Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Three Lenses on Linguistic Knowledge in Pretrained Transformers: A Phenomenon-Centered Review

Yizhe Wang
University of Science and Technology of China
Zhenhua Ling
zhling@ustc.edu.cn

University of Science and Technology of China

Research Article

Keywords: pretrained language models, linguistic knowledge, mechanistic interpretability, probing, compositional generalization

Posted Date: September 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10776566/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Three Lenses on Linguistic Knowledge in Pretrained Transformers: A Phenomenon-Centered Review

Yizhe Wang[1] and Zhenhua Ling[1*]

[1] Interdisciplinary Research Center for Linguistic Sciences; National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230026, Anhui, China .

*Corresponding author(s). E-mail(s): zhling@ustc.edu.cn;
Contributing authors: wangyz@ustc.edu.cn;

Abstract

Pretrained Transformer language models display substantial linguistic abilities, but evidence about their linguistic knowledge remains divided across representational, behavioral, and mechanistic traditions. This review compares these evidential lenses across three focal cases in English pretrained Transformers: subject–verb agreement, negation, and compositional generalization. It synthesizes a corpus of studies coded for target variable, model regime, and behavioral contrast. Three distinct configurations emerge. Agreement provides the clearest convergence. In selected matched paradigms, grammatical number can be decoded, manipulated, and linked to verb-form behavior. Negation shows a detection–application mismatch. Marker and scope information are recoverable from internal states, but model predictions do not reliably reflect the meaning of the negated proposition. Compositional generalization shows fragmentation. Positive evidence appears at each evidential level, but the studies target different forms of composition and do not support a single integrated account. We interpret these configurations through two properties: whether a relevant structure can be isolated in internal states and whether manipulating it changes matched behavior. This accessibility-and-use diagnostic identifies what current evidence establishes, where its limits lie, and which matched cross-level tests are needed next.

Keywords: pretrained language models, linguistic knowledge, mechanistic interpretability, probing, compositional generalization

1 Introduction

What do large language models know about language? The question is easy to ask and hard to answer. These models produce fluent and largely grammatical text, often with coherent meaning (Brown et al. 2020), and it is tempting to infer that they have learned how language works. But output is not a transparent guide to the knowledge behind it. Models can produce fluent text while still failing on basic linguistic phenomena, such as the effect of negation on sentence meaning (Ettinger 2020). A correct response can also be produced for reasons unrelated to what a test is meant to probe (McCoy et al. 2019).

Rather than treating behavior as decisive evidence, the field has approached this question through three distinct lines of inquiry. What linguistic information is encoded in a model's internal states? What generalizations are expressed in its behavior? And what computation produces that behavior? Each question has grown into its own research tradition. We treat these traditions as three evidential lenses on linguistic knowledge, respectively representational, behavioral, and mechanistic Each lens has a characteristic limitation. Representational studies ask what linguistic information can be recovered from a model's internal states, but recoverability does not show that a feature is used. Behavioral studies ask whether the model's outputs honor grammatical and semantic generalizations, but a correct output does not reveal the computation behind it. Mechanistic studies ask which components causally drive behavior, but their findings are often task-local and sensitive to the chosen intervention. These limitations are well documented, and existing work has usually responded by refining individual methods. This review asks a different question. Under what conditions do the three lenses support the same conclusion about a linguistic phenomenon?

This gap is also evident in the survey literature, which is typically organized by method rather than by phenomenon. Belinkov (2022) examines probing classifiers, the main representational tool. Chang and Bergen (2024) survey behavioral evidence across a wide range of English phenomena. Broader treatments of mechanistic interpretability review causal methods for analyzing neural networks (Räuker et al. 2023), and wider syntheses follow the same organization, from a primer on BERT analyses (Rogers et al. 2020) to overviews of Transformer internals (Ferrando et al. 2024). The closest prior review is López-Otal et al. (2025), which shares our restriction to base pretrained models and synthesizes internal analyses of linguistic knowledge across many languages, but excludes behavioral evidence by design. Its question is therefore one of location, asking where and in what form linguistic information is present inside the network.

The present review differs from these works on three points. First, it asks whether representational, behavioral, and mechanistic evidence, when all three exist for one phenomenon, support the same conclusion about it. Second, answering that question requires an evidence base no single-level survey admits, since it must join internal analyses with behavioral contrasts and with causal interventions tied to those contrasts. Third, its results describe the relation among the levels rather than any one level, and this relation differs in kind across phenomena. Table 1 summarizes the comparison. Because none of the cited surveys compares the three levels on a single phenomenon,

Survey	Organizing principle	Levels covered	Cross-level comparison per phenomenon
Rogers et al. (2020)	Model (analyses of BERT)	Representational, behavioral	No
Belinkov (2022)	Method (probing classifiers)	Representational	No
Chang and Bergen (2024)	Method (behavioral evaluation)	Behavioral	No
Räuker et al. (2023)	Method (causal interpretability)	Mechanistic	No
Ferrando et al. (2024)	Method (Transformer internals)	Mechanistic	No
López-Otal et al. (2025)	Internal analyses, organized by linguistic level	Representational, partly mechanistic	No
This review	Phenomenon	All three	Yes

Table 1 Positioning of the present review against prior surveys. Existing surveys provide deep coverage within one methodological tradition. The present review takes the phenomenon as the unit of analysis and compares the three evidential levels on each.

the finding that they relate differently for agreement, negation, and compositional generalization cannot be assembled from prior work.

We pursue this question through three phenomena, subject–verb agreement, negation, and compositional generalization. They were chosen, before coding began, for two reasons. First, each has been central to debates about what language models know. Agreement anchored the earliest targeted evaluations of neural language models (Linzen et al. 2016). Negation became a standard test of whether fluent models compute the meaning they appear to express (Ettinger 2020; Kassner and Schütze 2020). Compositional generalization carries the long-standing question of whether neural networks generalize systematically (Fodor and Pylyshyn 1988; Lake and Baroni 2018). Second, together they span three increasingly demanding forms of structured linguistic use. Subject–verb agreement is a morphosyntactic dependency in which a discrete grammatical feature must be identified and carried to a verb-form prediction. Negation is a semantic operation in which a marker must be detected, its scope fixed, and its truth-conditional consequence applied. Compositional generalization is the open-ended case, in which familiar parts must be recombined productively across new expressions, tasks, or intermediate computations.

The comparison therefore focuses on one major dimension of linguistic knowledge, the ability to support structured recombination. This narrows the broad question of linguistic knowledge without reducing language to recombination alone. Together, these focal cases vary in recombinatory demand, linguistic type, and operational precision. Agreement has a well-defined target variable, grammatical number (Corbett 2006). Negation separates marker, scope, and interpreted consequence (Horn 1989). Compositional generalization raises the broader problems of systematicity and productivity, for which there is no single operational target (Partee et al. 1984; Fodor and Pylyshyn 1988). It is accordingly better read as a family of capacities than as a single phenomenon. This asymmetry motivates the comparison. It lets the review ask whether the three lenses continue to align as the linguistic target changes, becoming

broader, less uniformly operationalized, and more demanding in the recombination it requires.

The review makes three contributions.

- **A phenomenon-centered cross-level comparison.** Existing surveys have usually been organized by method, asking what probes reveal, what behavioral benchmarks show, or what causal interventions can identify. We instead ask whether these forms of evidence address the same linguistic target. This comparison shows that the relation among evidential levels differs in kind across phenomena. The relation is convergence for agreement, a detection–application mismatch for negation, and fragmentation for compositional generalization.
- **A phenomenon-specific map of the encoding–use problem.** Probing is the standard tool for asking what a model encodes, and a recurring concern is that what a probe can recover need not be what the model uses. We show that this concern does not take one uniform form. For agreement, encoding and use are matched on grammatical number. For negation, the relation between marker detection and proposition-level interpretation remains unresolved. For compositional generalization, no common variable has been isolated, so the question cannot yet be tested. The review maps what each evidential configuration establishes and the questions that remain open. It shows that the limits of probing are phenomenon-specific rather than general.
- **Accessibility and use as a diagnostic hypothesis.** The observed pattern motivates the hypothesis that mechanistic tractability may depend not only on available tools, but also on whether the target structure is isolable in internal states and causally linked to matched behavior. The framework therefore yields concrete tests for future work, in which one variable is decoded, manipulated, and behaviorally tested within a single design.

The paper proceeds as follows. Section 2 sets out the scope of the review, the criteria defining the evidence corpus, and the procedure used to collect, assess, and code the literature, and presents the analytical framework. Section 3 introduces the three methodological traditions in more detail. Sections 4, 5, and 6 present multilevel analyses of subject–verb agreement, negation, and compositional generalization. Section 7 draws the cross-cutting patterns together, states the accessibility and use hypothesis alongside its rivals, and sets out implications for evaluation practice and for human–model comparison. Section 8 concludes.

2 Survey Methodology and Analytical Framework

2.1 Scope

The review concentrates on English pretrained Transformer models. Both masked and autoregressive models are covered, and they are distinguished where their behavior differs. The primary object of this review is linguistic knowledge in base pretrained models. Because fine-tuning can alter representations acquired during pretraining (Merchant et al. 2020), fine-tuned and instruction-tuned models are not themselves the primary target of the synthesis. Studies of such systems enter the corpus in two

roles. They probe linguistic knowledge directly, or they establish a boundary condition by showing how behavior changes when training or prompting departs from plain pretraining. Each evidence-map entry records the model regime and any fine-tuning, instruction-tuning, or prompting intervention. Recent preprints are treated as provisional where the fast-moving interpretability literature has not yet reached archival form. Work outside this scope is cited as supporting literature. The conditions for such citation are set out in Section 2.3.

2.2 Search and Assessment

Literature was identified through an iterative, phenomenon-led process. Google Scholar served for broad discovery. The ACL Anthology served for targeted retrieval from core computational linguistics venues. Backward and forward citation chasing started from benchmark papers, methodological surveys, and directly relevant articles. Search terms combined three groups. Model labels included BERT, GPT, Transformer, pretrained language model, and large language model. Method terms included probing, representational analysis, behavioral evaluation, minimal pairs, causal analysis, and mechanistic interpretability. Phenomenon terms included subject–verb agreement, grammatical number, negation, polarity, NPI licensing, compositionality, and systematic generalization.

Assessment was continuous with the search rather than a separately staged screening. Each candidate was assessed at the point of discovery. The assessment relied on the abstract and introduction. Further sections were consulted where the contribution remained unclear. A small number of ambiguous candidates were resolved by discussion between the authors. Candidates judged potentially relevant to the evidence base entered a tracked pool. Queries were refined and citation chains followed as the pool grew. Quantitative reporting covers the tracked pool of candidate evidence studies (Figure 1).

2.3 Corpus Criteria and Supporting Literature

The criteria in this subsection govern membership in the evidence corpus. The corpus is the set of studies on which the cross-level comparison rests. The criteria do not govern the bibliography as a whole. The bibliography also contains supporting literature, described at the end of this subsection.

A study qualified for the evidence corpus when it met all of the following criteria.

- It studies pretrained Transformer language models.
- It provides representational, behavioral, or mechanistic evidence bearing on one of the three focal phenomena.
- It states what it measures, whether a target variable, a behavioral contrast, or an intervention outcome. The statement must be explicit enough to be coded along the dimensions of Section 2.5.

Qualification did not guarantee inclusion. Several qualifying candidates often reported the same kind of evidence on the same target and evidential level. In such cases, representative studies were selected for citation. Among candidates reporting

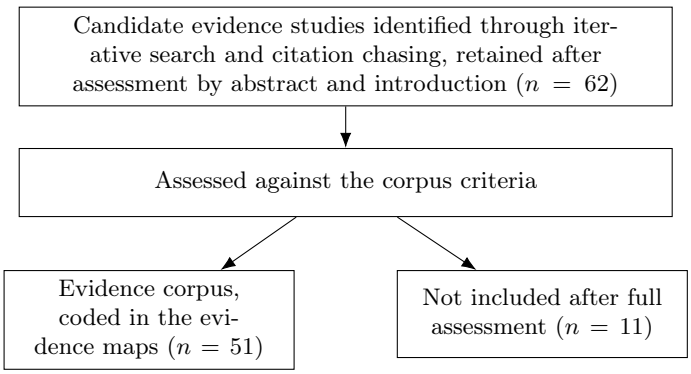

Fig. 1 Selection of the evidence corpus. Assessment was continuous with the iterative search. Records viewed and set aside before collection were not logged, so reporting covers the tracked candidate evidence studies. Each candidate was assessed against the corpus criteria and either included in the corpus or not included, with the reason recorded. Supporting literature is cited in the ordinary way and is outside this count.

the same kind of evidence, the selection favored three properties. A study could link its target variable to matched behavior. It could cover a model regime otherwise unrepresented. Or it could establish a boundary condition other studies do not.

Beyond the corpus, the review cites four kinds of supporting literature. Linguistic and cognitive theory grounds the definition of each phenomenon and the human–model comparison. Foundational pre-Transformer studies are cited where they establish a benchmark or analysis paradigm. The model regime is named at each such citation. Methodological work that develops or critiques analysis techniques without targeting the focal phenomena informs the assessment of what each lens can establish. Prior surveys are cited for positioning. Supporting literature is cited in the ordinary way. It is not coded in the evidence maps and is not counted in Figure 1 which reports the evidence-candidate track only.

2.4 Selection Outcome

Figure 1 summarizes the outcome of this process for the evidence-candidate track. The tracked pool contains 62 candidate studies. Each was assessed against the corpus criteria. The evidence corpus comprises 51 studies. The remaining 11 candidates were not included. Their evidence was covered by representatively cited studies, or the scope rules applied on closer reading. The procedure is therefore explicit and criterion-based in its assessment. The synthesis built on the corpus is integrative and thesis-driven. The resulting corpus is a targeted basis for comparing how evidence accumulates across levels. It is not a complete bibliography of agreement, negation, or compositionality.

2.5 Coding and Analytical Framework

Each corpus study was coded along five dimensions. These are the evidential level it provides, the model regime it examines, the target variable it operationalizes, the behavioral contrast it measures, and whether it links the target variable to matched behavior within one design. The corpus comprises all studies meeting the criteria of

Lens	Agreement	Negation	Compositional generalization
Representational	Number linearly decodable and manipulable (Sect. 4.1)	Marker and partial scope recoverable, no proposition-level variable (Sect. 5.1)	Geometry tracks complexity, no validated compositional variable (Sect. 6.1)
Behavioral	Near ceiling on canonical items, graded degradation under attractors and low support (Sect. 4.2)	Marker registered, semantic consequence applied unreliably (Sect. 5.2)	Separable and scaffolded recombination succeed, unaided chaining inconsistent (Sect. 6.2)
Mechanistic	Causal circuits identified, implementation varies with architecture and scale (Sect. 4.3)	Task-specific localizations, no shared mechanism (Sect. 5.3)	Two-hop intermediates traceable, no general circuit (Sect. 6.3)
Configuration	Convergence	Detection–application mismatch	Fragmentation

Table 2 Analytical framework and summary of findings. Rows are the three lenses and columns are the three phenomena. Each cell compresses the evidence reviewed in the indicated section, and the final row names the evidential configuration that the three levels jointly yield for each phenomenon.

Section 2.3 that are cited as evidence in Sections 4 through 6. The resulting study-level evidence maps appear as Tables 4, 5, and 6. They are placed at the end of those sections and display representative entries per level.

Table 2 presents the analytical framework of the review together with a summary of its findings. The rows are the three lenses and the columns are the three phenomena. Each cell states what the corresponding body of evidence establishes. The final row names the evidential configuration that emerges when the three cells of a column are read together. The remainder of the paper substantiates this table.

3 Three Methodological Traditions

The three traditions introduced in Section 1 provide complementary forms of evidence, each addressing limitations of the others. This section considers each tradition in turn. It outlines the principal methods and identifies the evidential limitation that motivates comparison with the other lenses. For this review, we classify each study by the type of evidence it provides rather than by its primary aim.

3.1 Representational Evidence: What is Encoded?

Representational studies examine the internal states of language models and ask whether specific linguistic features can be recovered from them (Conneau et al. 2018; Belinkov and Glass 2019; Rogers et al. 2020). The most common tool is the probing classifier. A supervised model is trained to predict linguistic annotations, such as part-of-speech tags, dependency relations, or parse-tree depth, from frozen contextual embeddings (Hewitt and Manning 2019; Tenney et al. 2019b,a; Liu et al. 2019). Recent work extends this paradigm to larger autoregressive models and to a wider range of syntactic phenomena (Simon et al. 2025). Complementary evidence comes

from geometric analyses (Coenen et al. 2019) and representational similarity analysis (Chrupała and Alishahi 2019). These methods test whether linguistic distinctions form separable regions, salient directions, or hierarchical structures in embedding space.

A long-standing debate concerns whether high probing accuracy reflects the model's own organization or the probe's capacity to learn the target task on its own (Pimentel et al. 2020; Belinkov 2022). The field has developed several responses. Hewitt and Liang (2019), for instance, introduce control tasks and the selectivity metric. Selectivity compares a probe's accuracy on the linguistic task with its accuracy on a random control, isolating the contribution of the representation from that of the probe. Voita and Titov (2020) reframe probe quality in terms of minimum description length, measuring how compactly the labels can be extracted rather than how accurately. This framing yields more stable comparisons across probes and representations. The debate has not settled on a single protocol, but its collective lesson is clear. Probe accuracy is not a transparent index of what a representation contains, since its interpretation depends partly on the probe.

A further line of work tests representational claims by intervention rather than by decoding alone. Amnesic probing removes a target feature from the representation and measures the effect on model predictions, using behavioral change as evidence relevant to functional use (Elazar et al. 2021). The removal is typically achieved by iterative nullspace projection, which repeatedly trains linear classifiers and projects out the directions they rely on (Ravfogel et al. 2020). These methods move probing closer to a causal test, but they do not eliminate the recoverability–use problem. Removing a linearly encoded feature can leave behavior unchanged when the information remains available in a form the projection does not reach. Conversely, any behavioral change must be tied to the intended linguistic variable rather than to unintended disruption caused by the projection.

The broader inferential limitation is therefore unchanged. Probes and geometric analyses can show that linguistic information is recoverable, but recoverability is not the same as functional use. A feature may be present in a representation without participating in the computation that produces the model's output. Recent work narrows the distance between probed information and behavior. He et al. (2024), for example, decode grammaticality labels for BLiMP minimal pairs from intermediate representations, linking a probed signal to a behavioral contrast. Evidence of this kind remains correlational. It shows that recoverable information covaries with behavior, not that the model uses it. Establishing use requires showing that manipulating the feature changes behavior in a predictable way. The causal methods described in Section 3.3 address this requirement, provided that a well-characterized behavioral contrast is available. Behavioral methods, to which we turn next, provide that endpoint.

3.2 Behavioral Evidence: What is Expressed?

Behavioral studies set the model's internal states aside and ask what it expresses in its outputs. Two measurement paradigms dominate. The first is direct probability measurement over minimal pairs, as in BLiMP (Warstadt et al. 2020) and SyntaxGym (Gauthier et al. 2020). These benchmarks present sentence pairs that differ in

a controlled linguistic contrast and assess whether the model assigns higher probability to the well-formed variant. The second is metalinguistic prompting, in which the model is asked for an explicit judgment, for example whether a sentence is grammatical (Behzad et al. 2023; Beguš et al. 2025). Hu and Levy (2023) compare the two and find that metalinguistic prompting yields weaker and less consistent evidence of linguistic knowledge than direct probability measurement.

Minimal-pair evaluation has a clear lineage. Early targeted evaluations tested whether recurrent models predict the correctly inflected verb in the presence of intervening nouns (Linzen et al. 2016). Later work removed lexical confounds by using semantically empty sentences, showing that models can track structural number rather than collocational cues (Gulordava et al. 2018). These studies established the minimal-pair logic that BLiMP and SyntaxGym later scaled to many constructions. A complementary measure is surprisal, the negative log probability a model assigns to a word in context. Surprisal supports graded predictions about processing difficulty and connects model behavior to human reading times (Wilcox et al. 2021). Minimal-pair and surprisal measures are related, since both read a behavioral preference from the model's probability distribution rather than from an explicit judgment.

Behavioral methods reveal what a model expresses, but not how the expression is produced. A model may succeed by computing the intended structural relation. It may also rely on shortcuts that align with the test items, such as lexical co-occurrence or surface word-order regularities in the training data (McCoy et al. 2019). Behavioral scores are also sensitive to evaluation protocol, prompting strategy, sequence length, and item construction (Ide et al. 2025). Behavioral evidence is therefore necessary but not sufficient. It tells us what the model does, not what representation or computation produces it.

3.3 Mechanistic Evidence: What is Implemented?

Mechanistic interpretability asks which internal components of a Transformer make a causal contribution to a behavior, and how those components combine into a circuit (Olsson et al. 2022; Wang et al. 2023; Ferrando et al. 2024). Its central methods are interventional. Patching substitutes activations from one input into the forward pass of another and measures how much of a behavior is recovered or disrupted. It supports tests of both sufficiency and necessity. Ablation removes or zeroes specific components and measures the resulting change, which speaks mainly to necessity. Causal mediation analysis decomposes the effect of an input feature on an output into paths that pass through particular intermediate variables. By manipulating components rather than correlating them with outputs, these methods support claims that go beyond correlation. They can tie particular attention heads, neurons, layers, or residual-stream directions to particular behavioral effects (Feder et al. 2022; Goldowsky-Dill et al. 2023; Amini et al. 2023).

These interventions have been refined to localize effects more precisely. Path patching restricts the substitution to specific routes through the network, isolating the contribution of one component to another rather than to the output as a whole (Goldowsky-Dill et al. 2023). Attribution patching approximates the effect of patching every component using a small number of forward and backward passes, which makes

circuit search tractable at scale (Syed et al. 2024). Automated methods extend this approach to discover candidate circuits without hand-specified hypotheses (Conmy et al. 2023). A separate line of work intervenes to test use rather than localization. Adding a feature direction to the residual stream and measuring the change in output probes whether the model reads that direction during generation (Turner et al. 2024). These methods enlarge the toolkit. However, none eliminates the interpretive burden, since a localized component or an effective steering direction still has to be shown to correspond to the linguistic variable of interest.

Any intervention requires a choice of unit, but conventional units may not correspond to the relevant causal variable. Individual neurons are frequently polysemantic, responding to several unrelated features. This arises because a network can represent more features than it has dimensions by encoding them in superposition (Elhage et al. 2022). An intervention defined over neurons or whole layers may therefore act on a mixture of features rather than the one of interest. Sparse autoencoders and related dictionary-learning methods address this problem. They decompose activations into a larger set of sparsely active, more nearly monosemantic directions that can serve as units for probing and intervention (Huben et al. 2023; Zhang et al. 2025). Such methods sharpen the representational basis for causal analysis, but isolating a feature is a precondition for intervention, not a substitute for it.

Even with an appropriate unit, causal methods carry their own risks, the most basic of which is that an intervention may not measure what it appears to. Modifying activations can move the model outside its training distribution or recruit pathways that normal processing would not use. Makelov et al. (2024) term these effects interpretability illusions, cases in which an intervention appears to reveal a mechanism but is an artifact of the manipulation. Two further concerns compound this one. Ablating a component can trigger compensatory behavior elsewhere in the network, the Hydra effect (McGrath et al. 2023), so that a component that appears dispensable was in fact contributing. Circuit faithfulness metrics introduce additional researcher degrees of freedom. They depend on many design choices, including the patching method, the corruption distribution, the granularity of components, and whether the intervention removes or restores a signal (Miller et al. 2024). Establishing that a discovered circuit is faithful, complete, and robust to alternative interventions therefore remains a demanding standard.

A final distinction concerns what a successful intervention shows. Localizing a behavior to a component identifies where a computation happens, not what algorithm the component implements. Two models can localize the same behavior to different components, and one model can implement the same behavior through different components across inputs. A circuit is therefore most informative when it is tied to a specific variable and shown to carry that variable across the conditions in which the behavior holds. Localization without this link describes the network's anatomy rather than its computation.

No lens is self-sufficient. Representation shows what is encoded but not whether it is used. Behavior shows what is expressed but not how. Mechanism shows what is implemented, but depends on intervention choices that can shape the result. Table 3 summarizes the three lenses, their main methods, and the limitation that motivates the

move from each lens to the next. These complementary limitations motivate the standard applied in the rest of the review. A claim about linguistic knowledge is strongest when all three lenses are aligned on the same target and address complementary evidential limitations.

Convergence, as used here, presupposes target alignment. The three lenses bear on the same target only when studies share a linguistic construct, model regime, and behavioral contrast. When these conditions are not met, apparent agreement or disagreement may reflect evidence addressing different targets rather than a property of the model. Target alignment is therefore a condition for assessing convergence, not a definition of linguistic knowledge. We distinguish two grades of target alignment. Under the strict grade, a single study decodes, manipulates, and behaviorally tests the same variable in the same model and on the same items. Construct, model regime, and behavioral contrast are therefore matched by design. Under weaker alignment, separate studies share a construct and behavioral contrast but differ in model regime, items, or both. In the evidence maps, a study is labeled "matched within study" only if it meets the strict criterion. We base convergence claims on such matched studies whenever they are available. Studies that differ in model regime or items can still offer cross-study triangulation. However, this weaker evidence supports conclusions only about the shared variable, not about a direct matched comparison.

4 Subject–Verb Agreement: A Well-Aligned Case

Subject–verb agreement is a narrowly specified morphosyntactic dependency. Across pretrained Transformers, grammatical number is often linearly recoverable from hidden states. In selected paradigms, causal interventions also link this feature to agreement behavior (Section 4.1). Behaviorally, models succeed on canonical items but degrade under attractor interference, structural distance, and reduced distributional support (Section 4.2). Mechanistically, the relevant circuit varies with input complexity, architecture, and scale rather than remaining fixed (Section 4.3).

4.1 Grammatical Number Is Recoverable and Functionally Used

In the agreement paradigms reviewed, grammatical number is often linearly recoverable from Transformer hidden states. Simple classifiers can decode it from contextual embeddings. It can also be separated into nominal and verbal encodings and localized to identifiable representational subspaces (Hao and Linzen 2023). Recoverability alone, however, does not establish that the model uses the recovered information.

Intervention studies bridge this gap. For instance, Lasri et al. (2022) combine linear probing with amnesic intervention. After identifying a linear number feature in BERT, they erase it from the representation and find the predicted decline in number-agreement performance. This result supports a functional role for the probed feature rather than mere recoverability. Hao and Linzen (2023) extend the intervention from erasure to directional manipulation. They reflect BERT's subject-position hidden state across the hyperplane separating singular from plural representations. In effect, this procedure substitutes one number value for the other. This manipulation

Lens	Main methods	What it can establish	Characteristic limitation
Representational	Probing classifiers (Hewitt and Manning 2019; Belinkov 2022), geometric and similarity analyses (Coenen et al. 2019; Chrupała and Alishahi 2019), amnesic probing (Elazar et al. 2021)	Whether a linguistic feature is recoverable from internal states, and where it is located	Recoverability does not establish functional use. Probe accuracy partly reflects the probe itself (Hewitt and Liang 2019)
Behavioral	Minimal-pair probability comparison (Warstadt et al. 2020; Gauthier et al. 2020), surprisal (Wilcox et al. 2021), metalinguistic prompting (Hu and Levy 2023)	What generalizations the model expresses in its output distribution	A correct output does not reveal the computation behind it. Shortcuts can mimic competence (McCoy et al. 2019)
Mechanistic	Activation and path patching (Goldowsky-Dill et al. 2023; Syed et al. 2024), ablation, causal mediation (Feder et al. 2022), sparse autoencoders (Huben et al. 2023)	Which components causally contribute to a behavior, and through which routes	Findings are task-local and sensitive to intervention choices. Interventions can create illusions (Makelov et al. 2024; Miller et al. 2024)

Table 3 The three lenses compared. Each lens establishes something the others cannot, and each carries a limitation that the next lens is positioned to address. Sections 4–6 apply this framework to three phenomena.

reverses the model's verb-form preference. Predictions go from 91% correct before the intervention to up to 85% incorrect after it. They also find that the relevant encoding is concentrated at the subject position in early layers and at the verb position in later layers.

This representational–causal link also appears in autoregressive models, though the implementing components differ by architecture. Ferrando and Costa-Jussà (2024) identify a circuit in Gemma 2B in which an attention head writes a number signal into the residual stream and final-layer MLP neurons read it. Africa (2025) identify a comparable circuit in GPT-2 Small. At benchmark scale, Arora et al. (2024) apply

distributed alignment search (Geiger et al. 2024) across the Pythia family and recover the agreement subspace consistently across model sizes. Taken together, these studies suggest that grammatical number can be causally localized in several model settings. Feature-level circuit discovery extends this line. Marks et al. (2025) propose a scalable method for identifying circuits of sparse autoencoder features, motivated by the observation that neurons and attention heads are often polysemantic, and validate it by discovering and evaluating feature circuits on a suite of subject–verb agreement tasks. The resulting agreement circuits are built from fine-grained, more nearly interpretable units rather than from neurons or attention heads.

4.2 Models Succeed on Canonical Agreement but Degrade Systematically

Behavioral evaluations document both the conditions under which pretrained Transformers succeed at agreement and those under which their performance becomes fragile. The positive baseline is well established. On simple adjacent-agreement items, Transformer models typically assign higher probability to the correctly inflected verb. The targeted stimuli for testing this preference were introduced with recurrent models (Marvin and Linzen 2018). An early direct assessment applied such stimuli to BERT and found the same preference (Goldberg 2019). The pattern was subsequently confirmed in BLiMP (Warstadt et al. 2020) and SyntaxGym (Gauthier et al. 2020), and in systematic comparisons across model classes (Hu et al. 2020).

The first boundary condition is structural distance in the presence of attractors. When subject and verb are separated by intervening material, especially material containing a noun of mismatching number, performance drops in a graded fashion. Finlayson et al. (2021) document this pattern across GPT-2, Transformer-XL, and XLNet. The preference for the grammatical verb declines when prepositional phrases or relative clauses intervene. It declines further when the intervening material contains an attractor. Ryu and Lewis (2021) provide convergent evidence from GPT-2. They find that surprisal on the verb in ungrammatical sentences is reduced when a distractor noun matches the verb in number.

The second boundary condition is distributional support. Models perform better when the correct verb form is supported by local semantic or lexical context (Newman et al. 2021; Lasri et al. 2022), and they show biases toward more frequent forms under lexical ambiguity (Yu et al. 2020). Wei et al. (2021) show that this frequency dependence is causal. In controlled training experiments with BERT, both the absolute frequency of a verb form and its frequency relative to the competing inflection shape agreement predictions at inference time. Their results suggest that BERT often applies the agreement dependency correctly but struggles when number features must be inferred for infrequent lexical items. This behavioral drop on low-support items has a representational correlate. Lasri et al. (2022) show that when BERT's agreement accuracy falls on semantically anomalous sentences, the linearly recoverable number signal weakens with it. Thus, representational and behavioral effects move together in this setting. Lower recoverability is accompanied by poorer agreement behavior.

The behavioral evidence therefore indicates a graded performance profile. Pretrained Transformers handle canonical adjacent agreement near ceiling, and their

accuracy falls off in a structured way as distributional support weakens or structural interference grows. This pattern aligns with the representational evidence in Section 4.1. As noted above, the recoverable number signal weakens exactly where behavior degrades. Grammatical number is recoverable and can be functionally used, but its use depends on how clearly the relevant feature is encoded.

4.3 Circuits Vary with Complexity, Architecture, and Scale

The circuit that implements agreement varies with input complexity, architecture, and scale. Cross-level convergence for this phenomenon therefore holds at the level of the variable, grammatical number, rather than at the level of any particular circuit. This line of analysis predates Transformers. Individual units carrying grammatical number were first identified in LSTM language models (Lakretz et al. 2019), establishing that agreement can be tracked to sparse internal components. The Transformer studies reviewed here inherit that question under a different architecture.

Circuit recruitment expands with input complexity. Finlayson et al. (2021) apply causal mediation analysis to GPT-2, Transformer-XL, and XLNet, extending to syntactic agreement the framework first used to trace gender bias through model components (Vig et al. 2020). They find that adjacent agreement can be supported by a relatively small set of neurons in mid-layer feedforward components, while non-adjacent agreement engages a broader set that grows with structural complexity. Circuit-discovery work in GPT-2 Small reaches a compatible conclusion. A compact circuit supports base agreement, and additional heads join it for plural subjects, pronouns, and negation contexts (Africa 2025). These additional heads could be read as construction-specific machinery. Within a single model, however, complexity-dependent expansion occurs over substantially reused machinery. Functional localization across seven open-weight models shows that distinct agreement constructions, including subject–verb, anaphor, and determiner–noun agreement, recruit overlapping sets of causally implicated units (Kryvosheieva et al. 2026). The emerging picture is therefore a shared pool of agreement resources whose recruitment broadens for more demanding inputs.

The implementing components also differ across architectures. Mueller et al. (2022) compare syntactic agreement neurons in masked and autoregressive models under controlled causal interventions. They find that BERT-style and XGLM-style models rely on substantially non-overlapping sets of neurons for the same task. This finding is consistent with the broader observation that linguistic information is more distributed in BERT and more concentrated in a few layers in XLNet (Dalvi et al. 2020). Functionally equivalent agreement behavior can be implemented by physically distinct components, so a circuit identified in one architecture should not be assumed to transfer to another.

Scale also reshapes agreement circuits. Finlayson et al. (2021) report that larger variants distribute the computation across more components. In smaller models, structural knowledge is concentrated in a few neurons with strong inflectional preferences. In larger ones it is spread across more, individually weaker neurons, with the causal effect attributable to any single neuron decreasing as size grows. This instantiates a more general property of over-parameterized Transformers, which carry substantial

representational redundancy (Dalvi et al. 2020). For agreement, increased scale does not make the mechanism cleaner or more modular. Instead, it distributes the relevant computation more widely.

4.4 Summary: Agreement as a Bounded Success Case

Subject–verb agreement provides the clearest focal case in this review. Selected studies offer matched within-study evidence, while other studies provide consistent results across model families. Grammatical number is recoverable from hidden states by probing and shown to be functionally used by causal intervention. Behaviorally, models perform strongly on canonical items and degrade systematically under attractor interference, structural distance, and reduced distributional support. The three levels concern the same target.

Across studies and model families, grammatical number remains the common variable. What varies is its implementation, which shifts with architecture, structural complexity, and scale, growing more distributed as models grow. Agreement is therefore a bounded success case. The strongest convergence arises in matched studies. Cross-model evidence provides triangulation at the level of grammatical number, not evidence for a shared circuit.

5 Negation: A Detection–Application Mismatch

Negation is the second phenomenon. It represents the semantic-operator case in the comparison. A marker such as *not* takes scope over a proposition and reverses its truth conditions, as in *the door is open* versus *the door is not open*. The computation has two parts that can come apart. The model must detect the marker and determine its scope. It must then apply the resulting truth-conditional consequence to predictions that depend on it. We therefore treat negation as the application of one operator to a complement, with detection and application as separable steps.

The reviewed evidence provides stronger support for detection than for semantic application. Reviewed studies recover marker and scope information from hidden states, but have not established a recoverable representation of the interpreted negated proposition (Section 5.1). Behaviorally, models register the marker yet apply its consequence unreliably (Section 5.2). Mechanistic studies localize different negation-related processes because they operationalize different target computations (Section 5.3). The pattern is therefore a detection–application mismatch. The evidence behind it is unevenly distributed. Behavioral studies dominate, while representational and mechanistic work on negation remains thinner and more recent. Table 5 maps the studies supporting this configuration.

5.1 Marker and Scope Information Are Recoverable, but Proposition-Level Representation Remains Unresolved

Negation leaves recoverable traces in Transformer hidden states. Classifiers can detect marker presence from representations of negated verbs and their arguments in BERT (Celikkanat et al. 2020). Tokens within the semantic scope of negation also contain

Level	Representative studies	Model regime	Target variable	Behavioral contrast	Cross-level link
R	Lasri et al. (2022)	Masked (BERT)	Grammatical number	Number-agreement accuracy under amnesic erasure	Matched within study
B	Goldberg (2019), BLiMP (Warstadt et al. 2020), SyntaxGym (Gauthier et al. 2020), Hu et al. (2020)	Masked and autoregressive	Grammatical number	Number-agreement minimal pairs on canonical items	Same variable as R and M
B	Ryu and Lewis (2021); Newman et al. (2021); Yu et al. (2020); Wei et al. (2021)	Masked and autoregressive	Grammatical number under attractors and low support	Attractor, distance, and frequency effects	Same variable as R and M
M	Hao and Linzen (2023)	Masked (BERT)	Grammatical number (subject-position subspace)	Verb-conjugation prediction, directional flip	Matched within study
M	Finlayson et al. (2021)	Autoregressive / permutation-LM (GPT-2 family, Transformer-XL, XLNet)	Grammatical number (agreement neurons)	Attractor and distance preference	Matched within study
M	Ferrando and Costa-Jussà (2024); Africa (2025); Arora et al. (2024)	Autoregressive (Gemma 2B, GPT-2 Small, Pythia)	Grammatical number (number-signal circuit)	Agreement preference	Circuit to behavior
M	Marks et al. (2025)	Autoregressive (Pythia, Gemma 2)	Agreement feature circuits (sparse autoencoder features)	Subject–verb agreement task suite	Matched within study
M	Kryvosheieva et al. (2026)	Autoregressive (seven open-weight models)	Agreement-responsive units across construction types	Agreement performance under ablation	Within study
M	Mueller et al. (2022)	Masked and autoregressive multilingual models (mBERT, XGLM)	Grammatical number (agreement neurons)	Agreement preference	Same variable, distinct circuits

Table 4 Evidence summary for subject–verb agreement. Grammatical number is the target variable at each evidential level, and several studies link it to matched behavioral contrasts within a single design.

more information about marker presence and nearby negative-polarity items than tokens outside that scope (Kletz et al. 2023b). Controlling linear distance suggests that part of this effect reflects structural scope rather than proximity alone.

The interpretation of these results remains limited. Some within-scope advantages extend to unrelated lexical items. Probes may therefore partly track clause structure rather than a representation specific to negation (Kletz et al. 2023b). More importantly, the decoded targets are marker presence and scope-related information. They are not equivalent to a representation of the proposition after negation has been semantically applied. No reviewed intervention establishes a causal role for such a proposition-level variable.

5.2 Models Apply Negated Meaning Inconsistently

Behavioral studies repeatedly find that models respond to negation markers less reliably than their truth-conditional role requires. This pattern recurs across evaluation formats and model regimes. In cloze completion, BERT often produces continuations compatible with the affirmative interpretation of a negated sentence (Ettinger 2020), and masked models can assign similar probability distributions to affirmative and negated factual prompts (Kassner and Schütze 2020). These tests read the failure off the probability distribution. A related limitation appears under stricter tests in which models must distinguish a sentence directly from its negation. These tests probe meaning–text correspondence rather than distributional plausibility (Jang et al. 2022). Beyond masked models, autoregressive and instruction-tuned systems show difficulty across lexical, factual, and reasoning-oriented negation benchmarks (Truong et al. 2023). In supervised classification, affirmative commonsense statements are handled more accurately than negated counterparts, with evidence of reliance on superficial cues (García-Ferrero et al. 2023).

Improved test design moderates the strongest versions of this conclusion. Multi-sentence contexts can reveal graded polarity sensitivity that isolated cloze tests miss (Gubelmann and Handschuh 2022). Stricter minimal pairs show clearer sensitivity in RoBERTa-large than in BERT-base, although semantically forbidden tokens can remain highly ranked (Kletz et al. 2023a). Larger diagnostic sets likewise reveal more sensitivity than early evaluations with few negation items suggested (Shivagunde et al. 2023). Models are therefore not uniformly blind to negation. Sensitivity is graded, and stricter designs consistently reveal more of it than early evaluations suggested. Across the reviewed designs, the central asymmetry remains. Semantic application is less reliable than marker detection.

Evidence concerning scale is heterogeneous. Larger models continue to fail on some benchmarks (Truong et al. 2023), while relationships between size and performance vary across model families and prompts (Zhang et al. 2023). Inverse scaling has also been reported for selected negation tasks (McKenzie et al. 2023), and negated prompts can induce factual inconsistency in generation (Varshney et al. 2025). These studies do not establish a universal effect of parameter count. Together these studies support a narrow conclusion. Scale alone has not produced consistent improvement across the evaluated settings.

Two explanations remain plausible. Negation may be sparse or unnecessary in common training and evaluation corpora, allowing affirmative priors to dominate (Hossain et al. 2022). Consistent with this first explanation, augmenting the training objective with an unlikelihood term for negated statements substantially reduces negation errors in BERT (Hosseini et al. 2021). The second explanation concerns learning dynamics. Retrieval of an affirmative answer and inversion under negation may develop at different rates (Zhang et al. 2023). Current evidence does not reliably distinguish an encoded interpretation that is overridden from an interpretation that is never computed.

5.3 Negation Mechanisms Depend on the Target Task

Mechanistic analyses have begun to localize negation-related effects, but they operationalize different computations. Work applying activation patching and ablation in GPT-2 Small finds sensitivity to affirmative–negated minimal pairs concentrated in a small set of mid-layer attention heads (Al Mofael et al. 2026). Work on syllogistic reasoning identifies components that affect truth-value logits (Saraipour and Zhang 2025). Analyses of sentiment classification instead find broader late-layer contextual integration (Hatua 2025). These findings are not directly comparable because marker sensitivity, truth-value transformation, and sentiment integration are distinct targets.

Related polarity phenomena[1] can be more tractable. Sensitivity to the licensing conditions of negative polarity items was documented first in recurrent language models (Jumelet and Hupkes 2018) and then examined in BERT with multiple analysis methods (Warstadt et al. 2019). More recent work localizes NPI licensing with causal tools (Bylinina and Tikhonov 2022).

5.4 Summary: Negation as a Detection–Application Mismatch

Negation presents partial evidence at each level, but these forms of evidence cannot yet be directly combined. The marker and aspects of its scope are recoverable from hidden states. In the reviewed corpus, no study identifies a recoverable representation of the interpreted negated proposition. Behaviorally, models register the marker yet apply its consequence unreliably, and improved test design narrows but does not remove the shortfall. Mechanistic studies identify components affecting negation-related behavior, but each localizes a different computation. The resulting evidence does not yet describe a single mechanism.

The defining pattern is a mismatch between detecting a negation marker and applying its consequence. Evidence for marker detection is available across the three lenses. The application side is not, and the two have not been measured on the same proposition-level target. Negation therefore shows a detection–application mismatch rather than a directly matched encoding–use dissociation. The source of the failure remains open. It may lie in encoding the proposition or in using it.

[1] NPI licensing constrains where items such as *any* may appear. It is related to negation through polarity but is not identical to truth-conditional negation. The NPI studies therefore provide evidence about polarity sensitivity only. They neither establish a representation of the interpreted negated proposition nor directly resolve the detection–application mismatch.

Level	Representative studies	Model regime	Target variable	Behavioral contrast	Cross-level link
R	Celikkanat et al. (2020); Kletz et al. (2023b)	Masked (BERT)	Marker presence and scope	—	None
B	Ettinger (2020); Kassner and Schütze (2020)	Masked (BERT)	Truth-conditional interpretation	Affirmative vs negated cloze and factual prompts	None
B	Gubelmann and Handschuh (2022); Kletz et al. (2023a); Shivagunde et al. (2023); Jang et al. (2022)	Masked (BERT, RoBERTa)	Truth-conditional interpretation under stricter test design	Multi-sentence contexts, strict minimal pairs, larger diagnostic sets, meaning–text correspondence	None
B	Truong et al. (2023); García-Ferrero et al. (2023); Varshney et al. (2025)	Autoregressive and instruction-tuned	Truth-conditional interpretation	Lexical, factual, reasoning, classification, and generation contrasts	None
B	Zhang et al. (2023); McKenzie et al. (2023)	Autoregressive / scaling range	Negation under scaling	Negation QA, inverse scaling	None
B	Hosseini et al. (2021)	Masked (BERT, targeted objective)	Truth-conditional interpretation under modified training	Negated factual prompts before and after unlikelihood training	None
B	Warstadt et al. (2019)	Masked (BERT)	NPI licensing	Acceptability and cloze diagnostics across licensing environments	None
M	Al Mofael et al. (2026)	Autoregressive (GPT-2 Small)	Marker sensitivity	Affirmative–negated minimal pairs	Within study, marker only
M	Saraipour and Zhang (2025)	Autoregressive (GPT-2 Small)	Binary truth-value transformation	Syllogism truth value	Within study, task-specific target
M	Hatua (2025)	Autoregressive (GPT-2, 12-layer)	Contextual integration in sentiment processing	Sentiment polarity under negation and other contextual shifts	Within study, sentiment-specific target
M	Bylinina and Tikhonov (2022)	Masked and autoregressive	NPI licensing	Licensing preference under intervention	Within study, NPI-specific target

Table 5 Evidence summary for negation. The representational target is marker presence or scope, whereas the main behavioral target is truth-conditional interpretation. Behavioral studies dominate the evidence base, and no reviewed study links marker detection, proposition-level representation, and behavioral application within the same interpreted proposition.

6 Compositional Generalization: Fragmented Evidence Across Forms of Recombination

Compositional generalization is the third phenomenon. It represents the open-ended recombination case in the comparison. The classical principle of compositionality concerns how meanings combine (Partee et al. 1984), but the capacity relevant here extends to recombination more generally, whether the parts are meanings, facts, or the outputs of earlier steps (McCurdy et al. 2024). Unlike agreement and negation, it often requires the computation and combination of intermediate results. Within neural network research, this broader capacity has itself been decomposed into distinct generalization abilities, including systematicity, productivity, and substitutivity (Hupkes et al. 2020). Each benchmark reviewed below targets only a subset of these abilities. This variation is one source of the fragmentation documented in this section.

The literature addresses this capacity through three lines of work that do not target the same thing. Representational studies find hidden states sensitive to phrase structure and to compositional complexity. Yet no reviewed study isolates a compositional variable and links it to behavior through intervention, and the complexity that has been measured tracks nonlinear rather than linear geometry (Section 6.1). Behavioral studies show stronger performance when operations are separable or when a decomposition is supplied. Performance is less consistent when models must execute unaided sequential chains. In the reviewed evaluations, increased scale has not consistently removed this limitation (Section 6.2). Mechanistic studies trace two-step composition to a detectable, causally active intermediate, but only for the shortest chains. When the embedding geometry permits an alternative, the model instead follows a non-compositional route (Section 6.3). These lines of work are related, but evidence from one does not establish the others. Their positive results therefore align only in restricted subcases.

A methodological caveat applies to both positive and negative findings. Behavioral success on a complex task does not by itself show that the success was achieved compositionally (McCurdy et al. 2024). A model may have learned a general recombination procedure, it may have exploited similarity between training and test distributions, or the test combinations may already have been available in pretraining (Kim et al. 2022). Failure is equally inconclusive, since the intended generalization may be underdetermined by the training data (Jabbar et al. 2025). The question is not whether pretrained Transformers can produce compositionally correct outputs on some tasks. It is whether those outputs reflect a robust and general capacity for chained recombination.

6.1 Representations Reflect Compositional Complexity, but the Relevant Feature Remains Unvalidated

Transformer representations are sensitive to phrase structure. An affine transformation of child representations can predict parent-phrase representations with moderate accuracy across several model families (Liu and Neubig 2022). This result captures regularities in phrase representations, but it does not distinguish semantic composition from other phrase-level associations, such as idiomatic or lexicalized expressions. Representational geometry also varies with compositional complexity. Across Pythia

models and checkpoints, nonlinear intrinsic dimension tracks the complexity of linguistic stimuli more closely than linear dimensionality (Lee et al. 2025). This finding suggests that linear summaries may omit relevant structure.

No reviewed study identifies a common compositional variable, manipulates it, and demonstrates a matched change in systematic generalization. The representational conclusion must therefore remain narrow and specific. Compositional complexity is reflected in activation geometry, yet no study in the reviewed corpus identifies a causally validated representation of systematic recombination.

6.2 Behavioral Evidence Shows Task-Dependent Recombination

Behavioral evidence comes from several benchmark families. SCAN evaluates generalization from familiar commands to novel action sequences (Lake and Baroni 2018). CFQ constructs question-answering splits that separate familiar compounds from new combinations (Keysers et al. 2020). COGS tests new combinations of lexical items and syntactic structures in semantic interpretation (Kim and Linzen 2020). These benchmarks were introduced with models trained from scratch. The behavioral evidence considered here comes from their later application to pretrained models (Qiu et al. 2022; Drozdov et al. 2023).

Performance is often stronger when component operations can be applied independently than when each step depends on the output of a preceding step. In-context evaluations across composite tasks show that scaling helps more consistently on separable operations than on sequential chaining (Xu et al. 2024). Within the same regime, a second limit concerns distribution rather than task structure. A generalization gap between in-distribution and compositional splits persists, narrowing with scale but not closing (Hosseini et al. 2022). The gap is also sensitive to demonstration design, since the selection and structure of in-context examples shape how well models generalize to new compositions (An et al. 2023). External decomposition can substantially improve performance. Chain-of-thought prompting elicits intermediate steps before the final answer (Wei et al. 2022). Least-to-most prompting decomposes a problem into subproblems solved in sequence (Zhou et al. 2023), and its dynamic variant uses syntactic decomposition to generate the intermediate parses (Drozdov et al. 2023). Skills-in-context prompting goes further and supplies both component skills and worked examples of their composition (Chen et al. 2024). These methods show that scaffolding can improve performance by eliciting explicit intermediate steps. They do not establish that base pretrained models apply the same procedures without scaffolding.

Under standard prompting, models can answer subquestions correctly yet fail to combine those answers, producing a compositionality gap (Press et al. 2023). The distributional limit observed in context appears here as well. Larger models improve more consistently on in-distribution semantic-parsing splits than on compositional splits (Qiu et al. 2022). Even large instruction-tuned models show sharp performance decay as required composition depth increases. This pattern is consistent with reliance on memorized subcomponents, although it does not establish that explanation (Dziri et al. 2023). These findings indicate that parameter scale does not yield uniform gains

under distribution shift, although they do not establish that scale can never improve compositional generalization.

Both success and failure therefore require the careful reading set out at the start of this section. Evaluations that restrict co-occurrence shortcuts provide the stronger evidence. Nested referring-expression tasks, for example, reveal category-dependent latent composition when entity co-occurrence cues are limited (Yang et al. 2025). Taken together, the behavioral literature supports a mixed, task-dependent pattern rather than a uniform conclusion about compositional generalization.

6.3 Short-Chain Mechanisms Are Traceable

Mechanistic studies provide their clearest evidence in two-step factual or referential tasks. In LLaMA-2-7B, an intermediate referent becomes decodable in middle-layer residual states before the final answer is produced. Interventions on attention modules at these layers alter the final prediction (Li et al. 2024). This finding links an intermediate representation to output in a restricted multi-hop setting.

The same surface task can also be solved by different internal strategies. In two-hop tasks, an intermediate variable is detectable when the model follows a compositional route. It may be absent, however, when embedding geometry permits a direct input-to-output mapping (Khandelwal and Pavlick 2025). The presence of a correct answer therefore does not guarantee a compositional mechanism.

These findings show that pretrained Transformers can implement short chains under selected conditions. They do not identify a general circuit for recursive composition, quantifier scope, or systematic generalization across benchmark splits in the reviewed corpus. Training objectives designed to encourage systematicity can produce stronger generalization in restricted instruction-learning paradigms (Lake and Baroni 2023). Modest changes to standard Transformer training likewise improve systematic generalization on compositional benchmarks (Csordás et al. 2021). This result shows that architecture, objective, and data jointly constrain performance. It does not isolate ordinary pretraining as the sole source of the limitation. The view that large-corpus pretraining provides only weak meta-learning for composition is compatible with this evidence, but remains an interpretation rather than a demonstrated mechanism (Jacob et al. 2024).

6.4 Summary: Compositional Generalization as Fragmented Evidence

Compositional generalization presents positive findings across several distinct targets. Representations are sensitive to compositional complexity, but no study isolates a common compositional variable or validates it causally. Behaviorally, models recombine familiar elements under separable tasks, scaffolding, and supportive distributions, and falter under unaided sequential chaining. Mechanistically, short two-hop chains can contain a causally active intermediate, but no reviewed study identifies a general circuit for systematic recombination.

The defining configuration is fragmentation. Phrase geometry, benchmark recombination, and two-hop computation are distinct targets, so their positive results do

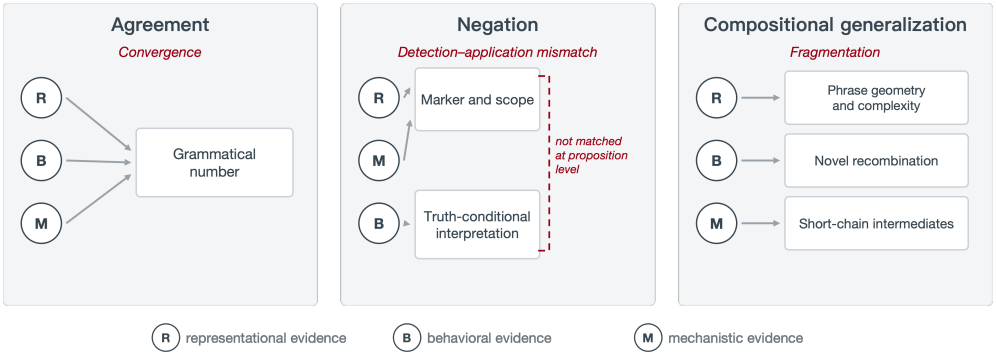

Fig. 2 Three evidential configurations for linguistic knowledge in pretrained Transformers. Agreement shows convergence on grammatical number. Negation shows a detection–application mismatch between marker and scope information and proposition-level interpretation. Compositional generalization shows fragmentation across phrase geometry and complexity, recombination benchmarks, and short-chain intermediates. The diagram is conceptual rather than quantitative.

not yet assemble into one account of systematic recombination. The primary unresolved question is accessibility, because the relevant structure has not been shown to be available in a form that analysis can isolate. Section 7 takes up what a shared paradigm would require to settle it.

7 Discussion and Open Questions

7.1 Three Evidential Configurations

The three focal cases yield three qualitatively different relations among the evidential levels, as summarized in Figure 2. Agreement shows the clearest convergence, particularly in selected matched paradigms. Negation shows a detection–application mismatch. Compositional generalization shows fragmentation.

These configurations should be interpreted against the target-alignment standard in Section 3. The three focal cases meet that standard to different degrees. Agreement meets the standard most closely in selected matched studies. Convergence concerns grammatical number as a shared variable rather than a single shared circuit. Negation meets it less fully, because detecting an operator is not the same construct as representing its interpreted consequence. Compositional generalization meets it least fully, since phrase formation, distribution-shift generalization, and multi-step computation are often treated as one capacity while operationalizing different targets.

Differences in target alignment are a central finding of the review. Representational studies report marker and scope information for negation, whereas behavioral studies report unreliable application of negated meaning. Read as claims about one capacity, these results conflict. Read against their targets, they do not, because the first concerns the marker and the second the interpreted proposition. Part of the apparent disagreement in the literature is resolved once the distinct targets of these studies are specified. This diagnosis is available only because alignment was assessed during coding rather than required for corpus entry. Requiring a shared target at corpus

Level	Representative studies	Model regime	Target variable	Behavioral contrast	Cross-level link
R	Liu and Neubig (2022)	Masked and autoregressive (several families)	Phrase composition (affine map)	—	None
R	Lee et al. (2025)	Autoregressive (Pythia)	Compositional complexity (nonlinear intrinsic dimension)	—	None
B	Qiu et al. (2022)	Pretrained, fine-tuned and prompted	Systematic recombination on compositional splits of CFQ and COGS	In-distribution vs compositional splits	None
B	Xu et al. (2024); Press et al. (2023); Dziri et al. (2023); Yang et al. (2025)	Autoregressive and instruction-tuned	Chained recombination	Separable vs sequential tasks, compositionality gap, depth decay, nested reference	None
B	Hosseini et al. (2022); An et al. (2023)	In-context (autoregressive)	Compositional generalization under demonstrations	Compositional vs in-distribution splits, demonstration design	None
B	Wei et al. (2022); Zhou et al. (2023); Drozdov et al. (2023); Chen et al. (2024)	Prompted autoregressive	Scaffolded decomposition	Scaffolded vs unaided performance	None
M	Li et al. (2024); Khandelwal and Pavlick (2025)	Autoregressive (LLaMA-2-7B and others)	Two-hop intermediate referent	Two-hop factual answer	Within study, short chains only

Table 6 Evidence summary for compositional generalization. The benchmark papers introducing SCAN, CFQ, and COGS used models trained from scratch and are cited as paradigm foundations rather than corpus evidence. The behavioral rows therefore cite applications of these paradigms to pretrained models. Positive evidence appears at all three evidential levels, but the target variables differ. Phrase composition, benchmark recombination, in-context generalization, scaffolded decomposition, and short-chain intermediate states do not yet form a matched cross-level account of systematic recombination.

Phenomenon	Best-aligned target	Main unresolved issue
Agreement	Grammatical number, linking representation, intervention, and verb-form preference within single studies	Whether the matched evidence generalizes beyond selected paradigms, since the implementing circuit shifts with architecture and scale
Negation	The negation marker, recoverable from hidden states and causally implicated in minimal-pair sensitivity	Marker detection and proposition-level interpretation have not been measured on the same target
Compositional generalization	The two-hop intermediate referent, matched within single studies for the shortest chains	No common variable spans phrase geometry, benchmark recombination, and longer chains, so a causally validated account of systematic recombination is missing

Table 7 Best-aligned target and main unresolved issue for each phenomenon. The corresponding cross-level configurations are shown in Figure 2.

selection would have excluded these informative misaligned cases. Table 7 summarizes the best-aligned target and the main unresolved issue for each phenomenon.

One component of alignment is the model regime. The reviewed studies do not support a single general relationship between parameter scale and cross-level convergence. Larger agreement models can perform better while distributing causal effects across more components (Finlayson et al. 2021). Negation performance varies nonmonotonically across model families, prompts, and benchmarks (McKenzie et al. 2023; Zhang et al. 2023). Compositional performance improves under some separable or scaffolded tasks but less consistently under unaided sequential chaining (Press et al. 2023). These comparisons are observational, and parameter count covaries with architecture, data, optimization, instruction tuning, and prompting protocol. Scale should therefore be treated as one dimension of model regime rather than as an explanatory variable in its own right. Future comparisons should hold model family and evaluation constant where possible and report base and instruction-tuned systems separately.

7.2 Accessibility, Use, and Evidential Standards

Once a target is aligned, we assess the relation among evidential levels through two properties of that target. The first is whether current analyses can isolate a representation of the specified variable, often as a linear direction in activation space (Park et al. 2023). We do not claim that all features are linear. The claim is only that linguistic targets differ in how accessible they are to current analyses. The second property is whether manipulating that representation changes a behavior that depends on the same variable. A feature can be recoverable yet fail to govern the output, either because later computation does not read it or because other signals outweigh it.

The three configurations differ in the extent to which these properties are supported. In selected paradigms, agreement satisfies both conditions. Number-related directions can be decoded, altered, and linked to verb-form predictions. For negation, marker and partial scope information satisfy the first condition. However, behavioral evidence does not establish reliable application of their semantic consequences. Because representation and behavior have not been matched on the interpreted proposition, the failure cannot yet be localized to use alone. The evidence supports a mismatch between detecting the marker and applying its consequence. For compositional generalization, the first condition has not yet been demonstrated for a common target variable. No reviewed study isolates a common compositional variable and validates it causally, and existing complexity measures track nonlinear rather than linear geometry (Lee et al. 2025). Negation and compositional generalization therefore leave different open questions. For negation, the question lies downstream of accessible marker and scope signals. For compositional generalization, the question begins with accessibility itself.

This distinction is an evidential framework, not a complete causal theory. A failure to isolate a variable can reflect an absent representation, nonlinear or distributed coding, an unsuitable analysis, or an ill-defined target. A weak behavioral effect can reflect failure to use a representation, competition from other signals, or an evaluation that does not measure the intended consequence. Accessibility and use become informative explanatory candidates only when these alternatives are separated experimentally.

The framework also reframes the familiar encoding–use problem. Prior studies show that recoverability is not generally sufficient for behavioral use. Linearly encoded features can sometimes be removed without measurable behavioral degradation (Elazar et al. 2021). Across Transformer models, syntactic probe accuracy also need not predict performance on targeted evaluations (Agarwal et al. 2025). These findings justify a general caution about interpreting probes, but the present comparison shows that the caution takes different forms across phenomena.

A phenomenon-specific encoding–use gap requires a stricter test. The same variable should be decoded, manipulated, and linked to a matched behavioral change. Only agreement approaches this standard, and only in selected paradigms. The cross-phenomenon evidence therefore does not rank the size of one uniform encoding–use gap. It identifies what each configuration establishes and what remains unresolved, as Table 7 summarizes.

7.3 Rival Explanations and How to Separate Them

The proposed accessibility-and-use conditions face an immediate objection. Across the three focal cases considered here, accessibility covaries with recombinatory demand. Agreement pairs a relatively isolable target with a comparatively less demanding form of structured use. Compositional generalization pairs the least isolable target with the most demanding form. The present cases therefore cannot rule out a simpler account, on which cross-level convergence declines merely because the task becomes harder.

Three explanations are in play. The first is methodological maturity. Convergence may track the tools available, so that phenomena appear inaccessible only until more expressive analyses reach them. The second is construct precision. Convergence may

track how sharply the linguistic target is defined, so that fragmentation reflects loose constructs rather than model properties. The third is the hypothesis advanced here, that accessibility and use contribute independently of task difficulty and construct definition.

Negation illustrates why the present evidence cannot decide among these. Marker information is accessible while semantic behavior remains unreliable, which at first looks like direct support for the third explanation. But the accessible signal and the unreliable behavior concern different targets. Under construct precision, the pattern reflects that mismatch of targets rather than any property of the model. Under methodological maturity, a proposition-level representation may exist and simply lie beyond current probes. The case therefore fits all three accounts until representation and behavior are measured on the interpreted proposition itself.

The three explanations make different empirical predictions. If methodological maturity explains the pattern, cross-phenomenon differences should diminish as more expressive analyses are applied within otherwise matched designs. Sparse autoencoders and related methods are promising instruments for this test, since they seek feature directions that conventional probes may miss (Elhage et al. 2022; Huben et al. 2023; Zhang et al. 2025). For agreement, these methods have already been applied (Section 4.1). If construct precision explains the pattern, differences should diminish once the linguistic target, model regime, and behavioral contrast are held constant. If accessibility and use contribute independently, differences should persist after this alignment, and they should then be traceable either to a failure to isolate the target variable or to a failure of interventions on that variable to change matched behavior.

These predictions are testable with matched cross-level designs, and Section 7.7 sets out what such designs require. The accessibility-and-use hypothesis is valuable not because current evidence is stronger, but because it makes sharper commitments. It specifies where it would fail.

7.4 Implications for NLP Practice

Whatever ultimately explains the three configurations, they already carry consequences for evaluation and auditing practice. Each configuration warrants a different level of confidence in a different kind of evidence, and evaluation practice should reflect these differences.

The first consequence concerns benchmark interpretation. A high score on a behavioral benchmark shows what a model expresses, not how the expression is produced. Negation makes the risk concrete. A model can register a negation marker while completing a negated prompt with its affirmative content, so aggregate accuracy on mixed test sets can conceal a systematic failure of semantic application. Where a phenomenon separates detection from application, evaluation reports should score the two item types separately. The same logic extends to compositional tasks, where scaffolded and unaided performance should be reported apart, since improvement under decomposition prompting does not establish unaided capacity.

The second consequence concerns the evidential weight of probing in model audits. A probing result carries strong evidential weight only when the probed variable has been linked to matched behavior. For grammatical number this link has been

established. For negation it holds only at the level of the marker. For compositional generalization it holds only for the shortest chains, not for systematic recombination. The link is therefore graded across capabilities rather than simply present or absent. An audit that cites probing evidence should state which of these situations applies to the capability under review, and an audit that cannot state it should treat the probing result as descriptive rather than evidential.

The third consequence concerns which capability claims can currently be verified with causal tools. The agreement literature shows that a narrow morphosyntactic feature can be located, manipulated, and tied to output. This tractability should not be assumed to transfer. The reviewed evidence does not yet establish causal validation for proposition-level negation or open-ended recombination. Claims of verified competence in these domains therefore exceed what current methods can support. Stating this boundary is itself useful, since it tells practitioners where a verification claim can be checked and where it can only be asserted.

7.5 Implications for Human–Model Comparison

Language models can serve as informative comparison systems for psycholinguistic theory, but the value of the comparison depends on the evidential configuration. For agreement, models show graded effects of number, attractors, and structural distance that resemble aspects of human sentence processing (Ryu and Lewis 2021; Wilcox et al. 2021). Cross-level evidence gives this comparison a specific meaning. It identifies a shared computational variable, without showing that humans and models implement the same mechanism. Convergent parallels have been reported for nested agreement dependencies, although in recurrent rather than Transformer models (Lakretz et al. 2021).

Human negation processing unfolds over time and depends on supportive context (Kaup et al. 2006; Nieuwland and Kuperberg 2008; Tian et al. 2016). The difference is the endpoint. Humans eventually reach interpreted negated meanings in ordinary comprehension, whereas for models the marker and the interpreted proposition have not been linked in a matched design. The comparison is therefore incomplete rather than negative.

For compositional generalization, humans productively extend familiar structures to new combinations. The reviewed pretrained models, by contrast, often depend on benchmark construction, scaffolding, and distributional support (Fodor and Pylyshyn 1988; Lake et al. 2019). Models trained with objectives designed to encourage systematicity can approach human performance on restricted tasks (Lake and Baroni 2023). Such results identify conditions under which particular forms of generalization emerge. They do not justify a global judgment about either human or model compositionality.

Data efficiency further complicates the comparison. Children receive fewer linguistic tokens than large models, but learn through multimodal, interactive, and socially grounded experience (Gilkerson et al. 2017; Warstadt et al. 2023; Frank 2023). Warstadt and Bowman (2022) examine at length what model results can and cannot show about human acquisition under these differences. Model size also need not improve cognitive fit, since larger language models can predict human reading times less well than smaller ones (Oh and Schuler 2023). A human–model comparison is

therefore only as informative as its specification. It should state the focal case, model regime, behavioral measure, and explanatory purpose.

7.6 Limitations of the Present Review

This review has four limitations. First, although the corpus was assembled under the explicit criteria of Section 2, the synthesis is integrative and selective rather than exhaustive. Its conclusions may therefore reflect the availability and selection of studies that connect multiple evidential levels. Second, the three phenomena differ in construct breadth and methodological maturity. Their comparison cannot establish that one linguistic capacity is intrinsically more difficult than another, a point developed in Section 7.3. Third, the evidence combines masked and autoregressive base models with selected instruction-tuned and prompt-intervention studies used as boundary evidence. We identify these regime differences, but the underlying literature does not always permit controlled comparison. Fourth, the review is restricted primarily to English. Agreement, negation, and compositionality are realized differently across languages, which may change both their behavioral demands and their representational accessibility (Bender 2011; Blasi et al. 2022; Anh et al. 2024; Ismayilzada et al. 2025).

7.7 Future Directions

The predictions outlined in Section 7.3 require matched cross-level designs. The most informative next step is to build such designs. A study should define one target variable and construct a behavioral contrast that depends on it. It should then identify, manipulate, and evaluate a candidate representation within one model and using the same items. Agreement already approximates this design in selected paradigms and can serve as a template. For negation, the target should be the interpreted proposition rather than marker presence alone. For compositional generalization, representation and intervention should be evaluated using the same controlled recombination task as the behavioral assessment.

Future work should also distinguish failures of representation from failures of computation. A model may lack an appropriate internal variable, fail to propagate an available variable, or lack sufficient inference-time computation to apply it. Additional intermediate steps can expand the functions that a model expresses (Merrill and Sabharwal 2024). Diagnosing missing capacity therefore requires ruling out insufficient inference-time computation. Scaffolded evaluation alone cannot rule out this alternative.

Multilingual evidence is essential because typological variation can decouple properties that English conflates. Each of the three focal cases offers a concrete test. For agreement, languages with rich inflectional morphology distribute number marking across several exponents rather than concentrating it on one verb form (Anh et al. 2024). If the English result reflects the underlying variable rather than its surface expression, a number feature would be expected to remain isolable and causally active even when its realization is distributed. Existing causal work on multilingual models makes this test feasible. Agreement neurons have been compared across mBERT and XGLM (Mueller et al. 2022). Agreement-responsive units also partially overlap across

languages, with greater overlap for syntactically similar languages (Kryvosheieva et al. 2026).

For negation, negative-concord languages permit several negative markers within a single negated proposition (Zeijlstra 2004). Marker count and propositional polarity therefore come apart in a way that standard English does not permit. Such languages offer a discriminating prediction. A model relying only on marker detection would be expected to be disrupted by concord. A model representing the interpreted proposition would not show the same pattern. For compositional generalization, freer word order weakens the relation between linear position and structural role. Recombination tests in such languages can therefore help separate structural composition from positional heuristics.

These tests also address a distributional concern. Because English dominates pretraining corpora (Bender 2011; Blasi et al. 2022), English results cannot distinguish properties of a linguistic target from consequences of abundant training data. Cross-lingual designs that hold the focal case constant while varying its realization and corpus frequency could separate these explanations (Ismayilzada et al. 2025).

Finally, new focal cases should decouple the correlations present in the current selection. A particularly informative target would combine high recombinatory demand with an easily isolated representation. Another would show low apparent demand but an accessible feature that is not used behaviorally. Factual and relational knowledge may provide a starting point because linear accessibility has been reported without an obvious relation to recombinatory demand (Hernandez et al. 2024). Such cases could test whether accessibility and use provide explanatory leverage beyond construct precision and task difficulty.

8 Conclusion

This review provides a phenomenon-centered framework for evaluating linguistic knowledge in pretrained Transformers. It compares representational, behavioral, and mechanistic evidence within three focal cases. The comparison yields three distinct configurations. Agreement shows convergence. Negation shows a detection–application mismatch, and compositional generalization shows fragmentation. These configurations show that cross-level evidence must be interpreted relative to a specified linguistic target.

Agreement provides the strongest evidence because grammatical number can be decoded, manipulated, and linked to verb-form behavior in selected matched paradigms. Negation reveals a different pattern: models can detect markers and partial scope information without reliably applying the resulting proposition-level meaning. Compositional generalization presents a third pattern, in which positive evidence arises across several tasks but does not yet converge on a common variable or mechanism. Together, these findings show that the relation between what a model encodes, expresses, and implements is target-specific rather than a general property of any single method.

The practical implication is straightforward. Claims about linguistic knowledge should be strongest when the same variable is measured, manipulated, and

behaviorally tested within one design. This standard can guide model evaluation, mechanistic auditing, and human–model comparison. Extending it to new focal cases, languages, and nonlinear representations will help determine why evidential configurations vary across linguistic targets. The present review offers a testable framework for making that comparison.

Declarations

Funding. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Data availability. This review reports no new primary data. The evidence maps (Tables 4–6) summarize the coding of the corpus studies, all of which are published works cited in the reference list. The coding scheme is described in Section 2.5.

Author contributions. Yizhe Wang conceived and designed the review, performed the literature search, assessment, and coding, and wrote the manuscript. Zhenhua Ling contributed to the selection of the case-study phenomena, took part in resolving ambiguous inclusion decisions, and contributed to the revision of the manuscript. Both authors read and approved the final manuscript.

References

Africa DD (2025) Identifying a circuit for verb conjugation in gpt-2. arXiv preprint arXiv:250622105

Agarwal A, Jian J, Manning CD, et al (2025) Mechanisms vs. outcomes: Probing for syntax fails to explain performance on targeted syntactic evaluations. In: Christodoulopoulos C, Chakraborty T, Rose C, et al (eds) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Suzhou, China, pp 33737–33757, https://doi.org/10.18653/v1/2025.emnlp-main.1712, URL https://aclanthology.org/2025.emnlp-main.1712/

Al Mofael A, Kuhn LM, Alkadi G, et al (2026) Interpreting negation in gpt-2: Layer-and head-level causal analysis. In: 2026 IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, pp 0042–0050, https://doi.org/10.1109/CCWC67433.2026.11393646

Amini A, Pimentel T, Meister C, et al (2023) Naturalistic causal probing for morpho-syntax. Transactions of the Association for Computational Linguistics 11:384–403. https://doi.org/10.1162/tacl_a_00554, URL https://aclanthology.org/2023.tacl-1.23/

An S, Lin Z, Fu Q, et al (2023) How do in-context examples affect compositional generalization? In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, pp 11027–11052, https://doi.org/10.18653/v1/2023.acl-long.618, URL https://aclanthology.org/2023.acl-long.618/

Anh D, Raviv L, Galke L (2024) Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test. In: Kuribayashi T, Rambelli G, Takmaz E, et al (eds) Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics. Association for Computational Linguistics, Bangkok, Thailand, pp 177–188, https://doi.org/10.18653/v1/2024.cmcl-1.15, URL https://aclanthology.org/2024.cmcl-1.15/

Arora A, Jurafsky D, Potts C (2024) CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In: Ku LW, Martins A, Srikumar V (eds) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Bangkok, Thailand, pp 14638–14663, https://doi.org/10.18653/v1/2024.acl-long.785, URL https://aclanthology.org/2024.acl-long.785/

Beguš G, Dąbkowski M, Rhodes R (2025) Large linguistic models: Investigating llms' metalinguistic abilities. IEEE Transactions on Artificial Intelligence 6(12):3453–3467. https://doi.org/10.1109/TAI.2025.3575745

Behzad S, Sakaguchi K, Schneider N, et al (2023) ELQA: A corpus of metalinguistic questions and answers about English. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, pp 2031–2047, https://doi.org/10.18653/v1/2023.acl-long.113, URL https://aclanthology.org/2023.acl-long.113/

Belinkov Y (2022) Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics 48(1):207–219. https://doi.org/10.1162/coli_a_00422, URL https://aclanthology.org/2022.cl-1.7/

Belinkov Y, Glass J (2019) Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics 7:49–72. https://doi.org/10.1162/tacl_a_00254, URL https://aclanthology.org/Q19-1004/

Bender EM (2011) On achieving and evaluating language-independence in nlp. Linguistic Issues in Language Technology 6:1–26. https://doi.org/10.33011/lilt.v6i.1239

Blasi D, Anastasopoulos A, Neubig G (2022) Systematic inequalities in language technology performance across the world's languages. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association

for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 5486–5505, https://doi.org/10.18653/v1/2022.acl-long.376, URL https://aclanthology.org/2022.acl-long.376/

Brown TB, Mann B, Ryder N, et al (2020) Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS '20

Bylinina L, Tikhonov A (2022) Transformers in the loop: Polarity in neural models of language. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 6601–6610, https://doi.org/10.18653/v1/2022.acl-long.455, URL https://aclanthology.org/2022.acl-long.455/

Celikkanat H, Virpioja S, Tiedemann J, et al (2020) Controlling the imprint of passivization and negation in contextualized representations. In: Alishahi A, Belinkov Y, Chrupała G, et al (eds) Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, Online, pp 136–148, https://doi.org/10.18653/v1/2020.blackboxnlp-1.13, URL https://aclanthology.org/2020.blackboxnlp-1.13/

Chang TA, Bergen BK (2024) Language model behavior: A comprehensive survey. Computational Linguistics 50(1):293–350. https://doi.org/10.1162/coli_a_00492, URL https://aclanthology.org/2024.cl-1.9/

Chen J, Pan X, Yu D, et al (2024) Skills-in-context: Unlocking compositionality in large language models. In: Al-Onaizan Y, Bansal M, Chen YN (eds) Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, pp 13838–13890, https://doi.org/10.18653/v1/2024.findings-emnlp.812, URL https://aclanthology.org/2024.findings-emnlp.812/

Chrupała G, Alishahi A (2019) Correlating neural and symbolic representations of language. In: Korhonen A, Traum D, Màrquez L (eds) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp 2952–2962, https://doi.org/10.18653/v1/P19-1283, URL https://aclanthology.org/P19-1283/

Coenen A, Reif E, Yuan A, et al (2019) Visualizing and measuring the geometry of bert. In: Wallach H, Larochelle H, Beygelzimer A, et al (eds) Advances in Neural Information Processing Systems, vol 32. Curran Associates Inc., Red Hook, NY, USA, pp 8594 – 8603, URL https://proceedings.neurips.cc/paper_files/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf

Conmy A, Mavor-Parker AN, Lynch A, et al (2023) Towards automated circuit discovery for mechanistic interpretability. In: Proceedings of the 37th International

Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS '23

Conneau A, Kruszewski G, Lample G, et al (2018) What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In: Gurevych I, Miyao Y (eds) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp 2126–2136, https://doi.org/10.18653/v1/P18-1198, URL https://aclanthology.org/P18-1198/

Corbett GG (2006) Agreement. Cambridge Textbooks in Linguistics, Cambridge University Press, Cambridge

Csordás R, Irie K, Schmidhuber J (2021) The devil is in the detail: Simple tricks improve systematic generalization of transformers. In: Moens MF, Huang X, Specia L, et al (eds) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 619–634, https://doi.org/10.18653/v1/2021.emnlp-main.49, URL https://aclanthology.org/2021.emnlp-main.49/

Dalvi F, Sajjad H, Durrani N, et al (2020) Analyzing redundancy in pretrained transformer models. In: Webber B, Cohn T, He Y, et al (eds) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 4908–4926, https://doi.org/10.18653/v1/2020.emnlp-main.398, URL https://aclanthology.org/2020.emnlp-main.398/

Drozdov A, Schärli N, Akyürek E, et al (2023) Compositional semantic parsing with large language models. In: The Eleventh International Conference on Learning Representations, Kigali, Rwanda

Dziri N, Lu X, Sclar M, et al (2023) Faith and fate: limits of transformers on compositionality. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS '23

Elazar Y, Ravfogel S, Jacovi A, et al (2021) Amnesic probing: Behavioral explanation with amnesic counterfactuals. Transactions of the Association for Computational Linguistics 9:160–175. https://doi.org/10.1162/tacl_a_00359, URL https://aclanthology.org/2021.tacl-1.10/

Elhage N, Hume T, Olsson C, et al (2022) Toy models of superposition. arXiv preprint arXiv:220910652 https://doi.org/10.48550/arXiv.2209.10652

Ettinger A (2020) What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics 8:34–48. https://doi.org/10.1162/tacl_a_00298, URL https://

aclanthology.org/2020.tacl-1.3/

Feder A, Keith KA, Manzoor E, et al (2022) Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. Transactions of the Association for Computational Linguistics 10:1138–1158. https://doi.org/10.1162/tacl_a_00511, URL https://aclanthology.org/2022.tacl-1.66/

Ferrando J, Costa-Jussà MR (2024) On the similarity of circuits across languages: a case study on the subject-verb agreement task. In: Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, pp 10115–10125, https://doi.org/10.18653/v1/2024.findings-emnlp.591

Ferrando J, Sarti G, Bisazza A, et al (2024) A primer on the inner workings of transformer-based language models. arXiv preprint arXiv:240500208 https://doi.org/10.48550/arXiv.2405.00208

Finlayson M, Mueller A, Gehrmann S, et al (2021) Causal analysis of syntactic agreement mechanisms in neural language models. In: Zong C, Xia F, Li W, et al (eds) 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). Association for Computational Linguistics, Online, pp 1828–1843, https://doi.org/10.18653/v1/2021.acl-long.144, URL https://aclanthology.org/2021.acl-long.144/

Fodor JA, Pylyshyn ZW (1988) Connectionism and cognitive architecture: A critical analysis. Cognition 28(1):3–71. https://doi.org/10.1016/0010-0277(88)90031-5, URL https://www.sciencedirect.com/science/article/pii/0010027788900315

Frank MC (2023) Bridging the data gap between children and large language models. Trends in Cognitive Sciences 27(11):990–992. https://doi.org/10.1016/j.tics.2023.08.007, URL https://www.sciencedirect.com/science/article/pii/S1364661323002036

García-Ferrero I, Altuna B, Alvez J, et al (2023) This is not a dataset: A large negation benchmark to challenge large language models. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, pp 8596–8615, https://doi.org/10.18653/v1/2023.emnlp-main.531, URL https://aclanthology.org/2023.emnlp-main.531/

Gauthier J, Hu J, Wilcox E, et al (2020) SyntaxGym: An online platform for targeted evaluation of language models. In: Celikyilmaz A, Wen TH (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Online, pp 70–76, https://doi.org/10.18653/v1/2020.acl-demos.10, URL https://aclanthology.org/2020.acl-demos.10/

Geiger A, Wu Z, Potts C, et al (2024) Finding alignments between interpretable causal variables and distributed neural representations. In: Locatello F, Didelez V (eds) Proceedings of the Third Conference on Causal Learning and Reasoning, Proceedings of Machine Learning Research, vol 236. PMLR, pp 160–187, URL https://proceedings.mlr.press/v236/geiger24a.html

Gilkerson J, Richards JA, Warren SF, et al (2017) Mapping the early language environment using all-day recordings and automated analysis. American journal of speech-language pathology 26(2):248–265. https://doi.org/10.1044/2016_AJSLP-15-0169.

Goldberg Y (2019) Assessing bert's syntactic abilities. URL https://arxiv.org/abs/1901.05287, arXiv:1901.05287

Goldowsky-Dill N, MacLeod C, Sato L, et al (2023) Localizing model behavior with path patching. arXiv preprint arXiv:230405969

Gubelmann R, Handschuh S (2022) Context matters: A pragmatic study of PLMs' negation understanding. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 4602–4621, https://doi.org/10.18653/v1/2022.acl-long.315, URL https://aclanthology.org/2022.acl-long.315/

Gulordava K, Bojanowski P, Grave E, et al (2018) Colorless green recurrent networks dream hierarchically. In: Walker M, Ji H, Stent A (eds) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 1195–1205, https://doi.org/10.18653/v1/N18-1108, URL https://aclanthology.org/N18-1108/

Hao S, Linzen T (2023) Verb conjugation in transformers is determined by linear encodings of subject number. In: Bouamor H, Pino J, Bali K (eds) Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp 4531–4539, https://doi.org/10.18653/v1/2023.findings-emnlp.300, URL https://aclanthology.org/2023.findings-emnlp.300/

Hatua A (2025) Mechanistic interpretability of GPT-2: Lexical and contextual layers in sentiment analysis. In: NeurIPS 2025 Workshop on Efficient Reasoning, San Diego, URL https://openreview.net/forum?id=l8rcexhR9K

He L, Chen P, Nie E, et al (2024) Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). ELRA and ICCL, Torino, Italia, pp 4488–4497, URL https://aclanthology.org/2024.lrec-main.402/

Hernandez E, Sen Sharma A, Haklay T, et al (2024) Linearity of relation decoding in transformer language models. In: International Conference on Learning Representations, Vienna Austria, pp 10504–10526

Hewitt J, Liang P (2019) Designing and interpreting probes with control tasks. In: Inui K, Jiang J, Ng V, et al (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 2733–2743, https://doi.org/10.18653/v1/D19-1275, URL https://aclanthology.org/D19-1275/

Hewitt J, Manning CD (2019) A structural probe for finding syntax in word representations. In: Burstein J, Doran C, Solorio T (eds) 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). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4129–4138, https://doi.org/10.18653/v1/N19-1419, URL https://aclanthology.org/N19-1419/

Horn LR (1989) A Natural History of Negation. University of Chicago Press, Chicago

Hossain MM, Chinnappa D, Blanco E (2022) An analysis of negation in natural language understanding corpora. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Dublin, Ireland, pp 716–723, https://doi.org/10.18653/v1/2022.acl-short.81, URL https://aclanthology.org/2022.acl-short.81/

Hosseini A, Reddy S, Bahdanau D, et al (2021) Understanding by understanding not: Modeling negation in language models. In: Toutanova K, Rumshisky A, Zettlemoyer L, et al (eds) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, pp 1301–1312, https://doi.org/10.18653/v1/2021.naacl-main.102, URL https://aclanthology.org/2021.naacl-main.102/

Hosseini A, Vani A, Bahdanau D, et al (2022) On the compositional generalization gap of in-context learning. In: Bastings J, Belinkov Y, Elazar Y, et al (eds) Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), pp 272–280, https://doi.org/10.18653/v1/2022.blackboxnlp-1.22, URL https://aclanthology.org/2022.blackboxnlp-1.22/

Hu J, Levy R (2023) Prompting is not a substitute for probability measurements in large language models. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, pp 5040–5060, https://doi.org/10.18653/v1/2023.emnlp-main.306, URL https://aclanthology.org/2023.emnlp-main.306/

Hu J, Gauthier J, Qian P, et al (2020) A systematic assessment of syntactic generalization in neural language models. In: Jurafsky D, Chai J, Schluter N, et al (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 1725–1744, https://doi.org/10.18653/v1/2020.acl-main.158, URL https://aclanthology.org/2020.acl-main.158/

Huben R, Cunningham H, Smith LR, et al (2023) Sparse autoencoders find highly interpretable features in language models. In: NeurIPS 2023 Workshop on Attributing Model Behavior at Scale, New Orleans

Hupkes D, Dankers V, Mul M, et al (2020) Compositionality decomposed: How do neural networks generalise? Journal of Artificial Intelligence Research 67:757–795

Ide Y, Nishida Y, Vasselli J, et al (2025) How to make the most of LLMs' grammatical knowledge for acceptability judgments. In: Chiruzzo L, Ritter A, Wang L (eds) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Albuquerque, New Mexico, pp 7416–7432, https://doi.org/10.18653/v1/2025.naacl-long.380, URL https://aclanthology.org/2025.naacl-long.380/

Ismayilzada M, Circi D, Sälevä J, et al (2025) Evaluating morphological compositional generalization in large language models. In: Chiruzzo L, Ritter A, Wang L (eds) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Albuquerque, New Mexico, pp 1270–1305, https://doi.org/10.18653/v1/2025.naacl-long.59, URL https://aclanthology.org/2025.naacl-long.59/

Jabbar A, Condoravdi C, Potts C (2025) Distinguishing fair from unfair compositional generalization tasks. In: Christodoulopoulos C, Chakraborty T, Rose C, et al (eds) Findings of the Association for Computational Linguistics: EMNLP 2025. Association for Computational Linguistics, Suzhou, China, pp 20796–20807, https://doi.org/10.18653/v1/2025.findings-emnlp.1133, URL https://aclanthology.org/2025.findings-emnlp.1133/

Jacob R, Sam M, Danielle W (2024) From frege to chatgpt: Compositionality in language, cognition, and deep neural networks. In: de Brigard Felipe, Walter S (eds) Neuroscience and Philosophy. Vol. 2. MIT Press

Jang M, Mtumbuka F, Lukasiewicz T (2022) Beyond distributional hypothesis: Let language models learn meaning-text correspondence. In: Carpuat M, de Marneffe MC, Meza Ruiz IV (eds) Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, Seattle, United States, pp 2030–2042, https://doi.org/10.18653/v1/2022.findings-naacl.156, URL https://aclanthology.org/2022.findings-naacl.156/

Jumelet J, Hupkes D (2018) Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In: Linzen T, Chrupała G, Alishahi A (eds) Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, Brussels, Belgium, pp 222–231, https://doi.org/10.18653/v1/W18-5424, URL https://aclanthology.org/W18-5424/

Kassner N, Schütze H (2020) Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In: Jurafsky D, Chai J, Schluter N, et al (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 7811–7818, https://doi.org/10.18653/v1/2020.acl-main.698, URL https://aclanthology.org/2020.acl-main.698/

Kaup B, Lüdtke J, Zwaan RA (2006) Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? Journal of Pragmatics 38(7):1033–1050. https://doi.org/10.1016/j.pragma.2005.09.012, URL https://www.sciencedirect.com/science/article/pii/S0378216606000075, special Issue: Processes and Products of Negation

Keysers D, Schärli N, Scales N, et al (2020) Measuring compositional generalization: A comprehensive method on realistic data. In: ICLR, Online

Khandelwal A, Pavlick E (2025) How do language models compose functions? URL https://arxiv.org/abs/2510.01685, arXiv:2510.01685

Kim N, Linzen T (2020) COGS: A compositional generalization challenge based on semantic interpretation. In: Webber B, Cohn T, He Y, et al (eds) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 9087–9105, https://doi.org/10.18653/v1/2020.emnlp-main.731, URL https://aclanthology.org/2020.emnlp-main.731/

Kim N, Linzen T, Smolensky P (2022) Uncontrolled lexical exposure leads to overestimation of compositional generalization in pretrained models. arXiv preprint arXiv:221210769 https://doi.org/10.48550/arXiv.2212.10769

Kletz D, Amsili P, Candito M (2023a) The self-contained negation test set. In: Belinkov Y, Hao S, Jumelet J, et al (eds) Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, Singapore, pp 212–221, https://doi.org/10.18653/v1/2023.blackboxnlp-1.16, URL https://aclanthology.org/2023.blackboxnlp-1.16/

Kletz D, Candito M, Amsili P (2023b) Probing structural constraints of negation in pretrained language models. In: Alumäe T, Fishel M (eds) Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa). University of Tartu Library, Tórshavn, Faroe Islands, pp 541–554, URL https://aclanthology.org/2023.

nodalida-1.54/

Kryvosheieva D, de Varda AG, Fedorenko E, et al (2026) Different types of syntactic agreement recruit the same units within large language models. In: Liakata M, Moreira VP, Zhang J, et al (eds) Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, San Diego, California, United States, pp 209–227, https://doi.org/10.18653/v1/2026.acl-long.7, URL https://aclanthology.org/2026.acl-long.7/

Lake B, Baroni M (2018) Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In: International conference on machine learning, PMLR, Stockholm, Sweden, pp 2873–2882

Lake BM, Baroni M (2023) Human-like systematic generalization through a meta-learning neural network. Nature 623(7985):115–121. https://doi.org/10.1038/s41586-023-06668-3

Lake BM, Linzen T, Baroni M (2019) Human few-shot learning of compositional instructions. In: Goel AK, Seifert CM, Freksa C (eds) Proceedings of the Annual Meeting of the Cognitive Science Society. cognitivesciencesociety.org, Montreal, Canada, pp 611–617, URL https://escholarship.org/uc/item/5vr9p4ks

Lakretz Y, Kruszewski G, Desbordes T, et al (2019) The emergence of number and syntax units in LSTM language models. In: Burstein J, Doran C, Solorio T (eds) 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). Association for Computational Linguistics, Minneapolis, Minnesota, pp 11–20, https://doi.org/10.18653/v1/N19-1002, URL https://aclanthology.org/N19-1002/

Lakretz Y, Hupkes D, Vergallito A, et al (2021) Mechanisms for handling nested dependencies in neural-network language models and humans. Cognition 213:104699. https://doi.org/10.1016/j.cognition.2021.104699

Lasri K, Pimentel T, Lenci A, et al (2022) Probing for the usage of grammatical number. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 8818–8831, https://doi.org/10.18653/v1/2022.acl-long.603, URL https://aclanthology.org/2022.acl-long.603/

Lee JH, Jiralerspong T, Yu L, et al (2025) Geometric signatures of compositionality across a language model's lifetime. In: Che W, Nabende J, Shutova E, et al (eds) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vienna, Austria, pp 5292–5320, https://doi.org/10.18653/v1/2025.acl-long.265,

URL https://aclanthology.org/2025.acl-long.265/

Li Z, Jiang G, Xie H, et al (2024) Understanding and patching compositional reasoning in LLMs. In: Ku LW, Martins A, Srikumar V (eds) Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, Bangkok, Thailand, pp 9668–9688, https://doi.org/10.18653/v1/2024.findings-acl.576, URL https://aclanthology.org/2024.findings-acl.576/

Linzen T, Dupoux E, Goldberg Y (2016) Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Transactions of the Association for Computational Linguistics 4:521–535. https://doi.org/10.1162/tacl_a_00115, URL https://aclanthology.org/Q16-1037/

Liu E, Neubig G (2022) Are representations built from the ground up? an empirical examination of local composition in language models. In: Goldberg Y, Kozareva Z, Zhang Y (eds) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 9053–9073, https://doi.org/10.18653/v1/2022.emnlp-main.617, URL https://aclanthology.org/2022.emnlp-main.617/

Liu NF, Gardner M, Belinkov Y, et al (2019) Linguistic knowledge and transferability of contextual representations. In: Burstein J, Doran C, Solorio T (eds) 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). Association for Computational Linguistics, Minneapolis, Minnesota, pp 1073–1094, https://doi.org/10.18653/v1/N19-1112, URL https://aclanthology.org/N19-1112/

López-Otal M, Gracia J, Bernad J, et al (2025) Linguistic interpretability of transformer-based language models: a systematic review. arXiv preprint arXiv:250408001

Makelov A, Lange G, Geiger A, et al (2024) Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In: The Twelfth International Conference on Learning Representations, Vienna Austria, URL https://openreview.net/forum?id=Ebt7JgMHv1

Marks S, Rager C, Michaud EJ, et al (2025) Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In: The Thirteenth International Conference on Learning Representations, Singapore, URL https://openreview.net/forum?id=I4e82CIDxv

Marvin R, Linzen T (2018) Targeted syntactic evaluation of language models. In: Riloff E, Chiang D, Hockenmaier J, et al (eds) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp 1192–1202, https://doi.org/10.18653/v1/D18-1151, URL https://aclanthology.org/D18-1151/

McCoy RT, Pavlick E, Linzen T (2019) Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In: Korhonen A, Traum D, Màrquez L (eds) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp 3428–3448, https://doi.org/10.18653/v1/P19-1334, URL https://aclanthology.org/P19-1334/

McCurdy K, Soulos P, Smolensky P, et al (2024) Toward compositional behavior in neural models: A survey of current views. In: Al-Onaizan Y, Bansal M, Chen YN (eds) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, pp 9323–9339, https://doi.org/10.18653/v1/2024.emnlp-main.524, URL https://aclanthology.org/2024.emnlp-main.524/

McGrath T, Rahtz M, Kramar J, et al (2023) The hydra effect: Emergent self-repair in language model computations. arXiv preprint arXiv:230715771

McKenzie IR, Lyzhov A, Pieler MM, et al (2023) Inverse scaling: When bigger isn't better. Transactions on Machine Learning Research URL https://openreview.net/forum?id=DwgRm72GQF, featured Certification

Merchant A, Rahimtoroghi E, Pavlick E, et al (2020) What happens to bert embeddings during fine-tuning? In: Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Online, pp 33–44

Merrill W, Sabharwal A (2024) The expressive power of transformers with chain of thought. In: International Conference on Learning Representations 2024, Vienna, Austria

Miller J, Chughtai B, Saunders W (2024) Transformer circuit evaluation metrics are not robust. In: First Conference on Language Modeling, Philadelphia, PA, URL https://openreview.net/forum?id=zSf8PJyQb2

Mueller A, Xia Y, Linzen T (2022) Causal analysis of syntactic agreement neurons in multilingual language models. In: Fokkens A, Srikumar V (eds) Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), pp 95–109, https://doi.org/10.18653/v1/2022.conll-1.8, URL https://aclanthology.org/2022.conll-1.8/

Newman B, Ang KS, Gong J, et al (2021) Refining targeted syntactic evaluation of language models. In: Toutanova K, Rumshisky A, Zettlemoyer L, et al (eds) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, pp 3710–3723, https://doi.org/10.18653/v1/2021.naacl-main.290, URL https://aclanthology.org/2021.naacl-main.290/

Nieuwland MS, Kuperberg GR (2008) When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. Psychological Science 19(12):1213–1218. https://doi.org/10.1111/j.1467-9280.2008.02226.x

Oh BD, Schuler W (2023) Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? Transactions of the Association for Computational Linguistics 11:336–350. https://doi.org/10.1162/tacl_a_00548, URL https://aclanthology.org/2023.tacl-1.20/

Olsson C, Elhage N, Nanda N, et al (2022) In-context learning and induction heads. arXiv preprint arXiv:220911895

Park K, Choe YJ, Veitch V (2023) The linear representation hypothesis and the geometry of large language models. In: NeurIPS 2023 Workshop on Causal Representation Learning, New Orleans

Partee B, et al (1984) Compositionality. Varieties of formal semantics 3:281–311

Pimentel T, Valvoda J, Maudslay RH, et al (2020) Information-theoretic probing for linguistic structure. In: Jurafsky D, Chai J, Schluter N, et al (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 4609–4622, https://doi.org/10.18653/v1/2020.acl-main.420, URL https://aclanthology.org/2020.acl-main.420/

Press O, Zhang M, Min S, et al (2023) Measuring and narrowing the compositionality gap in language models. In: Bouamor H, Pino J, Bali K (eds) Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp 5687–5711, https://doi.org/10.18653/v1/2023.findings-emnlp.378, URL https://aclanthology.org/2023.findings-emnlp.378/

Qiu L, Shaw P, Pasupat P, et al (2022) Evaluating the impact of model scale for compositional generalization in semantic parsing. In: Goldberg Y, Kozareva Z, Zhang Y (eds) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 9157–9179, https://doi.org/10.18653/v1/2022.emnlp-main.624, URL https://aclanthology.org/2022.emnlp-main.624/

Räuker T, Ho A, Casper S, et al (2023) Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In: 2023 ieee conference on secure and trustworthy machine learning (satml), IEEE. IEEE Computer Society, Los Alamitos, CA, USA, pp 464–483, https://doi.org/10.1109/SaTML54575.2023.00039

Ravfogel S, Elazar Y, Gonen H, et al (2020) Null it out: Guarding protected attributes by iterative nullspace projection. In: Jurafsky D, Chai J, Schluter N, et al (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 7237–7256, https://doi.org/10.18653/v1/2020.acl-main.647, URL https://aclanthology.

org/2020.acl-main.647/

Rogers A, Kovaleva O, Rumshisky A (2020) A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics 8:842–866. https://doi.org/10.1162/tacl_a_00349, URL https://aclanthology.org/2020.tacl-1.54/

Ryu SH, Lewis R (2021) Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In: Chersoni E, Hollenstein N, Jacobs C, et al (eds) Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics. Association for Computational Linguistics, Online, pp 61–71, https://doi.org/10.18653/v1/2021.cmcl-1.6, URL https://aclanthology.org/2021.cmcl-1.6/

Saraipour K, Zhang S (2025) From indirect object identification to syllogisms: Exploring binary mechanisms in transformer circuits. In: The First Workshop on the Interplay of Model Behavior and Model Internals, Canada, URL https://openreview.net/forum?id=P8rwvwB3qH

Shivagunde N, Lialin V, Rumshisky A (2023) Larger probes tell a different story: Extending psycholinguistic datasets via in-context learning. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, pp 2094–2107, https://doi.org/10.18653/v1/2023.emnlp-main.130, URL https://aclanthology.org/2023.emnlp-main.130/

Simon PJD, Chemla E, King JR, et al (2025) Probing syntax in large language models: Successes and remaining challenges. In: Second Conference on Language Modeling, Montreal

Syed A, Rager C, Conmy A (2024) Attribution patching outperforms automated circuit discovery. In: Belinkov Y, Kim N, Jumelet J, et al (eds) Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, Miami, Florida, US, pp 407–416, https://doi.org/10.18653/v1/2024.blackboxnlp-1.25, URL https://aclanthology.org/2024.blackboxnlp-1.25/

Tenney I, Das D, Pavlick E (2019a) BERT rediscovers the classical NLP pipeline. In: Korhonen A, Traum D, Màrquez L (eds) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp 4593–4601, https://doi.org/10.18653/v1/P19-1452, URL https://aclanthology.org/P19-1452/

Tenney I, Xia P, Chen B, et al (2019b) What do you learn from context? probing for sentence structure in contextualized word representations. In: International Conference on Learning Representations, New Orleans

Tian Y, Ferguson H, Breheny R (2016) Processing negation without context–why and when we represent the positive argument. Language, Cognition and Neuroscience 31(5):683–698. https://doi.org/10.1080/23273798.2016.1140214

Truong TH, Baldwin T, Verspoor K, et al (2023) Language models are not naysayers: an analysis of language models on negation benchmarks. In: Palmer A, Camacho-collados J (eds) Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023). Association for Computational Linguistics, Toronto, Canada, pp 101–114, https://doi.org/10.18653/v1/2023.starsem-1.10, URL https://aclanthology.org/2023.starsem-1.10/

Turner A, Thiergart L, Leech G, et al (2024) Activation addition: Steering language models without optimization. Tech. rep., arXiv.org, United Kingdom, https://doi.org/10.48550/arXiv.2308.10248

Varshney N, Raj S, Mishra V, et al (2025) Investigating and addressing hallucinations of LLMs in tasks involving negation. In: Cao T, Das A, Kumarage T, et al (eds) Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025). Association for Computational Linguistics, Albuquerque, New Mexico, pp 580–598, https://doi.org/10.18653/v1/2025.trustnlp-main.37, URL https://aclanthology.org/2025.trustnlp-main.37/

Vig J, Gehrmann S, Belinkov Y, et al (2020) Investigating gender bias in language models using causal mediation analysis. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS '20

Voita E, Titov I (2020) Information-theoretic probing with minimum description length. In: Webber B, Cohn T, He Y, et al (eds) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 183–196, https://doi.org/10.18653/v1/2020.emnlp-main.14, URL https://aclanthology.org/2020.emnlp-main.14/

Wang KR, Variengien A, Conmy A, et al (2023) Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In: The Eleventh International Conference on Learning Representations, Kigali Rwanda

Warstadt A, Bowman SR (2022) What artificial neural networks can tell us about human language acquisition. In: Lappin S, Bernardy JP (eds) Algebraic Structures in Natural Language, vol 17, 1st edn. CRC Press, chap 2, p 17–60, https://doi.org/10.1201/9781003205388-2

Warstadt A, Cao Y, Grosu I, et al (2019) Investigating BERT's knowledge of language: Five analysis methods with NPIs. In: Inui K, Jiang J, Ng V, et al (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China,

pp 2877–2887, https://doi.org/10.18653/v1/D19-1286, URL https://aclanthology.org/D19-1286/

Warstadt A, Parrish A, Liu H, et al (2020) BLiMP: The benchmark of linguistic minimal pairs for English. Transactions of the Association for Computational Linguistics 8:377–392. https://doi.org/10.1162/tacl_a_00321, URL https://aclanthology.org/2020.tacl-1.25/

Warstadt A, Mueller A, Choshen L, et al (2023) Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In: Warstadt A, Mueller A, Choshen L, et al (eds) Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning. Association for Computational Linguistics, Singapore, pp 1–34, https://doi.org/10.18653/v1/2023.conll-babylm.1, URL https://aclanthology.org/2023.conll-babylm.1/

Wei J, Garrette D, Linzen T, et al (2021) Frequency effects on syntactic rule learning in transformers. In: Moens MF, Huang X, Specia L, et al (eds) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 932–948, https://doi.org/10.18653/v1/2021.emnlp-main.72, URL https://aclanthology.org/2021.emnlp-main.72/

Wei J, Wang X, Schuurmans D, et al (2022) Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS '22

Wilcox E, Vani P, Levy R (2021) A targeted assessment of incremental processing in neural language models and humans. In: Zong C, Xia F, Li W, et al (eds) 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). Association for Computational Linguistics, Online, pp 939–952, https://doi.org/10.18653/v1/2021.acl-long.76, URL https://aclanthology.org/2021.acl-long.76/

Xu Z, Shi Z, Liang Y (2024) Do large language models have compositional ability? an investigation into limitations and scalability. In: First Conference on Language Modeling, Philadelphia, PA, URL https://openreview.net/forum?id=iI1CzEhEMU

Yang S, Kassner N, Gribovskaya E, et al (2025) Do large language models perform latent multi-hop reasoning without exploiting shortcuts? In: Che W, Nabende J, Shutova E, et al (eds) Findings of the Association for Computational Linguistics: ACL 2025. Association for Computational Linguistics, Vienna, Austria, pp 3971–3992, https://doi.org/10.18653/v1/2025.findings-acl.205, URL https://aclanthology.org/2025.findings-acl.205/

Yu C, Sie R, Tedeschi N, et al (2020) Word frequency does not predict grammatical knowledge in language models. In: Webber B, Cohn T, He Y, et al (eds) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 4040–4054, https://doi.org/10.18653/v1/2020.emnlp-main.331, URL https://aclanthology.org/2020.emnlp-main.331/

Zeijlstra HH (2004) Sentential negation and negative concord. Phd dissertation, University of Amsterdam, Amsterdam, URL https://hdl.handle.net/11245/1.224908, lOT Dissertation Series 101, LOT Publications, Utrecht

Zhang Q, Wang Y, Cui J, et al (2025) Beyond interpretability: The gains of feature monosemanticity on model robustness. In: The Thirteenth International Conference on Learning Representations, Singapore

Zhang Y, Yasunaga M, Zhou Z, et al (2023) Beyond positive scaling: How negation impacts scaling trends of language models. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, Toronto, Canada, pp 7479–7498, https://doi.org/10.18653/v1/2023.findings-acl.472, URL https://aclanthology.org/2023.findings-acl.472/

Zhou D, Schärli N, Hou L, et al (2023) Least-to-most prompting enables complex reasoning in large language models. In: The Eleventh International Conference on Learning Representations, Kigali, Rwanda, URL https://openreview.net/forum?id=WZH7099tgfM