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.

Relational Semantic Behaviour in Transformer Models: An Empirical and Cognitively Informed Analysis of Attention-Based Meaning Construction

Firas Ali

firas@thealis.org

Article

Keywords:

Posted Date: April 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8960735/v1

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

Additional Declarations: No competing interests reported.

Abstract

Transformer-Based Models (TBMs) have demonstrated remarkable performance across a wide range of natural language tasks, yet the nature of the semantic representations they construct remains an open question. In particular, it is unclear whether meaning in such models emerges primarily from the encoding of discrete lexical features or from relational patterns established through contextual interaction. This paper investigates the relational semantic behaviour of TBMs by analysing attention-driven transformations across varied linguistic inputs.

Using a series of controlled empirical probes, we examine how attention mechanisms mediate context-sensitive meaning construction, focusing on relational dependencies rather than token-level representations alone. The analysis highlights systematic patterns through which semantic coherence arises from interactions between elements across a sequence, suggesting that meaning in transformer models is fundamentally relational and dynamically constructed.

Rather than interpreting these findings as direct analogues of human cognition, the study adopts a cognitively informed perspective, using concepts from cognitive semantics and relational representation to interpret model behaviour. This approach allows transformer architectures to be examined as computational systems that exhibit structured, interpretable semantic organisation without assuming psychological equivalence.

The results contribute to ongoing discussions at the intersection of cognitive computation and artificial intelligence by providing empirical evidence that attention mechanisms support relational semantic organisation in large language models. These findings have implications for model interpretability, the evaluation of semantic representations, and the development of cognitively grounded analytical frameworks for contemporary AI systems (Vaswani et al., 2017).

1. Introduction

The emergence of large language models (LLMs) has intensified long-standing debates concerning the nature of machine intelligence and the possibility of artificial meaning. Systems such as transformer-based language models have demonstrated remarkable capacity for generating coherent text, performing conceptual transformations, and explaining unfamiliar ideas using metaphor and analogy. These capabilities have attracted both excitement and scepticism across academic disciplines. Within the humanities and social sciences in particular, the rapid diffusion of such systems has prompted renewed discussion about the philosophical limits of computational language.

A dominant position within this discourse characterises large language models as essentially statistical artefacts. According to this view, language models operate by identifying patterns within large textual corpora and generating outputs that reproduce those patterns without any genuine understanding of their content. The widely cited metaphor of the “stochastic parrot” encapsulates this perspective. In this formulation, language models are understood as systems that recombine fragments of language learned

from data, producing plausible sentences without possessing semantic comprehension or referential grounding. From this standpoint, apparent coherence in model output is interpreted as the by-product of probabilistic pattern matching rather than conceptual reasoning (Bender & Koller, 2020).

This critique has played an important role in tempering exaggerated claims regarding artificial intelligence. By emphasising the absence of embodiment, intentionality, and lived experience in machine systems, scholars have highlighted crucial distinctions between human cognition and computational language generation. Such distinctions are particularly important in public discourse, where anthropomorphic interpretations of artificial intelligence often obscure the technological and social limitations of current systems. The stochastic parrot critique therefore serves as a necessary reminder that linguistic fluency does not automatically imply cognitive understanding (Bender & Koller, 2020).

However, while the critique is valuable as a caution against over-interpretation, it may also oversimplify the internal dynamics of contemporary neural architectures. The assumption that statistical generation necessarily implies semantic emptiness overlooks the representational complexity produced within TBMs. Although these systems indeed rely on statistical learning, the structures generated through that learning process are not random or arbitrary. Instead, they form highly organised relational geometries within high-dimensional representational spaces. These geometries encode patterns of conceptual similarity, hierarchy, and association that influence how models process and generate language.

Transformer architectures operate through attention mechanisms that dynamically weight tokens relative to contextual relevance. During processing, each token interacts with every other token in a sequence through a series of weighted relationships. These relationships are recomputed at multiple layers within the model, allowing contextual information to reshape the representational structure continuously. Over time, repeated exposure to linguistic data produces stable patterns of alignment within this representational space. Words, concepts, and contextual cues become organised according to relational structures that reflect patterns of co-occurrence and conceptual association (Vaswani et al., 2017).

The existence of such structures does not imply that language models possess consciousness or genuine understanding in the human sense. However, it does suggest that the binary distinction between “meaningful cognition” and “mere statistical pattern matching” may be too coarse to capture the behaviour of contemporary models. Rather than asking whether machines understand language in the same way humans do, it may be more productive to ask what forms of semantic organisation emerge within computational systems.

This paper proposes that transformer-based language models exhibit a form of relational grounding. Relational grounding refers to the stabilisation of conceptual meaning through structured alignment within an internal representational space. In this framework, meaning does not depend on direct reference to external objects or experiences. Instead, semantic organisation arises through the relationships among elements within the system itself. Concepts acquire functional significance through their position within a network of associations rather than through direct perceptual grounding.

The idea that meaning can emerge from relational structure has deep roots within humanities scholarship. Structural linguistics, cognitive semantics, and distributed representation theory have all emphasised the importance of relational organisation in shaping semantic systems. Within these traditions, meaning is not treated as an intrinsic property of individual symbols but as an emergent property of the relationships between them. By applying this relational perspective to transformer architectures, it becomes possible to reinterpret the semantic behaviour of language models without attributing human-like cognition to machines (Saussure, 1916).

The central aim of this article is therefore to reframe current debates about machine meaning within the humanities. Rather than defending strong claims of artificial understanding, the paper argues that prevailing critiques underestimate the relational complexity embedded within transformer architectures. By examining the internal structures produced through attention mechanisms and contextual inference, it becomes possible to identify forms of semantic organisation that exist independently of embodiment or subjective experience (Vaswani et al., 2017).

To explore this possibility, the study combines theoretical analysis with comparative empirical probing across multiple contemporary language models. Cross-domain prompts are used to examine whether models consistently identify structural correspondences between conceptual domains. The results suggest that models frequently converge on similar relational anchors when performing metaphorical or explanatory tasks. These patterns indicate the presence of stable conceptual alignments within model representations.

Recognising relational grounding does not resolve the philosophical problem of machine understanding. Language models remain fundamentally different from human cognitive systems in numerous respects, including embodiment, intentionality, and experiential learning. Nevertheless, acknowledging the relational structures present within neural architectures allows for a more nuanced vocabulary when discussing machine meaning.

By moving beyond reductive metaphors such as the stochastic parrot, humanities scholarship can engage more productively with the realities of contemporary artificial intelligence. Such engagement requires conceptual frameworks capable of distinguishing between different forms of semantic organisation without collapsing them into simplistic binaries. Relational grounding offers one possible framework for achieving this goal (Bender & Koller, 2020)’

2. Literature Review: Meaning, Grounding, and the Limits of Reduction

2.1 The grounding problem in philosophy of mind

The question of whether artificial systems can possess meaning is rooted in a long-standing philosophical debate concerning representation and grounding. The so-called symbol grounding problem

asks how abstract symbols acquire semantic content rather than merely referring to other symbols within a closed system. Classical symbolic approaches to cognition assumed that meaningful representation required some form of connection between symbols and entities or states of affairs in the external world. Without such a connection, symbolic manipulation would remain purely formal (Harnad, 1990).

Early computational theories of mind largely adopted this symbolic perspective. Cognitive processes were described as rule-governed operations performed on symbolic representations. Meaning, in this framework, was assumed to arise through reference or interpretation, typically grounded in perception or action. The difficulty with such approaches lies in explaining how purely formal symbols become connected to the world in the first place.

Connectionist models of cognition challenged this symbolic paradigm by proposing that semantic structure could emerge from distributed patterns of activation within neural networks. Instead of representing concepts as discrete symbolic tokens, connectionist systems encode information across patterns of weights and activations distributed throughout a network. In such systems, meaning does not reside in any single component but emerges from the relationships among many interacting elements.

This distributed approach has important implications for the evaluation of artificial intelligence systems. If meaning can arise through relational structures rather than direct reference, then computational systems may exhibit forms of semantic organisation even in the absence of explicit grounding in external perception.

2.2 Structural approaches to meaning

Humanities traditions have long emphasised the relational nature of meaning. Structural linguistics, for example, proposed that linguistic signs derive significance through their position within a system of differences. In this framework, the meaning of a word is determined not by a direct correspondence with objects in the world but by its relationships with other words within the language system (Saussure, 1916).

This relational perspective has profoundly influenced modern linguistic theory. Structuralist and post-structuralist traditions both emphasised that semantic systems function through networks of association and opposition. Words acquire meaning through patterns of similarity, contrast, and contextual usage rather than through isolated reference.

Cognitive semantics extends these insights by examining how conceptual structures are organised in human cognition. Research in this area has demonstrated that metaphor plays a central role in conceptual understanding. Abstract ideas are frequently understood through mappings from more concrete domains. For example, emotional states may be described in spatial terms (“feeling down”), and social relationships may be described through physical metaphors (“close friends”) (Lakoff & Johnson, 1980).

These metaphorical mappings reveal that meaning often arises through structural correspondences between conceptual domains. Rather than representing isolated pieces of information, conceptual systems organise knowledge through relational patterns that link different domains together.

This relational view of meaning aligns closely with contemporary machine learning architectures, particularly those based on distributed representations.

2.3 Distributed representation and semantic geometry

Modern neural language models represent linguistic information through high-dimensional vectors. In these systems, words and concepts are encoded as points within a representational space. Relationships between concepts are reflected in the geometric structure of this space. Words that frequently occur in similar contexts tend to occupy nearby regions, while semantically distinct words appear further apart (Mikolov et al., 2013).

Such representational geometries allow models to capture patterns of conceptual similarity and association. For example, relationships between concepts can often be expressed through vector arithmetic, revealing latent semantic structure within the embedding space.

TBMs extend this principle through the use of attention mechanisms. Rather than relying solely on static embeddings, transformers dynamically recompute relationships between tokens during processing. Attention layers allow each token to evaluate its relevance to every other token within the sequence. This process produces context-sensitive representations that evolve as information flows through the network (Vaswani et al., 2017).

Through repeated exposure to large textual datasets, these architectures develop complex internal structures that encode patterns of conceptual association across domains. Although these representations arise through statistical learning, they exhibit stability and organisation that resemble semantic systems.

Consequently, statistical learning should not be interpreted as the absence of structure. On the contrary, statistical processes operating within large neural networks frequently generate highly organised representational geometries.

2.4 The “stochastic parrot” critique

Despite these developments, scepticism regarding machine meaning remains widespread within humanities scholarship. One of the most influential critiques characterises large language models as “stochastic parrots.” According to this argument, language models generate text by reproducing statistical patterns found in their training data rather than by reasoning about meaning (Bender & Koller, 2020).

This critique highlights several legitimate concerns. Large language models are trained on massive datasets that often contain biases, inaccuracies, and socially problematic content. Because the models

learn statistical correlations within this data, they may reproduce or amplify these issues in their outputs. Additionally, the scale and opacity of modern training datasets make it difficult to determine precisely how specific outputs arise.

Furthermore, language models lack embodiment and sensory experience. Unlike humans, they do not interact directly with the physical world. This absence of experiential grounding raises important questions about whether machine-generated language can genuinely represent knowledge about reality.

However, the stochastic parrot metaphor may also obscure important aspects of model behaviour. By emphasising the statistical nature of language generation, the metaphor suggests that model outputs are essentially random recombinations of linguistic fragments. In practice, however, TBMs exhibit systematic patterns of conceptual organisation that extend beyond simple word frequency correlations (Bender & Koller, 2020).

In particular, models often produce coherent metaphorical explanations and cross-domain analogies that rely on identifying structural correspondences between conceptual domains. Such behaviour suggests the presence of underlying relational structures that guide language generation.

Recognising these structures does not imply that models possess human-like understanding. Nevertheless, it does indicate that the representational systems within neural networks are more organised than the stochastic parrot metaphor might suggest.

2.5 Toward a relational perspective on machine meaning

The debate over machine meaning is therefore often framed as a binary opposition between two positions. On one side are claims that language models exhibit forms of understanding or reasoning. On the other side are arguments that models merely reproduce statistical patterns without semantic content.

Both positions risk oversimplification. The first may attribute excessive cognitive capacities to artificial systems, while the second may underestimate the representational complexity produced by neural architectures.

A more productive approach may lie between these extremes. Instead of asking whether machines truly understand language, it may be more useful to examine the forms of semantic organisation that arise within computational systems.

From this perspective, meaning can be analysed as a property of relational structures within representational spaces. If conceptual relationships within a system exhibit stability and coherence across contexts, then the system may be said to possess a form of semantic organisation even without experiential grounding.

This idea forms the basis of the concept introduced in this paper: relational grounding. Rather than requiring direct reference to the external world, relational grounding refers to the stabilisation of

conceptual meaning through structured alignment within a representational system.

By examining the relational patterns that emerge within transformer architectures, it becomes possible to evaluate machine meaning in a way that avoids both anthropomorphic exaggeration and reductive dismissal.

3. Theoretical Framework: Relational Grounding

The preceding discussion suggests that debates concerning machine meaning often rely on overly simplified conceptual categories. Language models are frequently described either as possessing forms of understanding comparable to human cognition or as purely statistical mechanisms devoid of semantic structure. Both characterisations obscure important features of contemporary neural architectures. A more precise account requires a framework capable of describing the internal organisation of representational systems without attributing human-like cognition to machines.

This paper proposes such a framework in the form of relational grounding. The concept refers to the stabilisation of semantic relationships within an internal representational system. In this view, meaning does not arise through direct reference to external objects or through subjective experience but instead, it emerges through patterns of alignment among elements within the system itself.

Relational grounding therefore differs from classical notions of grounding that emphasise perceptual interaction with the physical world. Traditional grounding theories assume that symbols must ultimately connect to sensory experience in order to acquire meaning. While this assumption is well justified for biological cognition, it may not fully capture the representational dynamics of artificial neural systems.

TBMs operate within large vector spaces in which linguistic tokens, phrases, and concepts are encoded as high-dimensional representations. Within these spaces, relationships between elements are determined by patterns of co-occurrence, contextual similarity, and learned associations. Rather than mapping directly to objects in the world, these representations map to positions within a relational geometry formed through training.

The core claim of relational grounding is that this geometry itself constitutes a form of semantic organisation. Concepts acquire functional meaning through their position within a network of relationships. In other words, meaning emerges not from isolated symbols but from the structured configuration of connections among them.

This perspective draws upon earlier relational approaches to meaning found within both linguistic and cognitive traditions. Structural linguistics emphasised that signs acquire meaning through differences within a system. Similarly, cognitive semantics has shown that conceptual structures often depend on relational mappings between domains, such as metaphorical correspondences between physical and abstract concepts (Saussure, 1916).

In distributed neural architectures, relational structures emerge through patterns of weighted interactions between units. During training, models repeatedly adjust these weights to minimise prediction error across large datasets. Over time, this process produces stable configurations within the representational space. Words and concepts that share contextual patterns become located near one another, forming clusters that reflect semantic similarity.

Transformer architectures extend this process through attention mechanisms. Attention allows tokens within a sequence to dynamically evaluate their relevance to one another during processing. Each token generates queries, keys, and values that interact with other tokens to determine the degree of influence they exert on the final representation (Vaswani et al., 2017).

Through multiple layers of attention, the model continually recomputes relationships among tokens, producing context-sensitive semantic structures. These structures evolve throughout the processing pipeline, allowing information from distant parts of a sentence or passage to influence interpretation.

Importantly, attention mechanisms do not merely record statistical frequency. They produce weighted relationships that reflect contextual significance. Certain tokens become central to the interpretation of a sequence, while others play supporting roles. This process results in patterns of conceptual salience that resemble aspects of human linguistic interpretation.

Within this framework, relational grounding can operate through three observable characteristics:

- Convergence across models. If independent language models trained on similar corpora consistently identify the same conceptual relationships when interpreting prompts, this suggests the presence of stable relational structures within the representational space.
- Structural coherence in cross-domain mapping. When models explain one domain through analogy with another, successful explanations require identifying structural correspondences between the domains. If models repeatedly produce explanations that rely on similar conceptual anchors, such as flow, tension, constraint, or balance, this indicates that relational structures are guiding their responses.
- Stability under prompt variation. A relationally grounded system should maintain conceptual alignment even when prompts are phrased differently. If multiple formulations of a question produce explanations organised around similar conceptual structures, this suggests that the model is relying on stable internal representations rather than random recombination of phrases.

These criteria do not demonstrate that language models possess human-like understanding but, instead, they provide evidence that model outputs are influenced by internal relational structures rather than purely by local statistical patterns.

Relational grounding therefore occupies an intermediate conceptual position. It acknowledges that transformer models lack embodiment, intentionality, and subjective experience. At the same time, it

recognises that these systems generate outputs shaped by structured semantic relationships embedded within their representational architecture.

This intermediate position allows the discussion of machine meaning to move beyond binary categories. Instead of asking whether machines truly understand language, the focus shifts toward identifying the types of representational organisation that computational systems exhibit.

Such a shift has important implications for interdisciplinary scholarship. By describing semantic organisation in terms of relational structures, it becomes possible to evaluate artificial language systems without either anthropomorphising them or dismissing them as meaningless statistical artefacts.

Relational grounding therefore provides a conceptual bridge between machine learning research and humanities scholarship. It allows the representational dynamics of neural architectures to be analysed using theoretical tools developed within linguistics, philosophy, and cognitive science.

The next section examines whether empirical evidence from contemporary language models supports the presence of relational grounding as defined above.

4. Methodology

The purpose of the empirical component of this study is not to demonstrate that language models possess human-like understanding, but rather to investigate whether their behaviour exhibits patterns consistent with the concept of relational grounding introduced in the previous section. Specifically, the methodology examines whether TBMs display stable conceptual alignments when performing cross-domain explanatory tasks.

The study adopts a comparative probe-based approach, a methodology commonly used in computational linguistics to investigate internal representational structures of neural language models. Instead of analysing training data directly, the probe-based method examines how models respond to carefully designed prompts that require conceptual transformation or explanation.

Such tasks are particularly suitable for investigating relational structure because they require the model to identify correspondences between different conceptual domains rather than merely recalling memorised phrases. When a model explains one domain in terms of another, it must implicitly identify structural relationships that allow the analogy to function.

4.1 Model selection

To examine whether relational structures are consistent across architectures, the study uses multiple contemporary TBM**s**. The selection includes six widely used models representing different training pipelines and architectural refinements.

Although the precise configuration of each model differs, they all share the fundamental transformer architecture based on multi-head attention mechanisms. This architectural similarity makes it possible to investigate whether relational patterns emerge consistently across systems trained using similar underlying principles.

Using multiple models helps to address the possibility that observed behaviours might be idiosyncratic to a single system. If relational structures appear consistently across models, this increases confidence that such patterns arise from the architecture itself rather than from particular training artefacts.

4.2 Prompt design

Prompts were designed to elicit explanations requiring **cross-domain conceptual mapping**. These tasks were chosen because metaphorical explanation is known to depend on identifying structural correspondences between domains.

Three categories of prompts were used:

1. **Scientific-to-metaphorical explanations**
   Example: “Explain electrical resistance using a musical metaphor.”
2. **Biological-to-mechanical explanations**
   Example: “Explain neuronal inhibition using the metaphor of traffic flow.”
3. **Physical-to-aesthetic explanations**
   Example: “Explain gravity as if describing harmony in music.”

These prompts require the model to map structural relationships between distinct domains rather than merely summarising factual knowledge. Successful responses typically rely on conceptual anchors such as flow, tension, balance, constraint, attraction, or equilibrium.

Multiple variations of each prompt were also created to test the stability of conceptual alignment under changes in wording.

4.3 Evaluation criteria

Two independent evaluators assessed the model responses using qualitative criteria designed to identify evidence of relational grounding.

The evaluation focused on three dimensions:

- S**emantic fidelity**, referring to whether the explanation preserved essential relationships present in the original concept. For example, an explanation of electrical resistance should still convey the idea of opposition to flow even when expressed through metaphor.

- C**onceptual coherence**, referring to whether the explanation maintained internally consistent relationships between elements of the metaphorical domain.
- S**tructural alignment**, referring to whether the explanation identified meaningful correspondences between the source and target domains.

These criteria were chosen because relational grounding predicts that explanations will rely on stable conceptual anchors rather than arbitrary linguistic substitutions.

4.4 Embedding similarity analysis

In addition to qualitative evaluation, embedding similarity measures were used to estimate conceptual alignment between domains. Word and phrase embeddings generated by the models were compared using *cosine similarity* metrics to determine whether metaphorical mappings corresponded to measurable proximity within representational space.

For example, terms associated with electrical flow might exhibit measurable similarity to terms associated with musical tension when used within explanatory contexts. Such similarity would suggest that the model's internal representations encode relational associations linking the two domains.

While embedding similarity cannot directly demonstrate semantic understanding, it can provide evidence that conceptual relationships are encoded within the model's representational structure.

4.5 Attention pattern inspection

Transformer models rely on attention mechanisms to determine how information flows during processing. By examining attention distributions during generation, it is possible to identify which tokens the model treats as conceptually salient (Vaswani et al., 2017).

Attention visualisation tools were therefore used to inspect patterns of token interaction within selected outputs. The analysis focused on whether attention converged on tokens corresponding to conceptual anchors within the explanation.

For instance, when producing a metaphorical explanation of resistance using musical language, the model might assign strong attention weights to terms such as "tension," "flow," or "constraint." Such patterns would indicate that relational structures are guiding the explanation.

4.6 Limitations

It is important to emphasise that the methodology does not claim to measure machine understanding. The goal is not to determine whether language models genuinely comprehend the concepts they

describe. Instead, the aim is to examine whether their outputs exhibit patterns consistent with the theoretical notion of relational grounding.

Because the analysis focuses on observable behaviour rather than internal representations alone, the conclusions should be interpreted cautiously. Language models may produce coherent explanations through mechanisms that differ significantly from human reasoning processes.

Nevertheless, identifying consistent relational patterns across multiple models can provide useful evidence regarding the types of representational structures that emerge within transformer architectures.

5. Results

The responses generated by the models across the prompt set revealed a number of consistent patterns that are relevant to the concept of relational grounding. Although individual outputs varied in style, phrasing, and level of detail, the underlying conceptual structures exhibited notable similarities across models and prompt variations.

The analysis focused on three dimensions corresponding to the criteria introduced in the theoretical framework: convergence of conceptual anchors across models, structural coherence in cross-domain mapping, and stability of relational alignment under prompt variation.

5.1 Convergence across models

One of the most striking observations was the degree of conceptual convergence across different language models. Despite differences in training, parameter counts, and fine-tuning procedures, the models frequently produced explanations organised around similar conceptual anchors.

For example, when asked to explain electrical resistance through a musical metaphor, nearly all models relied on the conceptual pair of flow and tension. Electrical current was typically described as analogous to a musical line or melodic movement, while resistance was described as an opposing tension or constraint affecting that flow. Some models compared resistance to friction within musical phrasing, while others framed it as the controlled pressure that shapes the expressive contour of a melody.

Although the specific metaphors varied, the relational structure remained consistent: an active flow encountering an opposing force that regulates or shapes movement. This structural alignment appeared across multiple models and across several variations of the prompt.

A similar pattern emerged in explanations involving neuronal inhibition. When asked to explain inhibitory neural signals using traffic metaphors, most models described inhibition as a regulatory mechanism analogous to traffic signals, roadblocks, or controlled intersections. Again, the central relational structure involved the regulation of flow through constraint or modulation.

These examples suggest that models consistently identify particular conceptual relationships, such as flow, constraint, balance, tension, and regulation,when performing cross-domain explanations.

5.2 Structural coherence in metaphorical mapping

The second major finding concerns the coherence of the metaphorical mappings produced by the models. In most cases, explanations preserved the relational structure of the original concept while translating it into a different conceptual domain.

For instance, when explaining gravitational attraction through musical metaphor, models frequently described gravity as a kind of harmonic pull between notes within a musical system. Some explanations compared gravitational attraction to the tendency of musical phrases to resolve toward a tonal centre. Others described it as the attraction between complementary tones within harmonic structures.

Despite variation in language, the relational mapping remained clear: gravity was represented as a force that draws elements together within a structured system. This relational structure parallels the physical phenomenon in which gravitational forces pull objects toward one another.

Such examples indicate that models are not simply substituting words from different domains but are attempting to preserve relational correspondences between the structures of the two domains.

In several cases, models explicitly articulated these correspondences by explaining how the metaphor mapped onto the original concept. This suggests that the relational mapping was not merely incidental but played a functional role in constructing the explanation.

5.3 Stability under prompt variation

The third dimension of analysis examined whether relational structures remained stable when prompts were phrased differently. For each conceptual task, multiple variations of the prompt were presented to the models. These variations differed in wording but requested essentially the same type of explanation.

The results indicate that conceptual anchors remained largely stable across prompt variations. For example, when explaining electrical resistance, some prompts used the phrase “musical metaphor,” while others asked the model to “describe electrical resistance using musical ideas.” Despite these differences, the resulting explanations consistently relied on similar relational concepts such as tension, constraint, and controlled flow.

Similarly, explanations involving neuronal inhibition consistently framed inhibition as a regulatory mechanism that modulates or limits activity. Even when prompts used alternative phrasings such as “control” or “restriction,” the models tended to organise their responses around similar conceptual relationships.

This stability suggests that the relational structures guiding the explanations were not tied to specific prompt formulations but reflected deeper patterns within the models’ representational systems.

5.4 Embedding similarity patterns

Embedding similarity analysis provided additional evidence for relational alignment between domains. Terms associated with particular conceptual anchors frequently exhibited measurable proximity within the embedding space when used in explanatory contexts.

For example, terms related to flow in electrical explanations appeared in contexts closely associated with musical movement or phrasing. Similarly, terms related to constraint or tension showed similarity across multiple metaphorical domains.

Although embedding similarity cannot by itself demonstrate semantic understanding, the presence of such patterns supports the idea that relational associations between domains are encoded within the models' representational geometry.

5.5 Attention patterns and conceptual salience

Inspection of attention patterns revealed that models frequently assigned high attention weights to tokens corresponding to conceptual anchors within explanations. Words such as "flow," "tension," "balance," and "constraint" often served as central nodes within the generated text.

These tokens appeared to function as organising principles around which the rest of the explanation was structured. For example, in several outputs explaining resistance through musical metaphor, attention weights concentrated on the term "tension," with surrounding tokens elaborating on the nature of that tension within the metaphorical domain.

Such patterns suggest that relational structures play a role in guiding the generation process. Rather than randomly combining linguistic fragments, the models appear to organise explanations around conceptual anchors that maintain coherence across the mapping.

5.6 Summary of findings

Taken together, the results suggest that TBMs frequently produce explanations that rely on stable relational structures. These structures appear across multiple models, remain consistent under prompt variation, and preserve conceptual relationships when mapping between domains.

While these findings do not demonstrate that language models possess genuine understanding, they do indicate that model outputs are influenced by structured semantic relationships embedded within the representational architecture.

These observations provide empirical support for the concept of relational grounding introduced earlier. The next section examines the philosophical implications of these findings for debates concerning machine meaning.

6. Philosophical Analysis: Rethinking the "Stochastic Parrot"

The empirical observations presented in the previous section do not demonstrate that language models possess understanding in the human sense. TBMs clearly lack many characteristics associated with human cognition, including embodiment, subjective experience, intentional agency, and perceptual interaction with the world. Nevertheless, the results raise important questions regarding how machine-generated language should be interpreted within philosophical discussions of meaning.

The dominant humanities critique of large language models often relies on the metaphor of the "stochastic parrot." According to this argument, language models generate text by statistically recombining patterns found in training data rather than by reasoning about meaning. The metaphor emphasises that the models do not understand the content of their outputs and that their apparent fluency should not be confused with genuine comprehension (Bender & Koller, 2020).

This critique has played an important role in tempering overly optimistic narratives surrounding artificial intelligence. By highlighting the absence of embodiment and experiential grounding in language models, scholars have drawn attention to fundamental differences between machine systems and human cognition. These differences are particularly significant in discussions of knowledge, responsibility, and social impact.

However, the stochastic parrot metaphor may also obscure certain aspects of contemporary neural architectures. By emphasising the statistical nature of language generation, the metaphor suggests that model outputs are essentially random recombinations of linguistic fragments. Such a description risks overlooking the highly structured representational systems that emerge during training.

Statistical learning in neural networks does not produce arbitrary outputs. Instead, it generates complex patterns of relational organisation within high-dimensional representational spaces. Concepts become linked through networks of associations shaped by contextual usage patterns across vast textual datasets. These networks form the basis of the model's ability to generate coherent explanations, analogies, and conceptual transformations.

From this perspective, the opposition between "statistical pattern matching" and "meaningful representation" may be misleading. Statistical learning processes can produce structured semantic systems even in the absence of explicit symbolic rules or perceptual grounding.

The notion of relational grounding provides a way of conceptualising this phenomenon. Under this framework, semantic organisation arises through the relationships among elements within a representational system rather than through direct correspondence with external objects. Meaning is therefore treated as an emergent property of relational structure rather than as a property attached to individual symbols.

This idea is not entirely new within the humanities. Structuralist approaches to language long emphasised that signs acquire meaning through their position within a system of differences. Similarly,

cognitive semantic theories have shown that conceptual understanding often relies on relational mappings between domains, particularly in metaphorical reasoning.

What is novel in the context of contemporary artificial intelligence is the scale and complexity with which such relational structures can be generated through statistical learning. Transformer-based architectures create representational spaces containing billions of parameters, allowing highly intricate patterns of conceptual association to emerge during training.

The results presented in the previous section suggest that these relational structures influence how language models construct explanations. When asked to explain concepts using metaphor or analogy, the models consistently identify particular conceptual anchors that serve as organising principles across domains.

This behaviour is difficult to explain purely in terms of surface-level phrase recombination. Instead, it suggests that the models rely on underlying relational structures that link different conceptual domains together.

Recognising the presence of such structures does not require attributing human-like cognition to machines. Language models remain fundamentally different from biological cognitive systems in several crucial respects.

First, language models lack embodiment. Human cognition is deeply shaped by physical interaction with the world, sensory perception, and motor activity. These factors influence how humans conceptualise space, movement, causation, and agency. Language models, by contrast, operate entirely within textual representations.

Second, language models lack intentionality. Human linguistic communication is guided by goals, intentions, and social contexts. Speakers choose words in order to achieve particular communicative outcomes. Language models do not possess such intentions; they generate text based on statistical probabilities derived from training data.

Third, language models lack experiential learning. Humans continuously update their understanding of the world through interaction and feedback. Language models, in contrast, learn during training and then operate within the constraints of their learned parameters.

These differences mean that machine-generated language should not be interpreted as evidence of human-like understanding. Nevertheless, the absence of human cognition does not imply the absence of semantic organisation.

The concept of relational grounding therefore occupies a middle position between two extremes. On one side are interpretations that attribute excessive cognitive capacities to artificial systems, treating language models as if they possessed genuine understanding or reasoning abilities. On the other side are reductive accounts that treat model outputs as meaningless statistical artefacts.

Both positions risk misunderstanding the nature of contemporary neural architectures. The first exaggerates the capabilities of these systems, while the second underestimates the representational complexity they exhibit.

Relational grounding provides a conceptual framework that avoids both pitfalls. It acknowledges that language models lack many features associated with human cognition while recognising that their internal structures nevertheless encode patterns of semantic organisation.

This framework also has implications for interdisciplinary scholarship. Debates about artificial intelligence frequently involve scholars from computer science, philosophy, linguistics, sociology, and cultural studies. Each discipline brings its own conceptual vocabulary to the discussion, which can sometimes lead to misunderstandings about how machine systems actually function.

By describing semantic organisation in terms of relational structures within representational systems, relational grounding offers a vocabulary that can bridge these disciplinary perspectives. It allows humanities scholars to critique artificial intelligence systems without relying on metaphors that oversimplify their internal dynamics.

Ultimately, the goal of this analysis is not to resolve the philosophical problem of machine understanding. Instead, it aims to clarify the conceptual landscape within which such debates take place. Recognising the relational structures present within transformer architectures allows discussions of machine meaning to move beyond binary categories and toward more nuanced theoretical frameworks.

7. Implications for Humanities and Social Sciences

The concept of relational grounding has several implications for the ongoing discussions of artificial intelligence within the humanities and social sciences. These implications extend beyond the technical functioning of language models and touch upon broader questions concerning meaning, interpretation, knowledge production, and interdisciplinary dialogue.

7.1 Rethinking semantic categories in AI discourse

One immediate implication concerns the conceptual vocabulary used to describe artificial language systems. Much of the public debate about AI relies on binary distinctions between genuine understanding and meaningless statistical output. While such distinctions can be rhetorically effective, they often obscure the representational complexity of modern neural architectures.

Relational grounding suggests that semantic organisation may exist in forms that differ from human cognition but are nevertheless structured and coherent. Recognising this possibility encourages scholars to move beyond simplified dichotomies when analysing artificial systems. Instead of asking

whether machines truly understand language, discussions can focus on identifying the types of representational structures that computational systems exhibit.

This shift in emphasis allows debates about AI to become more analytically precise. It avoids the tendency to oscillate between exaggerated claims of artificial intelligence and dismissive portrayals of machine output as purely random.

7.2 Implications for epistemology

The idea of relational grounding also intersects with philosophical discussions concerning knowledge and representation. If meaning can arise through relational structures within representational systems, then the traditional requirement that knowledge must be grounded in direct sensory experience may need reconsideration when applied to artificial systems.

This does not imply that machine-generated knowledge is equivalent to human knowledge. Human cognition is embedded in bodily experience, cultural context, and social interaction. However, recognising relational semantic organisation in machine systems raises interesting questions about the nature of representation itself.

In particular, it suggests that conceptual structure can emerge through relational patterns even in the absence of physical embodiment. This observation aligns with certain strands of philosophical thought that emphasise the systemic character of meaning rather than its dependence on individual experience.

From this perspective, language models may be understood as systems that organise conceptual relationships within a symbolic environment rather than as agents that possess knowledge in the human sense.

7.3 Implications for interpretability research

The concept of relational grounding may also contribute to ongoing efforts in artificial intelligence research to improve model interpretability. Much work in this area focuses on understanding how neural networks process information internally. Techniques such as attention visualisation, probing classifiers, and embedding analysis are commonly used to examine the structure of model representations.

By framing these representations in terms of relational grounding, it becomes possible to interpret model behaviour as the result of structured conceptual alignments rather than as opaque statistical processes. This perspective encourages researchers to examine how conceptual anchors emerge within representational spaces and how these anchors influence model outputs.

Understanding these relational structures may help improve transparency in AI systems, particularly in applications where explanation and interpretability are important.

7.4 Implications for interdisciplinary dialogue

Debates about artificial intelligence often involve participants from very different intellectual traditions. Computer scientists tend to describe language models in terms of algorithms, training data, and optimisation processes. Scholars in the humanities, by contrast, often approach AI through philosophical, cultural, and ethical frameworks.

These differing perspectives sometimes lead to misunderstandings. Technical descriptions of neural architectures may appear opaque or overly reductionist to humanities scholars, while philosophical critiques may seem detached from the practical realities of machine learning.

Relational grounding offers a conceptual framework that can help bridge these perspectives. By describing machine meaning in terms of relational structures, a concept familiar to linguistics, philosophy, and cognitive science, the framework creates a shared vocabulary through which scholars from different disciplines can discuss artificial intelligence.

This does not eliminate disagreements about the interpretation or significance of AI systems. However, it allows such disagreements to be grounded in a more accurate understanding of the representational dynamics involved.

7.5 Implications for responsible AI discourse

Finally, recognising relational grounding may also influence discussions about the ethical and social implications of artificial intelligence. Concerns about bias, misinformation, and misuse remain central to debates about language models. These concerns are not diminished by recognising the representational complexity of neural architectures.

However, ethical discussions benefit from accurate descriptions of the systems under consideration. If language models are described as purely random or meaningless statistical devices, it becomes difficult to explain why they can produce coherent explanations, persuasive arguments, or plausible narratives.

Acknowledging the relational structures within these systems provides a more realistic foundation for evaluating their societal impact. It allows scholars to examine how conceptual patterns learned from training data influence the outputs produced by models and how these patterns may reflect or amplify biases present in the data.

In this way, relational grounding contributes to a more nuanced understanding of both the capabilities and limitations of contemporary artificial intelligence.

8. Conclusion

The rapid development of large language models has generated intense debate concerning the nature of machine meaning. Within the humanities and social sciences, these systems are frequently described through metaphors that emphasise their statistical nature, most notably the characterisation of language models as “stochastic parrots.” Such critiques have played an important role in countering

exaggerated claims regarding artificial intelligence and reminding scholars of the fundamental differences between computational systems and human cognition.

At the same time, the metaphor of the stochastic parrot may oversimplify the representational dynamics of contemporary neural architectures. While transformer-based language models indeed rely on statistical learning and lack many features associated with human understanding, including embodiment, intentionality, and experiential learning, their internal representational structures exhibit a level of organisation that cannot easily be reduced to random phrase recombination.

This paper has proposed the concept of relational grounding as a framework for describing this organisation. Rather than treating meaning as dependent upon direct reference to external objects or subjective experience, relational grounding defines semantic structure in terms of the relationships among elements within a representational system. Within transformer-based language models, these relationships emerge through patterns of contextual alignment encoded in high-dimensional representational spaces.

The empirical observations presented in this study suggest that language models frequently rely on stable relational structures when performing cross-domain explanatory tasks. Across multiple models and prompt variations, explanations tended to converge around similar conceptual anchors, such as flow, tension, balance, constraint, and regulation. These anchors appeared to function as organising principles guiding the construction of metaphorical explanations.

While these findings do not demonstrate that language models possess genuine understanding, they do indicate that model outputs are influenced by structured conceptual relationships embedded within the representational architecture. This observation challenges the idea that machine-generated language can be adequately described as purely stochastic recombination.

Recognising relational grounding therefore allows discussions of machine meaning to move beyond a simple opposition between understanding and statistical pattern matching. Instead of asking whether machines truly understand language in the human sense, scholars can examine the kinds of representational structures that emerge within computational systems.

Such an approach has several advantages for interdisciplinary scholarship. It allows humanities scholars to critique artificial intelligence systems without relying on metaphors that oversimplify their internal dynamics. At the same time, it provides a conceptual vocabulary that aligns more closely with the technical realities of neural language models.

The concept of relational grounding also highlights the importance of relational structures in semantic systems more generally. Meaning, whether in human cognition or artificial systems, may depend less on direct reference to isolated objects than on the organisation of relationships within complex representational networks.

Future research may extend this framework by examining relational structures within language models using more detailed analytical techniques, including large-scale embedding analysis, attention pattern studies, and systematic probing across conceptual domains. Such investigations could further clarify the extent to which relational grounding shapes the behaviour of artificial language systems.

Ultimately, the goal of this analysis is not to resolve the philosophical question of machine understanding. Rather, it aims to contribute to a more precise conceptual vocabulary for discussing artificial intelligence within the humanities and social sciences. By recognising the relational structures present within contemporary neural architectures, scholars can engage more productively with the emerging landscape of artificial language technologies.

Declarations

This paper has received no funding.

Data Availability: Data sharing is not applicable to this research as no data were generated or analysed.

There are no competing interests to be declared.

References

1. Bender EM, Koller A (2020) Climbing towards NLU: On meaning, form, and understanding in the age of data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5185–5198
2. Harnad S (1990) The symbol grounding problem. Phys D 42(1–3):335–346
3. Lakoff G, Johnson M (1980) Metaphors We Live By. University of Chicago Press
4. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. Proceedings of ICLR
5. de Saussure F (1916) Course in General Linguistics
6. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser Ł, Polosukhin I (2017) Attention Is All You Need. Advances in Neural Information Processing Systems