Compositional Meaning Representations in LLMs: a Critical Review of Probing Studies

Rémy Marro
Tilburg University, Department of Communication and Cognition,
Warandelaan 2, 5037AB Tilburg, The Netherlands
r.marro@tilburguniversity.edu

Abstract

Large language models (LLMs) appear successful in emulating compositional language, yet it remains unclear what these results entail about their underlying compositional semantic representations. The probing classifier paradigm has emerged as a tool to remedy this. This paper proposes to critically review the findings of 24 probing studies targeting a wide range of linguistic and semantic phenomena. It proposes a taxonomy of probing tasks based on the linguistic primitives they presuppose, distinguishing four tiers: lexical semantics, the syntax–semantics interface, propositional semantics, and discourse and pragmatics. A gradient in representational evidence emerges: LLMs robustly encode lexical information, display less consistent sensitivity to structural relations within sentences, and obtain unsatisfactory results on tasks requiring propositional content, speech acts, or pragmatic inference. The review underscores the need for a clearer theoretical grounding of what probing tasks measure and reflects on how probing can illuminate the compositional pathways available within current language models.

1 Introduction

The compositionality principle is considered to be a hallmark of human cognition and is said to be pivotal in humans' ability to routinely understand and produce novel linguistic input (Partee et al., 1995). This principle – also known as Frege's principle – is characterized by a one-to-one mapping between structure (i.e., syntax, but also word-internal or discourse structure) and semantic interpretation, abiding by the mantra "the meaning of the whole is determined by the meaning of its parts and the way they are syntactically combined". While contested, it accounts for two epochal features of human language: systematicity and productivity. Productivity refers to the capacity to generate an unbounded number of novel linguistic outputs from a finite set of elements, while systematicity denotes the structured interdependence and reciprocity among constituents, as exemplified by the ability to discern that the proposition "John loves Mary" recombines to "Mary loves John" (Hadley, 1994). These two properties have been taken to pose a fundamental challenge for connectionist models of cognition. Fodor and Pylyshyn (1988) therefore identified two horns of the compositionality dilemma: (1) whether neural networks could exhibit compositional behaviors, that is, behaviors that are both systematic and compositionally productive, and (2) if the latter would hold, whether neural networks would have merely implemented symbols, thus failing to provide an alternative hypothesis to the principle of compositionality.

Contemporary transformer-based language models (Vaswani et al., 2017) revived this dilemma by notably undermining its first horn, and by foregrounding the explanatory burden to the distributional view of language composition. Large language models (LLMs) have indeed been incredibly successful in emulating compositional language, as shown by their performance on a wide array of natural language processing tasks (see Chang and Bergen, 2024 for a state-of-the-art review). However, extended scrutiny on transformers' compositional behaviors has revealed limited ability to derive the meaning of larger constituents based on their sub-units and generalize outside of their training data (Furrer et al., 2021; Keysers et al., 2020; Kim and Linzen, 2020; Lake and Baroni, 2018). Thus, evaluating LLMs purely in terms of behavior has proven misleading in many regards (Mitchell and Krakauer, 2023; McCoy et al., 2019), such that the focus has shifted towards mechanistic interpretation of language processes.

The probing classifier paradigm has therefore emerged as one of the principal methods to explain LLMs' behavior in terms of the linguistic representations they could leverage. The method

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 198–210

consists of linking the internal states of a language model with an identifiable external property of its training sets by training a simple classifier on its intermediate layers (Belinkov, 2022). One of the key tenets of the probing paradigm resides in the attempt to map properties inherited from symbolic and compositional theory of meaning onto the reputedly continuous word and sentence embeddings of language models. Claims about representational evidence are therefore inseparable from the compositional pathway – their mental functions and input/output thereof – posited by linguistic theory. This paper aims to review the findings that have emerged from this experimental tradition with regards to the compositional theory of meaning they (covertly) endorse. It will critically evaluate what the probing paradigm truly tells us about compositionality in transformer-based language models and what their successes and failures allow us to conclude about the transferability of linguistic theories of composition to LLMs. It shows that formal theories of composition only partially transfer to the findings of probing studies.

2 Probing Classifier as Semantic Diagnostic

2.1 Prospects and Limitations of Probing

Belinkov (2022) formally defines a probing classifier as follows: f is a model that maps an input x to an output $\hat{y}$: $f : x \mapsto \hat{y}$. The latter model, referred to as the original model, is trained on an annotated dataset $\mathcal{D}_O = \{x^{(i)}, y^{(i)}\}$, itself referred to as the original dataset. The original model performance is qualified as PERF$(f, \mathcal{D}_O)$, which is a metric measuring the ability of the model f to faithfully map inputs x to its respective output y. As the model under scrutiny is likely to be a neural network with several layers l, $f_l(x)$ is said to be the representation of x at the layer l. A probing classifier is then the function $g : f_l(x) \mapsto \hat{z}$, where $\hat{z}$ is a linguistic property present in the classifier g training dataset, distinct from the model dataset and noted $\mathcal{D}_P = \{x^{(i)}, z^{(i)}\}$. The classifier performance is a function of g, f and the datasets $\mathcal{D}_O$ and $\mathcal{D}_P$: PERF$(g, f, \mathcal{D}_O, \mathcal{D}_P)$. Typically, the performance is expressed in terms of its F1 score, which measures the harmonic mean of precision and recall, offering a balanced view of a classifier's ability to correctly identify both positive and negative instances. Albeit highly operationalizable, this methodology in itself endorses non-trivial theoretical assumptions as to linking pre-existing theoretical constructs to substrates of LM computation (see Buder-Gröndahl, 2023 for comprehensive overview). In addition, classifiers have proven very sensitive to the formalism utilized during dataset annotation (Kuznetsov and Gurevych, 2020).

Within this framework, it is mostly assumed that the linguistic property is linearly encoded in a model's internal representation. It is therefore advocated to keep the classifier as elementary as possible (Conneau et al., 2018), materializing in simple logistic regressions or in multi-layer perceptron pipelines, as more complex pipelines "bear the risk that the classifier infers features that are not actually used by the network" (Hupkes et al., 2018, see also Pimentel et al., 2020a; Hewitt and Liang, 2019). These issues are all consonant with what this paper calls the *functional indicativity* of the classifier. That is, whether good performance on probing tasks g is in fact indicative of genuine functional relevance in f. Linguistic properties may still be incidentally encoded in a model's internal representation, even though the property is not functionally relevant for f. More worryingly, in a series of synthetic tasks, probes still extracted properties from non-relevant sequences and from random noise (Ravichander et al., 2021; Zhang and Bowman, 2018). While some have professed the relevance of more sophisticated probes (e.g. vertex probing: Pimentel et al., 2020b; Voita and Titov, 2020), subsequent methodological refinements have explored the implementation of control conditions to mitigate this risk. For instance, Hewitt and Liang (2019), underlying that a classifier might be indicative of g and not of f, conceived a control condition for POS tagging by associating word types with random outputs, ensuring that the tasks are deterministic functions of word types that do not depend on context. Furthermore, Elazar et al. (2021) proposed amnesic probing as a solution to the challenge of assessing functional relevance. In this subparadigm, the property $\hat{z}$ is suppressed from the intermediate layer in the control condition before analyzing the model performance on the probing task. While these counterpoints have led researchers to be overtly vocal on the limitations of the probing paradigm in interpretability research, this paper argues that such a tool remains centrally informative of what linguistic representations a transformer model is not relying on (Ravichander et al., 2021).

2.2 Semantic Theory as Empirical Hypothesis

As such, while probing results establish the recoverability of linguistic properties from a model's internal representations, they do not in themselves specify the composition functions by which these properties are computed or used. In Marr's tripartite framework (1982), probing therefore operates at the computational level: it relates internal states to abstract properties, without providing access to the algorithms or mechanisms that realize the computation. Critically, as suggested by the issue of functional indicativity, probing suffers from a one-to-many mapping between internal representational states and computational explanations. This problem closely parallels the one identified as the Granularity Mismatch Problem (GMP: Poeppel and Embick, 2017) when linking fine-grained theoretical constructs to coarser substrate of neural computation. In this sense, Embick and Poeppel (2015) argues that mechanistic explanation requires explicit considerations of the computational primitives involved in the experimental design, instead of attempting to infer precise functions directly from implementation-level evidence. This motivates the role of higher-level computational frameworks like formal semantics, which characterize cognitive phenomena in terms of input/output pairs and thereby constrain the space of admissible functions (Egan, 2018; Krakauer et al., 2017).

This perspective suggests that interpretation should not rely on individual instances of recoverability, but on shared computational-representational commitments across tasks. In particular, manipulating a certain property $\hat{z}$ implicitly commits the probe to a functional hypothesis by assuming functional relevance of $\hat{z}$ in f. Probing tasks therefore inherit assumptions about the input/output structure of composition functions from the linguistic theories they operationalize. However, these functional assumptions should be made explicit as to the semantic operation they target, making semantic theory not a mere tool for explainability, but an empirical hypothesis that guide the search for precise composition functions.

Given these methodological and explanatory caveats, probing results should not be interpreted as decisive evidence for the presence of representations, but as a prima facie *diagnostic* to systematically discard implausible compositional pathways in language models. While probing studies have limited explanatory power on their own, a battery of tasks grouped by theory-driven input/output pairs constrains substantively the kinds of semantic composition functions that transformer-based models can be said to implement. Accordingly, the aim of this review is not to test whether language models instantiate (specific) formal representations, but to evaluate whether the primitives transfer to the empirical probing literature in such a way that it would not rule out the composition function posited by these theories. This leads to the following research question: How do the computational primitives presupposed by probing tasks constrain what they can reveal about the compositional functions of transformer-based language models?

3 Semantic Framework and Probes Classification

3.1 Review Methodology

This section first details the literature selection methodology, then introduces the semantic framework used to classify probing tasks. To scaffold a comprehensive account of the probing tasks targeting compositional structure, this paper relied on the literature review principle, which proposes to gather all papers tied to a previously stated research question through a systematic search with key terms and inclusion/exclusion criteria. As the field of artificial intelligence is confronted with a vertiginous publication rate and a reliance on conference papers, it was decided to query the literature with broad search terms[1] on the search engine Google Scholar. This search query yielded 422 results. This review then relied on inclusion and exclusion criteria to distinguish relevant from irrelevant pieces of work. Through succinct screening, this review identified 29 studies (as of February 2026) that complied with the inclusion criteria displayed in 1. An in-depth reading reduced the number of studies to 24 that used probing classifiers to investigate linguistic properties relevant to an inquiry of compositional meaning in LLMs.

Despite its broad coverage, this review does not claim to be exhaustive (see Waldis et al., 2024 for complete review of probing tasks). The probing methodology developed under a variety of labels before stabilizing terminologically, such that the tool is sometimes referred to as *diagnostic classifiers* or *linear probes* in the literature. Extending the search query to former terminology would

[1] Search query was the following: "probing classifier" AND "transformer" AND semantic

Criteria	Inclusion	Exclusion
Model	(Textual) language model	Speech model, Vision model, Programming model
Type of study	Empirical	Review, Theoretical
Types of probes	Structural probes	Behavioral probes
Component probed	Intermediate layers	Attention heads, Neurons, Circuits
Main dependent variable	Linguistic property $\hat{z}$ (what)	Probing classifier g (how), Layer l localization (where)
Training effect	Pre-training	Fine-tuning
Language	English, Multilingual (Cross-linguistic)	Extension to other languages (Russian, Italian, Arabic)

Table 1: Inclusion and Exclusion Criteria for Literature Review

be at odd with the fine-grained decompositional approach of probing tasks endorsed in this paper. As a result, some relevant studies may not be retrieved by the specific search query adopted here (e.g. Jumelet et al., 2021; Metheniti et al., 2022; Miaschi et al., 2020). To remain faithful to the principles of systematic review and to avoid post hoc cherry-picking, no study was retroactively included.

3.2 Task Classification and Semantic Framework

If studies usually investigated multiple semantic phenomena at once, tasks were individually classified according to the semantic domain it targeted. Given the theoretical importance of input/output pairs, classifying probes according to the theoretical constructs targeted over the methodology or the models leveraged becomes instrumental in unravelling potential pathway of language processing. Therefore, this paper decomposes compositional structure into four tiers of meaning: lexical semantics, the syntax-semantics interface, propositional semantics, and discourse semantics. Under this view, each tier operates at a distinct level of granularity and considers different primitives and composition functions, which together form the basic units of the theoretical framework adopted in this review. Inspired by contemporary formal accounts, it assumes a view of meaning that is simultaneously type-driven, model-theoretic, and truth-conditional, but remains semantically underdetermined and gets enriched pragmatically.

At the lexical tier, meaning is analyzed in terms of minimal semantic features and ontological types (e.g., physical object, human, event), defined relationally within a structured system. Words are treated as sets of features whose meaning arises through hierarchical (hypernymy/hyponymy) and mereological (part–whole) relations (Pustejovsky, 1998). Composition at this level consists primarily in feature aggregation and inheritance across systematically related lexical items.

At the syntax–semantics interface, composition operates over structured mappings from syntax to semantic interpretation (e.g. Heim and Kratzer, 1998). The central primitives are syntactic categories, semantic types, predicates, and their arguments. Meaning is constructed through function–argument application, where predicates combine with their arguments in a structure-sensitive manner. Tasks targeting part-of-speech, dependency relations, semantic roles, or predicate morphosemantics implicitly presuppose this level of composition.

At the propositional tier, sentence meaning is treated as expressing a proposition that can be evaluated for truth with respect to a model of the world (Tarski, 1944). Composition at this tier consists in assembling these elements into truth-evaluable expressions, for instance with logical connectors. Consequently, logical reasoning tasks, entailment, quantification test whether models support composition functions that go beyond local predicate–argument saturation toward truth-conditional abstraction.

Finally, at the discourse and pragmatic tier, fully formed propositions are integrated into broader communicative settings. This last tier of meaning takes as primitives the speaker's intentions and hearer's mental states, as the derivation of meaning depends on contextually-driven inferences beyond what is expressed at the sentence level. Composition then takes the form of inference and pragmatic enrichment, whereby propositional content is modulated by context, inferential reasoning, and communicative goals (Searle and Vanderveken, 1985).

Importantly, classification was guided primarily by the input/output mapping presupposed by the experimental design of the probe, rather than by the precise information stream that might en-

able task resolution. That is, instead of isolating the semantic nature of $\hat{z}$ (whether lexical, propositional, pragmatic, etc.), greater weight was given to the shared representational basis that supports the relevance of $\hat{z}$ in f. For instance, anaphoric dependencies can sometimes be resolved through syntactic constraints, but in other cases require pragmatic or contextual information. However, the phenomenon itself concerns how lexical items are mapped onto logical form via argument structure and variable binding. From this perspective, anaphora operates at the syntax–semantics interface, as it takes individual words as input and returns a structured semantic representation, which may remain underspecified with respect to reference and later be resolved by an assignment function drawing on context. This stance yields a more uniform representational picture of each semantic level, where evidence for or against the existence of abstract compositional structures gradually builds up throughout studies. In addition, while some studies frontally engaged with linguistic constructs and could be easily assigned to the relevant tier, others had looser theoretical ties to compositionality and had to be re-evaluated with regards to the composition function they were targeting. For instance, some datasets investigating intentions or overall sentiment behind a line of dialog (e.g. Eric et al., 2019; Coucke et al., 2018) were re-evaluated as targeting the illocutionary force of a proposition.

4 Results

The results are synthesized in Table 2, which organizes the findings by semantic tiers, primitives, composition functions, and performance levels. Because this paper synthesizes results across heterogeneous experimental designs, reported metrics are not strictly comparable and are interpreted as reflecting global trends across model types and datasets rather than fine-grained empirical differences. In this context, exact metric thresholds are less informative than the presence or absence of convergent patterns. Non-convergent findings are therefore of particular interest, as they indicate unstable representational evidence and cast doubt on the transferability of the formal semantic model assumed by the probe.

4.1 Tier-by-Tier Description

At the lexical tier, transformer models like BERT (Devlin et al., 2019) were capable of assigning an entity label to a given token more reliably than other model types. Lexical relations identification (e.g. hypernymy or hyponymy, as seen in WordNet: Fellbaum, 1998) received overall good performance, but with inconsistencies depending on the nature of the relation probed. Relation type between head and modifier in compound nouns was also reliably recoverable, with an accuracy across upper layers of about 80 percent. It is worth noticing that no studies targeted noun internal formal structure (see e.g. Pustejovsky, 1998), nor scrutinized event features and how they co-compose with sentence meaning (Dowty, 1979), leaving room for further exploration in this direction.

The syntax-semantics interface tier represents the core compositional structure of the hypothesized framework. Studies reported high performance on Part-of-Speech tagging, thematic role or semantic properties assignment (F1 score > 90), with no contrasting results over multiple datasets. Transformers notably outperformed uncontextualized baseline models in these cases. Syntactic constituency and tree structure were also recoverable to a relatively high degree of certainty (71 < F1 score < 83), despite the notion of syntactic tree depth appearing absent from the embeddings. Syntactic dependencies establishment, pivotal for functional application in formal semantics, was high in some studies, but constituents could not be tied to a specific in-sentence grammatical function (Alt et al., 2020). Findings about coreference and anaphora resolution were mixed and unreliable, with certain datasets appearing as approachable (OntoNotes: Weischedel et al., 2013, F1 score > 85), and others not (Winograd: Levesque et al., 2012, F1 score < 65). This suggests that syntactically ambiguous cases are not consistently distinguished in internal representations. When it comes to inflecting tense and morphosyntactic features (case, gender, etc.), representations were overall high, but was easily perturbed by task manipulation (Mikhailov et al., 2021). Moreover, no abstract notion of morphosyntax could be recovered across typologically diverse languages, suggesting that morphosyntactic regularities are encoded in a language-specific rather than abstract form (Choenni and Shutova, 2022). Negation did not appear robustly encoded in internal states of LLMs (Chen and Gao, 2022). No studies attempted to identify semantic type directly and many interface phenomena were not investigated either (e.g. modals, interrogatives, etc.)

In the propositional tier, the three studies re-

Table 2: Alignment between semantic tiers, their assumed primitives, and evidence provided by probing tasks across studies.

	Lexical Semantics	**Syntax–Semantics Interface**	**Propositional Semantics**	**Discourse Semantics / Pragmatics**
Primitives	Concepts	Arguments, Predicates, Morphemes, Thematic Roles	Propositions, Truth-values	Implicatures, Illocutionary Force, Informational Status
Composition functions	Lexical Relations, Co-composition	Predication, Semantic Translation	Logical Operations, Syllogisms	Inferences
Good performance	Entity labeling (Tenney et al.; Zhao et al.; Wallat et al.)	Semantic roles (Tenney et al.; Wang et al.), Predicate properties (Edge probing: Tenney et al.)	–	–
Moderate performance	Lexical relations (Aspillaga et al.; Lin and Ng; Chen and Gao) Head Modifier Relation in Compounds (Kendrick et al.)	Tense and inflection (Jawahar et al.; Mikhailov et al.), Syntactic constituents (Tenney et al.; Jawahar et al.; Arps et al.; Starace et al.)	–	Figures of speech (Aghazadeh et al.;Schneidermann et al.; Klubička et al.; Dankers et al.), Informational status (Li et al.; Ju et al.)
Fair performance	–	Abstract morphosyntactic organization (Choenni and Shutova), Grammatical functions (Alt et al.), Coreference and anaphora (Tenney et al.; Saleh et al.; Wallat et al.; Chen and Gao), Predicate properties (Vertex probing: Chen and Gao)	Logical operators (Lyu et al.; Ryb et al.; Traylor et al.)	Speech acts (Saleh et al.; Chen and Gao), Discourse organization (Jawahar et al.; Saleh et al.)
Poor performance	–	Event apprehension (Wang et al.), Negation (Chen and Gao)	Monotonicity (Chen and Gao), Semantic odd-man-out (Jawahar et al.)	Discourse macrostructure (Saleh et al.)

Performance levels: ***Good****: F1* $\geq$ *90, no contrasting results;* ***Moderate****: 75* $\leq$ *F1* $<$ *90, some contrasting results;* ***Fair****: 65* $\leq$ *F1* $<$ *75, contrasting results;* ***Poor****: F1* $<$ *65. Dashes indicate the absence of dedicated probing tasks or consistently interpretable results.*

ported drastically different results. While Ryb et al. (2022) reliably recovered conjunction, condition or quantification, Traylor et al. (2021); Lyu et al. (2022) could not tease apart boolean connectives in LLMs embeddings. A tantamount task in this tier is the semantic odd-man-out first introduced by (Conneau et al., 2018). It varies a constant in a proposition by substituting a token by an equally probable one in the overall probability distribution of the model. In Jawahar et al. (2019), BERT was easily duped by such manipulation (F1 < 65). Similarly, LLMs did not reliably display sensitivity to upward or downward monotonic environment in a logical form (Chen and Gao, 2022). While many natural language inference datasets now focus on reasoning, relatively few studies targeted this tier.

Recasting propositions into a communicative context imposes discursive and meta-representational requirements that were only partially recoverable. For instance, diverse figures of speech were generally identified as such, but models systematically failed to identify the illocutionary force of a speech act in dialog (Chen and Gao, 2022; Saleh et al., 2020). Models were also dynamically updating the informational status of the entities featured in discourse, but could not internally isolate a stable representation of the discourse context responsible for such updates (Li et al., 2021). Moreover, traces of dialogs and discourse structural representations could not be recovered, as internal states of models only displayed weak encoding of the functional role of a sentence in a dialog. Abstraction of a sequence of sentences into a larger thematic structure was also absent from transformer embeddings (Saleh et al., 2020).

4.2 Overall Trends

Overall, a gradual drop of accuracy can be observed as probes target increasingly abstract linguistic constructs, particularly those involving larger constituents and higher-order composition, like propositions or discourse movements. Most notably, isolating the influence of a single word or a group of words on the syntactico-logical structure of a proposition remains difficult for LLMs (Arps et al., 2022; Jawahar et al., 2019; Chen and Gao, 2022; Li et al., 2021). By contrast, representations of concepts and basic relations appear to be consistently well encoded (Tenney et al., 2019; Jawahar et al., 2019). Compared to earlier neural network architectures, transformer-based models seem to represent relational information more robustly, as demonstrated by the higher prevalence of lexical and dependency-based information (Tenney et al., 2019) at the predicate level. On the other hand, a simple concatenation of static vectors can encode pragmatic features to a degree comparable to, if not higher than, transformer models (Saleh et al.,

2020). This suggests that the purported computational primitives of formal models fail to give a comprehensive account of the seemingly compositional behavior of LLMs: while probing studies do not contradict the presence of word-level semantic information in LLMs, the evidence for propositional or discourse-level representations remains inconclusive at best. These results therefore provide limited support for the existence of higher-order meaning composition like pragmatic inferencing or logical reasoning on propositions.

5 Discussion

5.1 Beyond Compositionality

Early distributional models had already proven successful in handling semantic composition tasks in a syntax-agnostic fashion through rudimentary operations on static word embeddings. Distributional models have shown that simple or weighted addition can approximate human intuition when compounding nouns (Lazaridou et al., 2013), model noun-adjective combination (Baroni and Zamparelli, 2010), and predict verbs' thematic structure (Lenci, 2011). While being outperformed in syntax-related tasks, a sum of Word2Vec vectors (Mikolov et al., 2013) also encodes a fair range of linguistic properties, and overall achieves comparable results to BERT on sentence-level probing tasks (Miaschi and Dell'Orletta, 2020; Lenci et al., 2022). Moreover, in a review of state-of-the-art sentence encoders, Li et al. (2023) have shown that static-vector systems fine-tuned on natural language inference datasets may even outperform transformer-based models on inter-sentence downstream and linguistic classification tasks. A recent study also showed that transformers might still be performing vector addition and other simple arithmetic on vectors to tackle tense inflection and lexical association tasks (Merullo et al., 2024). Taken together, the success of simple composition methods and non-contextualized-vector-based encoders should illustrate that fruitful composition may, in large part, originate from lexically accurate word embeddings rather than linguistically motivated composition methods (Boleda, 2020).

This raises questions about whether mechanistic accounts of LLMs should prioritize identifying traditional compositional primitives. POS tagging and other syntactic tasks remain the most probed representation in the literature (Waldis et al., 2024). The limited support for a propositional level of representations indicates that predicates may be lexically linked together locally, but would not be abstracted into a logical form according to a mapping from syntactic to semantic types. If constituents do not combine according to a part-whole structure, this would undermine the centrality of syntax and interrogate the relevance of identifying syntactic representations. In other words, a strict homeomorphism between structure and semantic interpretation readily loses its explanatory power. If this holds true, identifying syntactic structure in neural network embeddings might only be an approximate fitting of some idiosyncratic representations by the logistic regressor (see Michael et al., 2020; Buder-Gröndahl, 2024), and may not carry the same functional indicativity assumed in type-theoretic semantics. The formal framework considered therefore appears insufficient, on its own, to fully characterize the compositional mechanisms evidenced by probing results. This is consistent with LLMs' weaker performance on anaphora and co-reference probing tasks, which require more complex logical derivation than semantic role assignment. Similarly, at the discourse level, the lack of illocutionary force may also indicate that coherence and discourse flow originate from complex inter-sentence predicate-arguments arrangements and lexical regularities, and not from discursive movements based on intentions. Whether more recent and larger language models than the one surveyed in this paper – which mostly focus on BERT-style models Devlin et al., 2019; Liu et al., 2019; He et al., 2021 – could display more stable representational evidence of these higher-level structures warrant further research.

The investigation of LLMs' internal representations remains explanatorily underdetermined. As this review shows, probing studies consistently recover lexical concepts, semantic roles, and aspects of syntactic structure, while providing more limited evidence for propositional-level abstraction and illocutionary force. At the same time, no clear tendency emerge across benchmarks that would support a shift toward explicitly compositional representations. Taken together, these findings suggest that composition-like behavior in LLMs may be supported primarily by lexical and locally structure-sensitive regularities, rather than by fully propositional representations. A predominantly lexicalized, distributionally grounded form of composition therefore remains a plausible explanatory hypothesis, given the current probing evidence.

5.2 Role of Formal Theory and Future Directions

While reviewing the implications of LLMs' compositional representation for semantic theory is beyond the scope of this paper, its goal is therefore not to adjudicate between theories of compositionality, but to evaluate how different semantic commitments affect the interpretation of probing results. Reworking the primitives and composition functions posited might drastically change the picture drawn over this review, and other traditions of semantics might make a different exegesis of the results shared here. Formal semantics is not the only game in town for a science of meaning, and concurrent approaches have long-lastingly criticized its take on compositionality (Goldberg, 2015). Usage-based approaches, whose sentential and discourse primitives consist of form-to-meaning constructions, may represent a prima facie better fit to unravel LLMs' internal composition mechanisms (Baggio, 2021; Goldberg, 2024; Rambelli et al., 2019). Its application in explainability studies is an active line of research (Tayyar Madabushi et al., 2020; Pannitto and Herbelot, 2023), with initial evidence showing high degree of transferability between LLMs' knowledge and usage-based constructions (Chronis et al., 2023; Li et al., 2022). Two studies screened in this review directly applied this perspective to the identification of construction representation (Scivetti and Schneider, 2025; Ramezani et al., 2025). However, this paper argues that the limited probing evidence for proposition-level representations poses a challenge for attempts to directly map human semantic theories onto transformer architectures. Whether this reflects architectural constraints or limitations of current probing methodologies remains an open question.

These considerations suggest that mechanistic accounts of compositionality in LLMs may benefit from first characterizing the models' own representational affordances. This motivates a shift in the study of LLMs' underlying representations towards a more bottom-up and model-centered approach. A promising step in this direction is sparse autoencoder probing (Cunningham et al., 2023), which attempts to isolate disentangled semantic features in model representations without assuming human-like compositional primitives. Such techniques help characterize the internal structure of LLMs not as a mirror of human semantic competence, but as a functionally emergent system shaped by large-scale linguistic co-occurrence. These uncovered semantic representations should form the theoretical basis to formally sketch a computational and algorithmic theory of LLMs' compositional pathway. Formal approaches then come in handy to properly characterize the properties required by the composition functions derived from models' observed behaviors (Vilas et al., 2024). These higher-level theories should then be tested empirically through causal intervention and ablation tests to effectively assess the functional relevance of the identified representations (Geiger et al., 2025). In this context, traditional probing still has a card to play. It should serve as a post-hoc sanity check to ensure recoverability of posited representations in the network's intermediate layers, whereby failure to identify the assumed representations should immediately be a sign for the researcher to return to the theoretical work board.

6 Conclusion

The probing paradigm aims to identify which linguistic properties are recoverable from the internal states of LLMs. Albeit largely adopted in the field, it faces limitations when used to draw conclusions about compositional mechanisms. In particular, issues of functional indicativity and one-to-many mappings between internal states and computational explanations complicate the inference from property recoverability to function.

This review has argued that such limitations do not make probing uninformative, but rather redefine its role in explainability. Considered individually, probing studies provide partial evidence about recoverable properties. Considered collectively and interpreted through the primitives they (covertly) endorse, probing functions as a diagnostic tool for evaluating the transferability of specific compositional hypotheses. Across the reviewed literature, evidence supports lexical and local structural regularities, while providing little support for robust propositional or discourse-level structured representations.

Accordingly, the contribution of this review is not to adjudicate whether LLMs instantiate formal representations, but to clarify how the computational primitives presupposed by probing tasks constrain what can be concluded about their compositional pathways. When anchored to explicit theories of meaning, probing remains a valuable

tool to systematically eliminate implausible models of compositionality, even if it cannot by itself establish the precise functions implemented by language models.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Probing Linguistic Features of Sentence-Level Representations in Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1534–1545, Online. Association for Computational Linguistics.

David Arps, Younes Samih, Laura Kallmeyer, and Hassan Sajjad. 2022. Probing for Constituency Structure in Neural Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6738–6757, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carlos Aspillaga, Marcelo Mendoza, and Alvaro Soto. 2021. Inspecting the concept knowledge graph encoded by modern language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2984–3000, Online. Association for Computational Linguistics.

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

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1183–1193.

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Gemma Boleda. 2020. Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, 6(1):213–234. ArXiv:1905.01896 [cs].

Tommi Buder-Gröndahl. 2023. The ambiguity of bertology: what do large language models represent? *Synthese*, 203(1):15.

Tommi Buder-Gröndahl. 2024. What does parameter-free probing really uncover? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 327–336, Bangkok, Thailand. Association for Computational Linguistics.

Tyler A Chang and Benjamin K Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.

Zeming Chen and Qiyue Gao. 2022. Probing Linguistic Information for Logical Inference in Pre-trained Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10509–10517.

Rochelle Choenni and Ekaterina Shutova. 2022. Investigating Language Relationships in Multilingual Sentence Encoders Through the Lens of Linguistic Typology. *Computational Linguistics*, 48(3):635–672.

Gabriella Chronis, Kyle Mahowald, and Katrin Erk. 2023. A method for studying semantic construal in grammatical constructions with interpretable contextual embedding spaces. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 242–261, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, and 1 others. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David R Dowty. 1979. Word meaning and montague grammar. *Studies in Linguistics and Philosophy*.

Frances Egan. 2018. Function-theoretic explanation and the search for neural mechanisms. In *Explanation and Integration in Mind and Brain Science*, pages 145–163. Oxford University Press.

Yanai Elazar, 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.

David Embick and David Poeppel. 2015. Towards a computational (ist) neurobiology of language: correlational, integrated and explanatory neurolinguistics. *Language, cognition and neuroscience*, 30(4):357–366.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Daniel Furrer, Marc van Zee, Nathan Scales, and Nathanael Schärli. 2021. Compositional Generalization in Semantic Parsing: Pre-training vs. Specialized Architectures. *arXiv preprint*. ArXiv:2007.08970 [cs].

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and 1 others. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64.

Adele E Goldberg. 2015. Compositionality. In *The Routledge handbook of semantics*, pages 419–433. Routledge.

Adele E Goldberg. 2024. Usage-based constructionist approaches and large language models. *Constructions and Frames*, 16(2):220–254.

Robert F Hadley. 1994. Systematicity in connectionist language learning. *Mind & Language*, 9(3):247–272.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Irene Heim and Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Blackwell, Malden, MA.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'Diagnostic Classifiers' Reveal How Recurrent and Recursive Neural Networks Process Hierarchical Structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8235–8246, Torino, Italia. ELRA and ICCL.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics.

Saffron Kendrick, Mark Ormerod, Hui Wang, and Barry Devereux. 2025. Investigating noun-noun compound relation representations in autoregressive large language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 253–263, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Filip Klubička, Vasudevan Nedumpozhimana, and John Kelleher. 2023. Idioms, Probing and Dangerous Things: Towards Structural Probing for Idiomaticity in Vector Space. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 45–57, Dubrovnik, Croatia. Association for Computational Linguistics.

John W Krakauer, Asif A Ghazanfar, Alex Gomez-Marin, Malcolm A MacIver, and David Poeppel. 2017. Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, 93(3):480–490.

Ilia Kuznetsov and Iryna Gurevych. 2020. A matter of framing: The impact of linguistic formalism on probing results. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 171–182, Online. Association for Computational Linguistics.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1517–1526, Sofia, Bulgaria. Association for Computational Linguistics.

Alessandro Lenci. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–66, Portland, Oregon, USA. Association for Computational Linguistics.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Language resources and evaluation*, 56(4):1269–1313.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. *KR*, 2012:13th.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit Representations of Meaning 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 1813–1827, Online. Association for Computational Linguistics.

Ruiqi Li, Xiang Zhao, and Marie-Francine Moens. 2023. A Brief Overview of Universal Sentence Representation Methods: A Linguistic View. *ACM Computing Surveys*, 55(3):1–42.

Ruixi Lin and Hwee Tou Ng. 2022. Does BERT Know that the IS-A Relation Is Transitive? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 94–99, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Qing Lyu, Zheng Hua, Daoxin Li, Li Zhang, Marianna Apidianaki, and Chris Callison-Burch. 2022. Is "My Favorite New Movie" My Favorite Movie? Probing the Understanding of Recursive Noun Phrases. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5286–5302, Seattle, United States. Association for Computational Linguistics.

David Marr. 1982. *Vision : a computational investigation into the human representation and processing of visual information*. W.H. Freeman.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Language models implement simple Word2Vec-style vector arithmetic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5030–5047, Mexico City, Mexico. Association for Computational Linguistics.

Eleni Metheniti, Tim Van De Cruys, and Nabil Hathout. 2022. About time: Do transformers learn temporal verbal aspect? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 88–101, Dublin, Ireland. Association for Computational Linguistics.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic Profiling of a Neural Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alessio Miaschi and Felice Dell'Orletta. 2020. Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online. Association for Computational Linguistics.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6792–6812, Online. Association for Computational Linguistics.

Vladislav Mikhailov, Oleg Serikov, and Ekaterina Artemova. 2021. Morph Call: Probing Morphosyntactic Content of Multilingual Transformers. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 97–121, Online. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Ludovica Pannitto and Aurélie Herbelot. 2023. CALaMo: a constructionist assessment of language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 21–30, Washington, D.C. Association for Computational Linguistics.

Barbara Partee and 1 others. 1995. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020a. Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3138–3153, Online. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020b. 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.

David Poeppel and David Embick. 2017. Defining the relation between linguistics and neuroscience. In *Twenty-first century psycholinguistics*, pages 103–118. Routledge.

James Pustejovsky. 1998. *The generative lexicon*. MIT press.

Giulia Rambelli, Emmanuele Chersoni, Philippe Blache, Chu-Ren Huang, and Alessandro Lenci. 2019. Distributional semantics meets construction grammar. towards a unified usage-based model of grammar and meaning. In *First international workshop on designing meaning representations (DMR 2019)*.

Pegah Ramezani, Achim Schilling, and Patrick Krauss. 2025. Analysis of argument structure constructions in the large language model bert. *Frontiers in Artificial Intelligence*, 8:1477246.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Samuel Ryb, Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2022. AnaLog: Testing Analytical and Deductive Logic Learnability in Language Models. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 55–68, Seattle, Washington. Association for Computational Linguistics.

Abdelrhman Saleh, Tovly Deutsch, Stephen Casper, Yonatan Belinkov, and Stuart Shieber. 2020. Probing Neural Dialog Models for Conversational Understanding. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 132–143, Online. Association for Computational Linguistics.

Nina Schneidermann, Daniel Hershcovich, and Bolette Pedersen. 2023. Probing for Hyperbole in Pre-Trained Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 200–211, Toronto, Canada. Association for Computational Linguistics.

Wesley Scivetti and Nathan Schneider. 2025. Construction identification and disambiguation using BERT: A case study of NPN. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 365–376, Vienna, Austria. Association for Computational Linguistics.

John R Searle and Daniel Vanderveken. 1985. *Foundations of illocutionary logic*. CUP Archive.

Giulio Starace, Konstantinos Papakostas, Rochelle Choenni, Apostolos Panagiotopoulos, Matteo Rosati, Alina Leidinger, 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, Singapore. Association for Computational Linguistics.

Alfred Tarski. 1944. The semantic conception of truth: and the foundations of semantics. *Philosophy and phenomenological research*, 4(3):341–376.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. *arXiv preprint*. Version Number: 1.

Aaron Traylor, Roman Feiman, and Ellie Pavlick. 2021. AND does not mean OR: Using Formal Languages to Study Language Models' Representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 158–167, Online. Association for Computational Linguistics.

Ashish Vaswani, 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.

Martina G. Vilas, Federico Adolfi, David Poeppel, and Gemma Roig. 2024. Position: An inner interpretability framework for AI inspired by lessons from cognitive neuroscience. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 49506–49522. PMLR.

Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length. *arXiv preprint*. ArXiv:2003.12298 [cs].

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes a benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.

Jonas Wallat, Fabian Beringer, Abhijit Anand, and Avishek Anand. 2023. Probing BERT for Ranking Abilities. In *Advances in Information Retrieval*, pages 255–273, Cham. Springer Nature Switzerland.

Barry Wang, Xinya Du, and Claire Cardie. 2023. Probing Representations for Document-level Event Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12675–12683, Singapore. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, and 1 others. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23(170):20.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

Mengjie Zhao, Philipp Dufter, Yadollah Yaghoobzadeh, and Hinrich Schütze. 2020. Quantifying the Contextualization of Word Representations with Semantic Class Probing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1219–1234, Online. Association for Computational Linguistics.