LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

LINGUISTIC INFORMATION IN TRANSFORMER REPRESENTATIONS: A CRITICAL REVIEW OF PROBING METHODS, LAYER-WISE PATTERNS, AND EVIDENCE

Samra Hameed[*1], Dr Saima Akhtar Chattha[2], Shazma Khalid[3], Mifra Hussain[4]

*[*1]MS English Linguistics, COMSATS University Lahore.*
[2]Assistant Professor, CUI Lahore
[3,4]MS English Linguistics Scholar, COMSATS University Lahore

[*1]samrahameed923@gmail.com, [2]saimachattha@cuilahore.edu.pk, [3]shazmakhalid64@gmail.com, [4]hussainmifra2@gmail.com

DOI:https://doi.org/10.5281/zenodo.22708009

Keywords
transformers, linguistic representations, probing classifiers, mechanistic interpretability, syntax, morphology, discourse

Article History
Received: 01 June 2026
Accepted: 23 August 2026
Published: 08 September 2026

Corresponding Author: *
Samra Hameed

Abstract

The new paradigm of language modeling with transformers has made the hidden states of these models with context-dependent behavior seem to contain rich linguistic information, but claims about what the hidden states contain are frequently not as well supported by evidence. This article is a structured narrative review of methods and results of linguistic feature extraction from Transformer representations. The 21 primary works that were published or posted from 2017 to June 2026 were selected for the core synthesis, after relevant and source quality assessments, peer-review, and direct support of the articles cited for the synthesis's claims. There are three valid conclusions that can be drawn. First, there is surface, lexical, morphological, syntactic, semantic and discourse information that is often accessible from the internal state, but is dependent on the specific architecture, training objective, language, task, and probe. Secondly, BERT-family encoders often show a more or less hierarchical progression from surface information in early layers, to syntactic information in middle layers and to semantic and discourse-level information in later layers, which may not be a common pipeline for all large language models. Third, the typical cross-lingual and discourse abstractions are more easily accessed in intermediate layers, while recent evidence from multi-model studies indicates that inflectional features could be linearly accessible over wide layer ranges. But the accuracy of the probe does not prove that the feature is used causally; other tasks such as control, minimum description length estimates, matched baselines, cross-model replication and tests of intervention are needed for more robust conclusions. The review does not use cross-dataset performance rankings, but uses an evidence ladder to identify decodability, selectivity, robustness and causal use.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

INTRODUCTION

Transformer-based language models have revolutionized natural language processing (NLP) by yielding remarkable results in diverse NLP tasks like text classification, question answering, named-entity recognition, machine translation, and natural language inference. Transformers leverage self-attention to capture relationship between tokens in a full sequence, and enable efficient parallel processing (Vaswani et al., 2017) compared to previous recurrent architectures. Later models like BERT showed that the contextually rich representations achievable by large scale pretraining that can transfer well to many downstream tasks are provided by bidirectional models (Devlin et al., 2019). Such advances took the focus of research from assessing the capabilities of language models to examining their internal representations of linguistic information.

One of the main issues in this regard is whether Transformer models represent known levels of linguistic structure such as lexical, morphological, syntactic, semantic, and discourse structure. Contextualized representations are not equivalent to traditional word embedding's, since the representation of a token is transformed based on its context. Empirical studies have demonstrated that such contextual dependence tends to grow with the model depth, but the geometry and reachability of information produced by the model layer differs considerably across model architectures and model training goals (Ethayarajh 2019). The hidden states of a Transformer cannot therefore be considered a uniform representation space where all the linguistic properties are equally accessible.

Layer-wise analyses have often revealed general tendencies of the distribution of linguistic information. Research on BERT-like encoders shows that the surface properties and lexical information tend to be most accessible in earlier layers, while the syntactic information is typically strongest in lower-middle or middle layers, and semantic information or task-specific information is more prominent in later layers (Liu et al., 2019; Tenney et al., 2019; Rogers et al., 2020). In parallel, structural probing has shown that some dependency-tree distance and depth can be recovered from the context representation, using learned transformations (Hewitt & Manning, 2019). The following questions have been raised in more recent research on multilingual discourse, cross-lingual alignment, inflectional morphology, semantic geometry, and potential internal model representations and human language-processing measures.

With these results, the interpretation of probing results is still methodologically challenging. High probing accuracy indicates that it is possible to extract a linguistic feature from a representation, given certain experimental conditions, but does not mean that the model uses that feature to make predictions. With a sufficiently expressive probe, a pattern may be learned that contains information that is not naturally organized or functionally used by the underlying model, but is rather a part of the diagnostic task. Lexical shortcuts, label distribution, dataset construction, tokenization, probe capacity, model architecture and fine-tuning procedures are other factors that can affect probe results. Control tasks and selectivity measures can be used to assess whether a probe is gaining meaningful knowledge or memorization of arbitrary associations (Hewitt & Liang, 2019) and minimum description length can be used to assess how efficiently a target property can be extracted (Voita & Titov, 2020). However, these methods are mostly correlational in nature.

A distinction between representation and accessibility, and between representation and causal use, is, therefore, of crucial importance. A feature can be "hidden" and easily classifiable by a linear classifier, but not part of the model's actual decision making process. Evidence for stronger functional claims include intervention-based evidence, in which the removal, modification, or control of relevant information produces a predictable change in the behavior of the model.Evidence for stronger functional claims is intervention-based evidence, in which the removal, modification, or control of relevant information

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

produces a predictable change in the behavior of the model (Ravfogel et al., 2020; Amini et al., 2023). The same applies to explanations based on focus: an understandable attention pattern may not be indicative of the features that underlie a model's predictions (Jain & Wallace, 2019). When it is claimed that a model "knows" or "understands" or "uses" a linguistic structure, this claim then must be restricted to what the method adopted allows to be properly demonstrated.

Another challenge is to generalize across models and languages. Results from English BERT are not necessarily applicable to multi-lingual, decoder-only, instruction tuned or morphologically diverse languages nor to models with an encoder-decoder architecture. Numbers in the different layers of the different types of networks are not directly comparable and the same linguistic feature can be present in different layers in different distributions depending on tokenization, training data, model size, objective and downstream adaptation. The raw scores obtained from tests of different data sets are also not directly comparable to draw a general rank order of linguistic competence without carefully matching the evaluation conditions.

In the light of this context, the current article critically examines studies on linguistic information in Transformer representations of particular interest focus on probing methodology, layer-wise patterns, methodological controls and the quality of the conclusions drawn from various types of evidence. It is structured to review the literature in a narrative way and combines 21 main publications that have been published or posted from 2017 to June 2026. The review does not create a ranking on the basis of conflicting results across studies, but rather focuses on the direction and consistency of the results, the ability to replicate across models and languages, the diagnostic methods used, and the boundary conditions of each result.

The review considers three questions: What properties of language can be recovered with confidence from Transformer models? What diagnostic procedures can support correlational or robust or causal inferences about those properties? What architectural, linguistic, methodological and evaluative properties limit comparisons and generalization? The article provides a sequence of evidence from the decodable and selective to robust and causal uses. This framework supplies a more accurate ground for assessing internal-representation studies and paves the way to distinguishing between the things a model makes available in its occult states and the things it actually employs in its language-processing states.

Literature Review

Vaswani et al., 2017 showed that the Transformer architecture has transformed NLP by replacing the traditional recurrent computation with attention mechanisms that are easily computable in parallel. Transformers skip the idea of processing tokens in a sequence over time steps and enable each token to dynamically attend to all others using self-attention mechanisms, which is able to model long-range dependencies more effectively and with significantly higher efficiency during large-scale pretraining. This architectural revolution provided the basic infrastructure for the rapid growth of the large-scale pre-trained language model, which has revolutionized the understanding and generation of languages. On the basis of this success, encoder-only models like BERT showed that deep bidirectional pretraining can be used to achieve top performance on a large range of downstream tasks while making very little modification to the model's architecture (Devlin et al., 2019). By demonstrating the successful transfer of general-purpose representations from an enormous amount of unlabeled text to a wide range of language tasks such as sentiment classification, question answering, named entity recognition, and natural language inference, BERT convincingly illustrated that these representations are effective across a wide variety of language tasks. This amazing success of such models consequently changed the primary research domain from "does the model get high benchmark accuracy?" to "how does the model get high accuracy of the benchmark?"

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

These amazing engineering feats inevitably lead to a higher education level of scientific inquiry beyond just benchmark performance: What linguistic information is being represented within the network: Where is it stored: Does it actually get used during inference? Representation analysis and interpretability are at the core of the emerging area of natural language processing and are directly related to answering these basic questions. However, it is not necessarily the case that a deep neural language model can accurately predict a language without it structuring or organising it in a way that is human-interpretable or even conventionally structured. Thus, raw performance would not be indicative of syntactic, semantic, morphological or discourse level knowledge, nor would it indicate that a model is simply making use of statistical surface correlations. More and more researchers are able to separate information that is either passively encoded or decodable from a representation vector from information that is actively and causally used to generate the final prediction of a model. This is especially important as modern neural networks can have significant amounts of redundant information, and the fact that they generate a decodable message does not necessarily mean that they use that message to perform their task.

The early representation analyses only gave some limited and very useful empirical evidence about these internal dynamics. Linear probes on the frozen contextual representations yielded highly accurate results for a large range of linguistic labeling tasks, and extensive model-by-model analyses of models such as BERT revealed that the internal hierarchy of these models often resembled a classical NLP pipeline (Liu et al., 2019; Tenney et al., 2019). In these early experiments, it was proposed that the deep network may have several layers with differing accessibility to different forms of linguistic information. In particular, lower and middle layers were more strongly linked to relatively local, surface-level or syntactic properties while higher layers were more strongly linked to abstract semantic, pragmatic and task-related properties. This kind of information proved to be very significant because it gave empirical support to the investigation of the internal organization of contextual representations, rather than assuming the whole neural network to be an impenetrable, indivisible black box (Ramzan & Kahn, 2024).

Structural probes also showed that aspects of dependency-tree geometry could be successfully recovered directly from contextual vectors using a learned linear transformation (Hewitt & Manning, 2019). Theoretically important was this body of work, since it took things beyond mere classification tasks and questioned whether there might be rich information in hidden representations about the underlying hierarchical structure of linguistic relationships. If the syntactic distances and the tree depths can be recovered from a model representation reliably then it might mean that the network deals with the organization of information in a way that is structurally similar to the traditional linguistic theory, at least in part. But there is significant methodological nuance that must be exercised in the interpretation. It is important to note that the ability to recover a particular property is evidence of the property being available under the particular mathematical assumptions of a particular probe, but it does not necessarily show that the property is actually constructed, maintained or used by the underlying language model during its standard prediction process (Qureshi, 2021)

This distinction is of crucial importance because it is possible that the regularities (or spurious lexical shortcuts, or even just the target task itself) are only slightly present in the underlying representation and that it is only a sufficiently expressive probe that can learn these regularities. A very complex probe can be very effective even if the representation itself doesn't naturally or inherently organize information in the way that is represented by the probing task.

Moreover, representations can be quite architecture-specific and highly dependent on the schemes for tokenization and fine-tuning (Belinkov & Glass, 2019; Ethayarajh, 2019; Rogers et al., 2020). For instance, the linguistic information that may be derived from a representation can vary significantly depending on

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

the models' tuning for a specific downstream application. For example, various tokenization techniques can have a significant impact on the information that is easily accessible at the token level, especially in morphologically complex languages or multi-lingual contexts.

The next conceptually relevant question relates to exactly where in the network and how it is distributed the linguistic information resides. The individual layers may seem to have fixed, discrete linguistic roles; for example, the syntax might be in one layer and the semantics in another. But, representation analysis as a kind of empirical evidence, generally favors a much more complex interpretation, though. Linguistic properties are often realized at several layers at once and they may be linguistically realized at different levels of abstraction at the same time. Features of a language may be more or less accessible through the depth of the network, but not jump to a single discrete level. Furthermore, the exact same linguistic property can be encoded in different ways in different architectures, model sizes or training languages or objectives. Therefore, it is important to take into account the nature of the observations and to consider the results as general tendencies rather than as a rule for all neural language models.

Sweeping statements that a model “knows” the syntax of a model or that attention weights capture all the information needed to make a prediction simply cannot be backed by high probe accuracy or pretty images. While attention weights can often provide meaningful and intuitive patterns, they do not necessarily directly indicate model decisions. It is important to note that a visually interpretable attention pattern does not mean that the tokens the attention focuses on are causally responsible for producing the model's output. Similarly, a representation can contain a linguistic feature that it does not explicitly employ in making inferences. These worries lead to a general move away from descriptive interpretability and toward more rigorous analysis in terms of cause and intervention. Rather than the question of whether a linguistic property can be extracted from a vector, the question can be raised whether the manipulation, removal or alteration of this property directly affects the behavior of the model.

The three-part separation of representation, accessibility, and utilization offers a very convenient framework for the assessment of various claims concerning language models. Representation analysis can tell whether information is present or recoverable, probing can tell how easy it is to extract that information, and intervention-based methods can tell much more about whether the information does, in fact, contribute to a specific prediction. None of these methodological approaches can be completely satisfactory on its own. A comprehensive and reliable analysis must thus combine various types of evidence, such as evaluation of behavior, controlled probing experiments, layer-wise analysis, and causal interventions, when appropriate.

As a whole, Transformer and Encoder architectures have led to the construction of extremely powerful models, for which the structure of their internal language is a key scientific question. What existing research has established is that there is a large amount of information about the linguistic structure in the contextual representations, but it is also clear that decodability cannot be equated with the genuine understanding or causal use. Verdicts are accordingly most convincing when they are strictly confined to the experimental technique used. This literature review reformulates such claims into testable claims, whose scope is clearly defined by the empirical evidence on which they are based. This way, the study of representation research can go beyond the dichotomous "can the model represent a linguistic property?" and become more precise: Where is the information represented? How easy or difficult is it to access? When did it emerge? Does the model rely on it when predicting? Such questions are more sophisticated: they offer a much more solid basis for the study of the real linguistic knowledge contained in modern neural language models.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

Research Questions

RQ1. What are the linguistic properties that are reliably recovered from Transformer representations and what are the commonalities across the studies that investigate these representations?

RQ2. What diagnostic procedures could help to support correlational, robust, or causal claims about those properties?

RQ3. What are the architectural, linguistic, methodological and evaluative factors that hinder comparison and generalization?

Methodology

This article is a structured narrative review - not a systematic review or meta-analysis. The evidence pool was derived from the starting bibliography of the original manuscript, which contained 45 links; it was not possible to reconstruct how the starting bibliography was compiled and this cannot be described as a systematic search, as is reported elsewhere as a limitation. All records were checked for title, author, year, publication venue, source type and direct support of the sentence or table to which the record was cited. Original publisher and conference or arXiv records were used instead of the duplicate versions, ResearchGate mirrors, project pages or generic repository links. Task-specific, non-relevant records were removed from the primary synthesis.

The remaining Evidence Base comprises 21 main publications that have been published or posted from 2017 to June 2026. The criteria for a record to be included were threefold: A record could be included either if it was directly analysing the Transformer representations, or if it was contributing a probing method or an intervention method, or if it was giving a foundational description of a relevant architecture. Peer-reviewed journal and conference papers were given priority. Four recent manuscripts were selected as they explicitly discuss the modern large language models; they are clearly indicated as preprints or accepted manuscripts and used throughout as emerging but not final evidence.

Pooled performance estimate was not computed. Probe accuracy and F1 are not necessarily mutually compatible with any of the following aspects of the studies: probe sets, label inventories, class distributions, train/test splits, model scales, and tuning budgets.Probe accuracy and F1 are not necessarily commensurable with any of the following aspects of the studies: probe sets, label inventories, class distributions, train/test splits, model scales, and tuning budgets. Only quantitative values are therefore reported if they are necessary to define a study and can be traced back to the original source. The synthesis analyzes the direction of findings, the replicability of findings across models and languages, the methodology used, and the boundary conditions of the findings, and rates claims based on the evidence ladder introduced below.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

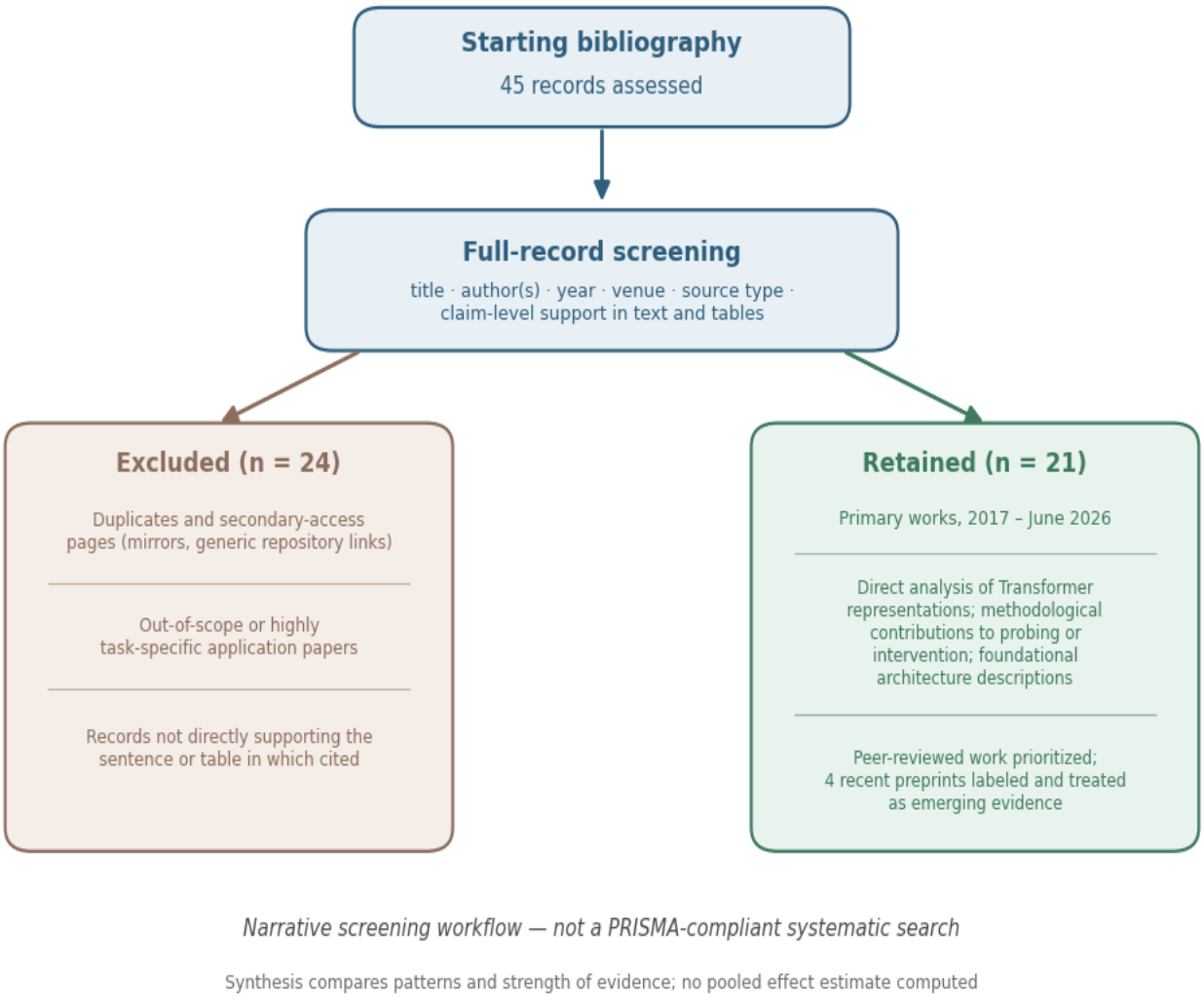

Figure 1 Screening of the Starting Bibliography

Note. The number of revisions refers to the revision process of this manuscript, not a systematic search following PRISMA principles. The excluded records were identified as duplicates, secondary access pages, applications that are out of scope, or applications that were not directly supporting the claim.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

Conceptual Framework

A Transformer takes a sequence of tokens as input and produces a sequence of hidden states. BERT embeds token, position and segment embeddings at the input of the encoder, and models the left and right context through the masked-language-modeling objective (Devlin et al., 2019). The resulting vectors are not "static" words, but are context sensitive. Ethayarajh (2019) demonstrated a layer-by-layer increase in context dependency of representation of same words and anisotropic representation spaces in the case of BERT, ELMo, and GPT-2. The baseline of cosine similarity, the linear separability and the geometric distance must therefore be interpreted in relation to appropriate baseline, not as clear measures of meaning.

What are the relationships and causal factors between multiple variables?

A diagnostic probe is a supervised model which is trained to predict a linguistic label from frozen hidden states. If it is successful, this means that the information can be retrieved by a linear decision boundary, but not that the language model would use that decision boundary when making inferences. Control tasks test whether a probe is able to merely memorize random labeling, while minimum description length (MDL) evaluates the efficiency of labels being extracted from the representation (Hewitt & Liang, 2019; Voita & Titov, 2020). The difference between the two is that causal methods manipulate either inputs, activations, or representational subspaces, and test if the behaviour of the model changes along the way as predicted. (Amini et al., 2023; Ravfogel et al., 2020). These methods deal with related, but different questions, and may yield different conclusions that might be licensed. (Table 1)

Structural Probes as Geometric Tests

Hewitt & Manning (2019) train a matrix B so that the squared Euclidean distance in the transformed space is consistent with dependency-tree distance, and the norm of a transformed vector is consistent with node depth:

$$d^B(h_i, h_j)^2 = \|B(h_i - h_j)\|_2^2 \approx d^T(w_i, w_j)$$

$$\|Bh_i\|_2^2 \approx \text{depth}^T(w_i)$$

This proves that tree relationships can be recovered by doing a low rank linear transformation. This matrix itself does not imply that it is used for dependency parsing within the network. Recent controlled experiments also reveal that linear word distance influences outputs from a structural-probe and that the syntactic depth and interactions between noun number and ungrammatical forms negatively impact the outputs (Diego-Simón et al., 2025).

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

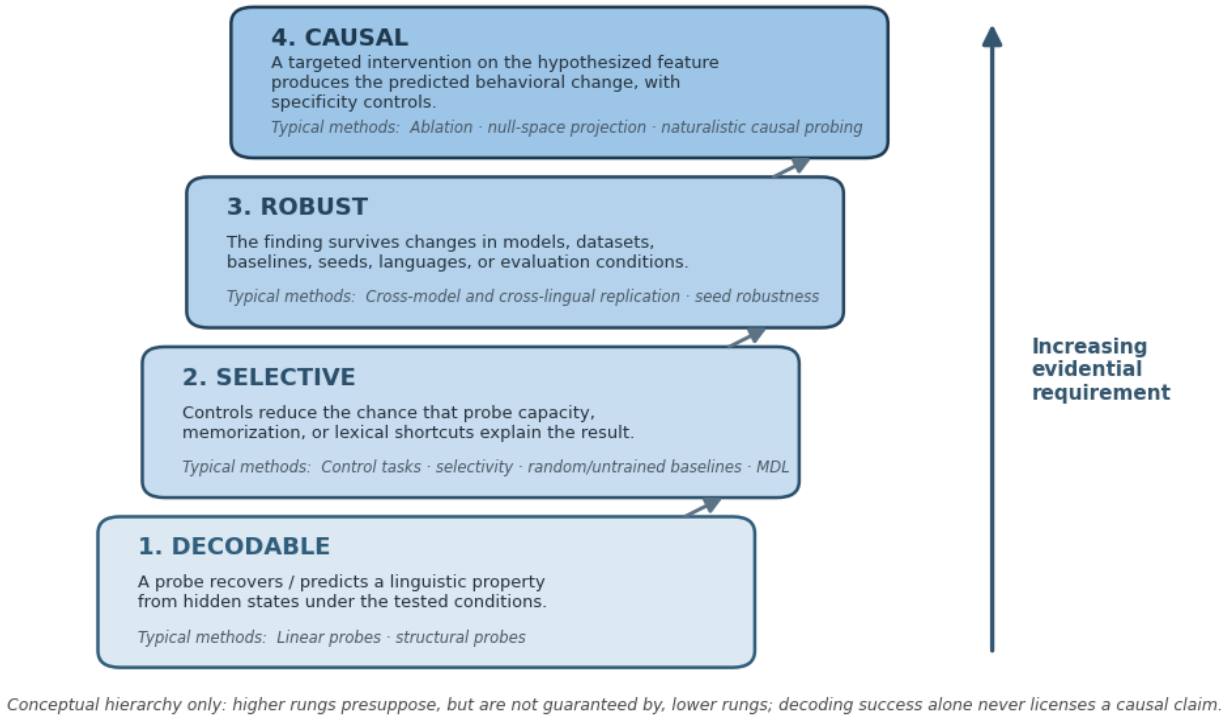

Figure 2 Evidence Ladder for Internal-Representation Claims

Note. The ladder is not a quantitative measure. Each step introduces a greater level of evidence and a causal claim almost always depends upon decoding and robustness checks that are not necessarily successful.

Findings

Mathematical tools such as surface, lexical, and morphological information. While models varied, lexical identity is more readily available and context has a diminishing impact as the distance increases (Ethayarajh, 2019). In a recent multi-model analysis that included 25 models and 6 languages, there was a convergence across layers, with inflectional features being linearly decodable across many layers, but lexical identity being best linearly decoded in the early layers and becoming less linearly accessible with depth (Li & Subramani, 2025). The same study showed that the stabilization of the structure of the inflection was earlier than the lexical identity, and that high accuracy on the probes is not necessarily a good measure of steering quality in compressed representational subspaces.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

The results corroborate a qualified conclusion: Transformers can convey morphosyntactic information in an accessible manner and they gradually "contextualize" lexical identity. They are not as strong as the claim that all the grammatical information is contained in a small linear subspace, nor as strong as the claim that it is the same for analytic languages and fusional languages and the same for agglutinative languages. Moderators that appear to be plausible include tokenization, morphological richness, model family, and probe target. Since the wide multi-model evidence is relatively new, there is still a need for cross-language/cross-architecture replication.

It is organized in syntax and layer-wise.

The best evidence in favour of layer-wise organization is provided by BERT-based studies. Tenney et al. (2019) demonstrated that, on average, parsing, named-entity recognition, semantic roles, and coreference are all localized in that order, with later layers able to update earlier layers. Liu et al. (2019) demonstrated that linguistic information is not necessarily transferable from a layer to an architecture, that is, it is not necessarily one-to-one. Based on this literature, Rogers et al. (2020) suggested that syntactic information is most easily acquired in the middle layers, whereas surface features are most easily acquired in the older layers and task-specific semantics are most easily acquired in the newer layers.

Probabilistic and architecture-bounded appropriate interpretation. It is not a language-invariant coordinate, for example, a 6-layer encoder and a 16-layer decoder do not necessarily correspond, and fine-tuning can move information. Finally, it is shown that even systematic errors can be masked by controlled syntax studies, in which strong average performance is observed on natural corpora. Wang (2016) reported that structural probes were not sensitive to word predictability, but were sensitive to nearby words and to structural depth, as did Diego-Simón et al. (2025). These findings do not rule out structural probing; they pinpoint which contrasts a probe must be able to withstand in order to serve as support for generalizations about hierarchical syntax.

The influence of semantics, discourse, and cross-lingual representation on learners' language acquisition.

There is emerging evidence for high-level discourse representations, but this may rely on the design of a label and an evaluation protocol. Eichin et al. (2025) built a global Discourse Relation Inventory (DI) and tested 23 LLMs in languages and with different annotation structures, finding that intermediate layers generalise across languages best for multilingual models. The study does not suggest that there is a common and cross-linguistic "discourse module" but shows that in the specific cross-framework classification tasks studied, intermediate representations proved to be most helpful.

This can be confirmed by independent cross-lingual evidence. More than 1000 language pairs were examined by Liu and Niehues (2025) and they identify middle layers as the most promising place to add an alignment objective to improve transfer to slot filling, machine translation and structured generation, particularly for lower-resource languages. While this evidence extends beyond passive decoding, it is generalised by the models and tasks used for training and evaluation—downstream transfer.

Recent extensions have been made to structural probing for semantic geometry. Diego-Simón et al. (2026) demonstrate that relations in minimalist semantic structure can be recovered from distances and directions in a learned polar subspace, the most discriminative signal in the middle layers. This is encouraging, but comes from a preprint which was released early for a small domain. It is not meant to be interpreted as proof of the universal applicability of a polar coordinate system for the semantics of natural language, but as proof that relational geometry is (linearly) recoverable under the conditions studied.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

Intrinsic Geometry and Cognitive Alignment

The intrinsic dimension of measures of representational complexity does not represent a complexity score of a label. Baroni et al. (2026) demonstrate that more complex psycholinguistic contrasts result in more complex intrinsic-dimension profiles across six models, and also have different depth profiles. That indicates that intrinsic dimension could be a proxy for processing requirements, but that's a preprint and doesn't specify a linguistic mechanism.

Evidence from cognitive modeling must be given equal attention and care in terms of the language used. Next word probabilities of internal layers of LLMs were correlated with human behavioural and neurophysiological data by Kuribayashi et al. (2025), and even outperformed the final layer probabilities. The earlier layers were more similar to faster measures like gaze durations and self-paced reading times, while the later layers were more similar to relatively slower measures like the N400 and MAZE responses. This does not show that Transformer layers align to human language stages at specific layers, nor that models think like humans.

Comparative Synthesis

Table 1 Diagnostic Methods and the Claims They Can Support

Method	Primary question	Essential controls	Permissible conclusion
Linear probe	Linear accessibility of a label	Random or untrained representations; lexical baselines; held-out items	Decodability only; may exploit shortcuts or probe capacity
Control-task probe	Selectivity beyond memorization	Randomized type-label mapping; matched probe capacity	Improves interpretation but remains correlational
MDL probe	Efficiency with which labels are learned	Online/ variational coding; matched data and architecture	More informative than accuracy alone; still not causal
Structural/geometric probe	Recoverable distance, depth, or relation geometry	Random baselines; controlled word distance and depth; rank tests	Sensitive to benchmark geometry and structural confounds
Intervention or ablation	Whether changing a feature changes behavior	Specificity tests; counterfactual validity; downstream outcome	Supports causal use if the intervention is targeted and faithful

Note. MDL = minimum description length. The categories summarize methodological guidance from Hewitt and Liang (2019), Voita and Titov (2020), Belinkov (2022), Amini et al. (2023), and Ravfogel et al. (2020).

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

Table 2 Qualified Synthesis of Layer-Wise Findings

Domain	Recurring pattern	Evidence base	Interpretive strength
Surface and lexical form	Often strongest in embeddings and early layers; lexical identity becomes more context-specific with depth	BERT, ELMo, GPT-2, and recent multi-model analyses	Moderate to strong; architecture and tokenization matter
Morphology	Frequently linearly decodable across broad layer ranges	Recent 25-model, six-language analysis	Promising but still concentrated in one large modern-model study
Syntax	Often most accessible in lower-middle or middle layers of BERT-family encoders	Edge probing, structural probing, controlled syntax	Strong for BERT; weaker as a universal LLM claim
Discourse/cross-lingual	Intermediate layers often provide the best aligned abstractions	23-model discourse study; 1,000+ language-pair alignment study	Convergent peer-reviewed evidence with task-specific boundaries
Cognitive fit	Different internal layers best predict different behavioral or neural measures	Reading-time, gaze, MAZE, and N400 analyses	Predictive alignment, not biological equivalence

Note. Layer labels will depend on the depth of each model and on its architecture. The terms 'early', 'mid' and 'late' should not be regarded as absolute layers for each model family.

The best conclusion across the studies is not that a single fixed pipeline has been identified, but that there is no fixed pipeline that can be identified. Instead, there are properties of representational accessibility, properties that tend to be more easily extracted at lower, middle, or upper levels in the encoders, depending on the properties involved. These overlaps exist, future layers are able to alter previous choices, and the model need not be like BERT. In a layer-wise claim, therefore, always include model family, normalized depth, probe, target label, dataset and control condition.

It is not methodologically correct to compare raw accuracy scores on different sets of data. These comparisons are only valid when the number of labels, class distributions, train/test splits, evaluation metrics, probe architectures, tuning budgets, languages or models sizes are equal. The numbers reported are "information markers" in the original experimental context and cannot be used as a general measure of linguistic competence. Tables 1 and 2 can therefore be compared to each other, but not to build the cross-dataset leaderboard.

Some methodological weaknesses and reporting standards:

Probe Capacity and Selectivity

Probe capacity is a two-edged sword: A probe that is too weak can miss out on information that is actually present, while a probe that is too strong is susceptible to learning the task itself instead of learning the

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

representation. Hewitt and Liang (2019) mention control tasks and selectivity, Voita and Titov (2020) mention minimum description length that is the amount of data or model capacity a label requires. The aim of reports should be to report on control-task performance, sample efficiency or MDL, random and untrained baselines, held out lexical items if applicable, confidence intervals, and robustness across seeds, in addition to accuracy or F1.

Attention and Explanation

While attention weights can be descriptive patterns, they should not be confused with the importance of features or causal explanations. Weak correlations between attention distributions and other importance measures have been shown by Jain and Wallace (2019), and fundamentally different attention distributions can lead to the same predictions. Claims of attention should thus be backed up with other methods of evidence like ablation, activation intervention, gradient-based analysis, or counterfactual testing.

A test of decodability and causal use

A feature does not have to be exploited by the model to be decodable, and a lack of a model's use of linearly decodable information does not imply that there are no nonlinear traces of a concept in the model. Iterative null space projection alters the linear information (Ravfogel et al., 2020), while naturalistic causal probing changes the morphosyntactic structure of the input sentence while keeping other information as stable as possible (Amini et al., 2023). Despite these approaches, it is still important to include specificity checks since interventions can influence multiple correlated features. When making causal claims, include statements about the intervention, hypothesized mediator, behavioral outcome and plausible alternative pathways.

The concept of architecture, language, and data boundaries

The results presented in English BERT-base do not imply that the same results will be obtained with multilingual encoders, instruction tuned decoders, mixture-of-experts models, or speech Transformers. The unit being probed can be changed in tokenization, the relationship between the unit and the grammatical features can be changed in morphology, and fine-tuning can reorganize representations. Multilingual studies are thus more informative if they report results in terms of language and typological property, rather than simply a macro-average. To facilitate comparison between models, the thickness of the layers should be made comparable and the data volume, the probes and the evaluation procedures should be kept comparable.

The preprint status and reproducibly aspect. The preprint status and aspect of reproducibly.

Here are a few preprints from 2025-2026 that are cited in several claims related to recent LLM. Their inclusion is pertinent and their status is to be explicit in the text and in the reference list, and the records to be amended if there are archival versions. Model checkpoints, prompts or templates, layer extraction decisions, code, random seed, dataset licenses and exclusions should also be reported by the authors of probing studies as these seemingly minor implementation decisions may have an impact on the representation analyses.

Implications and Future Research

The natural choice of the model for linguistic feature extraction should depend on the kind of evidence needed. If the objective is to provide reliable annotation or prediction, a light encoder with a well-established linear probe could be more than enough. Intervention with behavioural validation is required

for a causal scientific claim. If using for cross-lingual applications, it is recommended that multilingual evidence be used, and typologically diverse languages be used instead of English based layer maps. The diagnostic approach must be selected according to the target construct in each of the cases, and the conclusion allowed must be precision.

There are some specific areas of research that are critical. First, the basic BERT results should be reproduced with modern decoder-only, encoder-decoder, multilingual and instruction-tuned models with normalized depth and shared controls. Second, correlational probes need to be used with interventions to determine the effect of decoding the feature on prediction, generation, or transfer. Third, the study of the evaluation should extend beyond English and high-resource languages to involve systematic investigation of tokenization, morphological typology, code switching and discourse structures. If we make the kind of progress here, we shall be able to make attractive layer maps into a cumulative body of evidence for representational function.

Conclusion

Although a linguistic structure is often recoverable - as is the empirical fact - the analytical verbs used to describe this fact are not interchangeable: to passively "contain" information is not to explicitly "encoding" it, and to "encode" a feature is not necessarily to "use" it during an active model inference. The structured layer-wise accessibility of the BERT-family of encoders has been well supported by robust peer-reviewed research that shows linear recoverability of classical syntactic tree-geometry, as well as intermediate layer benefits in complex cross-lingual alignment and discourse-level tasks. Moreover, recent explorations of the potential of contemporary large language models have demonstrated the applicability of these core patterns to more sophisticated dimensions of language, such as fine-grained morphology, dimensionality of language representations and the polar semantic geometry.

Concurrently, systematic probe biases are quite apparent, as are significant architectural differences – both representational decodability and causal function are heavily dependent on the specific model, target language, downstream task, tokenization scheme, probing methodology, and evaluation conditions. For example, as models are scaled up or trained with various pretraining tasks, the distribution of linguistic features changes dynamically, showing that "localization" assumptions, which are fixed, are unable to represent the fluid nature of neural representation. As a result, it is always a scientific conclusion that is conditional: specific linguistic information is often available in the hidden states of Transformer models, but where in the hidden state, in what mathematical representation, and with which function it is used for comes always back to an explicit question that must be answered with appropriate controls specifically matched to the model, the language, the task and the specific theoretical claim.

It is not possible to simply demonstrate a problem by observing that some event is decodable and for the logic to then imply a genuine cause, in the absence of strong intervention-based evidence and appropriate specificity controls to guard against incidental artifacts or superficial lexical shortcuts. If no such rigorous verification were performed, it would be difficult to decipher what a model can express when it is called from the outside to be under the pressure of external input, and what it can express when it is not called from the outside, but is called otherwise – when it is unprompted and is in its own state.

References

Amini, A., Pimentel, T., Meister, C., & Cotterell, R. (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

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

Baroni, M., Cheng, E., de-Dios-Flores, I., & Franzon, F. (2026). Tracing the complexity profiles of different linguistic phenomena through the intrinsic dimension of LLM representations [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2601.03779

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

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

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

Diego-Simón, P. J., Chemla, E., King, J.-R., & Lakretz, Y. (2025). Probing syntax in large language models: Successes and remaining challenges [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2508.03211

Diego-Simón, P. J., Orhan, P., Chemla, E., Lakretz, Y., & King, J.-R. (2026). Polar probe linearly decodes semantic structures from LLMs [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2605.14125

Eichin, F., Liu, Y. J., Plank, B., & Hedderich, M. A. (2025). Probing LLMs for multilingual discourse generalization through a unified label set. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 18665–18684). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.915

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 55–65). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1006

Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 2733–2743). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1275

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4129–4138). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1419

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 3543–3556). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1357

Kuribayashi, T., Oseki, Y., Ben Taieb, S., Inui, K., & Baldwin, T. (2025). Large language models are human-like internally. Transactions of the Association for Computational Linguistics, 13, 1743–1766. https://doi.org/10.1162/tacl.a.58

Li, M., & Subramani, N. (2025). Model internal sleuthing: Finding lexical identity and inflectional features in modern language models [Preprint; accepted to ACL 2026]. arXiv. https://doi.org/10.48550/arXiv.2506.02132

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7

Liu, D., & Niehues, J. (2025). Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 15979–15996). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.778

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1073–1094). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1112

Qureshi, A. W. (2021). Politics as Rhetoric: A Discourse Analysis of Selected Pakistani Politicians' Press Statements. *Humanities & Social Sciences Reviews*.

Ramzan, M., & Khan, M. A. (2024). Linguistic Coherence as Cultural Insights in Prologue of the Holy Woman and Epilogue of Unmarriageable. *Contemporary Journal of Social Science Review*, *2*(04), 266-281.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7237–7256). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.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

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008). Curran Associates. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Voita, E., & Titov, I. (2020). Information-theoretic probing with minimum description length. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 183–196). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.14