training signal. As Bisk et al. (2020) emphasize in their framework of World Scopes[3], text-only pretraining (WS2) omits the foundational layers of human semantic acquisition: multimodal perception (WS3), embodied physical interaction (WS4), and social agency (WS5).

"Meaning does not arise from the statistical distribution of words, but from their use by people to communicate. Many of the assumptions and understandings on which communication relies lie outside of text. [...] In human development, interactive multimodal sensory experience forms the

[3] The World Scopes framework (Bisk et al., 2020) outlines five expanding levels of contextual grounding required for natural language understanding: WS1 (Corpus), WS2 (Internet), WS3 (Perception), WS4 (Embodiment), and WS5 (Social).

basis of action-oriented categories. [...] Embodiment—situated action taking—is therefore a natural next broader context." (Bisk et al., 2020, pp. 8718, 8722)

Because language models operate without bodily presence, physical agency, world models, or real-time environmental interaction, their conceptual representations lack a stable semantic core. This structural difference has empirical consequences. Investigating the internal conceptual organization of LLMs, Suresh et al. (2023) demonstrated that while models produce fluent surface text, their underlying conceptual structures are fundamentally brittle and lack internal coherence across different evaluation prompts: "The LLM may not have a coherent conceptual "core" driving its behaviors, and for this reason, may organize its internal representations quite differently with changes to the task instruction or prompt." (Suresh et al., 2023, p. 9).

Human conceptual structure remains stable across varied task demands because it is tied to sensorimotor systems and real-world interactions. In contrast, an LLM's semantic space shifts based on prompt context. The vector representation of a word is computed as a weighted average of surrounding tokens, meaning tokens lack stable semantic identity outside their immediate context window. What presents as creative conceptual blending in an LLM is often context-sensitive statistical alignment, unanchored by an underlying conceptual model. Furthermore, attention mechanisms lack the dynamic, active simulation capabilities characteristic of human conceptual blending. In human cognition, double-scope blending relies on "running" a blend, which means using background knowledge to simulate actions and evaluate counterfactual outcomes (Fauconnier & Turner, 2003). LLMs do not run parallel simulations over unobserved states of affairs; they calculate token probability distributions based on sequential optimization.

As James Carney observes, deep learning architectures operate within formal computational frameworks that do not incorporate the active, action-oriented loops central to 4E cognition: "Advances in areas like deep learning have been achieved without any reference at all to 4E perspectives. [...] [T]he 4E theorist must pay for a formalized notion of embodied cognition by relinquishing the claim that cognition is not computational in nature" (Carney, 2020, pp. 81–82). While attention mechanisms enable cross-domain information routing, they do not replicate the active environmental intervention through which human minds test hypotheses and build stable concepts.

4.3 Structural Disanalogies and the Limits of LLMs as Cognitive Models

To determine whether LLMs can legitimately function as cognitive models of human language, we must distinguish between functional simulation and structural identity. Modern LLMs demonstrate that disembodied statistical learning over text corpora can indeed approximate complex human linguistic behaviors. They quite convincingly manage syntax, construct metaphors, and combine concepts from disparate domains in ways that mirror human conceptual blending outputs.

However, this functional overlap masks deep structural disanalogies in the underlying mechanisms: Top-down prompting and grounded compression illustrate a key distinction between human and artificial cognition, as human syntax acts as a compressed prompt for implicit, entrenched mental space integrations anchored in physical, spatial, and bodily experience (Lakoff & Johnson, 1999; Fauconnier & Turner, 2003), whereas LLM "syntax" consists of learned spatial relationships within high-dimensional vector spaces driven by statistical optimization without reference to bodily experience or environmental

affordances. Furthermore, while human cognitive blending operates from a stable, grounded conceptual core that maintains conceptual coherence across contexts and task demands, LLM "semantics" relies on contextual fluidity, resulting in context-dependent vector states that shift significantly when probed with different prompting formats and producing brittle conceptual structures (Suresh et al., 2023). Finally, in contrast to human double-scope integration, which relies on running mental simulations, testing hypotheses through physical action, and actively refining concepts via environmental feedback (Bisk et al., 2020; Carney, 2020), transformers depend on passive pattern matching where self-attention computes static matrix multiplications across fixed context windows, lacking active environmental loops and goal-directed interventions.

For transformer architectures to achieve structural comparability with human conceptual integration, significant architectural changes would be required. Merely scaling up dataset sizes or parameter counts within disembodied, text-only systems (WS2) will not bridge this gap. As Bisk et al. (2020) argue, progress requires architectures grounded in multimodal sensory streams, embedded in physical or simulated environments, and capable of interactive, goal-directed behavior. Such systems would need to replace static context-vector updates with persistent and yet dynamically updated, embodied memory structures that learn through active environmental interaction.

In conclusion, attention-based integration in LLMs serves as a useful computational analog to human conceptual blending, but it is not structurally identical to human linguistic competence. LLMs illustrate how far statistical pattern matching over text can go in simulating *in silico* the surface structures of human language. However, because they lack sensorimotor grounding, real-time bodily agency, and stable conceptual coherence, they remain disembodied simulators of human linguistic artifacts rather than true models of embodied human cognition.

5 Concluding remarks

This thesis has evaluated the structural, computational, and cognitive claims surrounding state-of-the-art LLMs, charting the precise boundaries between surface behavioral emulation and internal architectural identity. By synthesizing insights across neurology, formal linguistics, 4E cognitive science, and machine learning, I have argued that modern transformer architectures do not instantiate the cognitive mechanisms that define human language.

In the discussion of formal syntax, it was established that human language acquisition is constrained by biologically channeled neural circuits. Human neuroimaging confirms that Broca's area selectively processes UG-compliant rules while rejecting counterfactual, unnatural grammars. Although transformers show statistical performance drops when trained on "impossible" languages, attributing this behavior to human-like linguistic universals relies on a formal fallacy. LLM performance disparities stem from tokenization artifacts, positional encoding constraints, and gradient optimization trajectories over discrete vector spaces, rather than biological structural priors.

In dissecting syntactic representations, I have highlighted the fundamental distinction between human formal competence, driven by an innate inductive bias for nested, recursive tree structures, and transformer sequence prediction. Standard self-attention mechanisms do not naturally induce abstract hierarchical rules; instead, they exploit high-

dimensional surface correlations and linear heuristics. When confronted with complex, out-of-distribution recursive constructions, standard models fail, requiring explicit structural interventions like pushdown layers to parse syntax. Moreover, as meta-representational engines, LLM parameters do not track referential truth conditions about an external world; semantic intentionality remains an interpretative projection imposed by the human user.

The analysis of semantics demonstrated the computational limits of disembodied statistical learning. Human conceptual structure relies on double-scope conceptual blending and 4E sensorimotor grounding, where abstract linguistic prompts activate pre-existing, embodied image schemas calibrated through physical agency and environmental interaction. While the scaled dot-product attention matrix performs mathematical vector routing across context windows, it operates within an ungrounded, disembodied vector space. Lacking sensorimotor grounding across broader World Scopes, LLMs exhibit severe conceptual instability when prompt conditions shift, failing to maintain stable conceptual cores or run active, goal-directed mental simulations.

To conclude, this research makes 3 distinct contributions to the philosophy of mind, philosophy of language, and cognitive computational modeling: it provides an interdisciplinarily informed critique of the performance-versus-cognition fallacy to prevent the uncritical attribution of human mental architecture to surface statistical artifacts; it offers an integrated theoretical framework separating formal linguistic fluency from functional cognitive competence; and it provides a fine-grained comparative analysis proving that dynamic vector reweighting cannot act as a substitute for dynamic sensorimotor conceptual grounding.

The boundary conditions identified in this work point toward clear limits for current AI architectures. As long as models remain confined to text-only pretraining, scaling parameter count or training corpus size will yield diminishing returns regarding true semantic comprehension, intentionality, and conceptual stability. Text-only models will remain sophisticated meta-representational simulators of human linguistic artifacts. To move toward computational architectures that achieve structural comparability with the human mind, cognitive science and machine learning must pivot toward embodied multimodal agents, persistently grounded memory structures, and hybrid architectures that combine soft pattern-matching with explicit recursive modules. Ultimately, language is not an isolated, self-referential matrix of ungrounded statistical dependencies; it is the evolutionary and cultural crest of an embodied, biologically constrained, and environment-enacted mind. True intelligence, semantic depth, and genuine conceptual integration do not emerge from the mere optimization of disembodied form; they require a body that acts, a world that responds, and a mind that intends.

6 References

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM FAccT Conference*, 610–623. https://doi.org/10.1145/3442188.3445922

- Berwick, R. C., & Chomsky, N. (2016). *Why Only Us: Language and Evolution*. Cambridge, Massachusetts: MIT Press.

- Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience Grounds Language. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718-8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

- Binz, M., Akata, E., Bethge, M. *et al.* (2025) A foundation model to predict and capture human cognition. *Nature* 644, 1002–1009. https://doi.org/10.1038/s41586-025-09215-4

- Bolhuis J.J., Tattersall I., Chomsky N., Berwick R.C. (2014) How Could Language Have Evolved? *PLoS Biol* 12(8): e1001934. https://doi.org/10.1371/journal.pbio.1001934.

- Boulenger, V., Hauk, O., & Pulvermüller, F. (2009). Grasping ideas with the motor system: Semantic somatotopy in idiom comprehension. *Cerebral Cortex*, *19*(8), 1905–1914. https://doi.org/10.1093/cercor/bhn217

- Carney, J. (2020). Thinking avant la lettre: A Review of 4E Cognition. *Evolutionary Studies in Imaginative Culture*, *4*(1), 77-90. https://doi.org/10.26613/esic.4.1.172

- Chomsky, N. (1957). *Syntactic Structures*. Mouton & Co

- Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The neural representation of sequences: From transition probabilities to algebraic patterns and linguistic trees. *Neuron*, 88(1), 2–19. https://doi.org/10.1016/j.neuron.2015.09.019

- Desai, R. H., Binder, J. R., Conant, L. L., Mano, Q. R., & Seidenberg, M. S. (2011). The neural career of sensory-motor metaphors. *Journal of Cognitive Neuroscience*, *23*(9), 2376–2386. https://doi.org/10.1162/jocn.2010.21596

- Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in natural speech. *Nature Neuroscience*, 19(2), 158–164. https://doi.org/10.1038/nn.4209

- Fauconnier, G., & Turner, M. (2003). *CONCEPTUAL BLENDING, FORM AND MEANING*.

- 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

- Hutchins E. (1995). Cultural Cognition. In: *Cognition in the Wild*. Cambridge, MA: MIT Press, 353-374.

- Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago Press.

- Kallini, A., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). *Mission Impossible: Language models struggle to learn unnatural counterfactual languages*. arXiv preprint arXiv:2402.17663.
- Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. Basic Books.
- Lakretz, Y., Dielemans, C., Subbiah, M., Lampinen, A., Baroni, M., & Dehaene, S. (2022). Can transformers process recursive nested constructions, like humans? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3225–3232.
- Leivada, E., Murphy, E., & Marcus, G. (2024). Differentiating performance from competence in Large Language Models. *Nature Machine Intelligence*, 6(1), 10–13.
- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models: A cognitive perspective. *Trends in Cognitive Sciences*, 28(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011
- McCoy, R. T., Frank, R., & Linzen, T. (2020). Does syntax need to grow on trees? Sources of hierarchical inductive bias in a LSTM language model. Transactions of the Association for Computational Linguistics, 8, 157–173. https://doi.org/10.1162/tacl_a_00313
- Murty, S., Sharma, A., Manning, C. D., & Bowman, S. R. (2023). Pushdown layers: Encoding recursive structure in transformer language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3234–3245.
- Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Büchel, C., & Weiller, C. (2003). Broca's area and the language instinct. *Nature Neuroscience*, 6(7), 774–781.
- Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint*. https://lingbuzz.net/lingbuzz/007180.
- Ruyant, Q. (2026). What do large language models represent? *Synthese*, 207(1), Article 17. https://doi.org/10.1007/s11229-025-05416-6
- Sato, M. (2025). *Conceptual Blending, Neural Dynamics, and Prompt-Induced Transitions in LLMs.* arXiv preprint arXiv:2505.10948
- Saygin, A. P., McCullough, S., Alac, M., & Emmorey, K. (2010). Modality-specific neural systems for biological motion processing. *Current Biology*, *20*(21), 1906–1910. https://doi.org/10.1016/j.cub.2010.09.063
- Suresh, S., Mukherjee, K., Yu, X., Huang, W.-C., Padua, L., & Rogers, T. T. (2023). *Conceptual structure coheres in human cognition but not in large language models* (arXiv:2304.02754). arXiv. https://doi.org/10.48550/arXiv.2304.02754
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.