Context-Dependent Semantics in Large Language Models: A Case for Transformer Architectures

Felix Borck

Goethe Universität Frankfurt am Main
felix.borck@googlemail.com

Abstract. This paper explores how the decoder-only Transformer Architecture (TA) in Large Language Models (LLMs) can be viewed as an operationalization of context-dependent semantic theories. Drawing upon David Braun's critique of David Kaplan's stable character functions, the discussion illustrates how TA's self-attention and embedding mechanisms align with a relational model of meaning, where words derive their sense from contextual cues rather than from static mappings. Empirical findings on self-similarity and anisotropy in GPT-based embeddings reinforce the conclusion that context influences meaning at a fundamental level, challenging long-standing assumptions about language. By integrating philosophical perspectives with detailed technical analyses, this paper highlights how the success of LLMs provides an impetus to reevaluate foundational theories of semantics, ultimately suggesting that the dynamic interplay between words and contexts lies at the core of linguistic competence—human or machine.

Keywords: Large Language Models, Context-Dependent Semantics, Transformer Architecture, Hallucinations, Semantic Relationalism, Self-Attention, Contextual Embeddings, Philosophy of Language

1 Introduction

In recent years, the remarkable performance of Large Language Models (LLMs) has sparked debates in both technical and philosophical circles. Early critiques - from labeling LLMs as 'stochastic parrots' [1] to highlighting the absence of genuine 'intentionality' in text output [12,10] - argue that these systems merely generate surface-level statistical correlations rather than exhibit real semantic understanding.

On the other hand, several researchers (e.g. [8,9,3]) counter that LLMs may indeed form sophisticated internal representations that parallel human-like semantic processes. While the question of whether LLMs truly "understand" language remains open, there is increasing evidence that context-dependence is key to their capabilities. That is, it is the relationship between a token (or word) and its larger linguistic environment that enables LLMs to produce language that appears both coherent and contextually appropriate.

The philosophical debate over context-dependence has precedents in David Kaplan's [5] theory of indexicals and demonstratives, which posits that a term's "character" is stable across contexts, while the specific "content" can vary. David Braun's [2] critique challenges the notion of "stable character" by arguing for a more context-sensitive and relational understanding of meaning. This paper synthesizes these philosophical positions with empirical evidence from state-of-the-art LLMs (e.g., GPT-3, GPT-4) to show how the Transformer Architecture (TA) implicitly embodies Braun's perspective, thus prompting a reevaluation of long-standing semantic assumptions. Throughout, 'LLM' denotes a foundation decoder-only model (e.g., GPT-3) prior to instruction-tuning/RLHF unless explicitly stated otherwise.

2 Theoretical Foundations in Semantics

2.1 Kaplan's Theory of Character and Content

Kaplan's framework draws a line between "character" (the conventional rule determining reference) and "content" (the particular referent in each context). For indexical expressions like "I" or "this," Kaplan claims that the character remains constant across contexts, while the content (e.g., which individual "I" refers to) varies with each utterance.

2.2 Braun's Critique: Context as a Relational Entity

David Braun challenges Kaplan's use of extensional functions to represent meaning. He proposes that a term's interpretation should be far more relational, capturing subtleties that shift with context. While Kaplan's theory works neatly for clear-cut indexicals, Braun argues that everyday words exhibit context-sensitive nuances that cannot be reduced to a single stable function. This perspective draws on Wittgenstein's later maxim that "the meaning of a word is its use in the language game" (PI §43). The Transformer operationalizes this language-game principle by continually adapting word representations to their contextual use.

2.3 Bridging to LLMs

LLMs such as GPT-4 routinely exhibit context-sensitive behaviors, where tokens within the same environment shape each other's representations. This observation resonates more with Braun's emphasis on relational meaning than with Kaplan's stable character. In effect, LLMs rely heavily on context to disambiguate words and to generate locally coherent discourse—an outcome that underscores the significance of dynamic, context-dependent semantics.

3 Empirical Evidence in Existing Literature

3.1 Self-Similarity and Anisotropy

Analyses by Ethayarajh [4] and subsequent studies have shown that GPT-based embeddings (e.g., GPT-2, GPT-3) demonstrate increasingly low self-similarity and high anisotropy in deeper layers. Low self-similarity means a single word's vector representation changes significantly across different contexts. High anisotropy implies that vectors for different words in the same context can converge, often becoming almost indistinguishable when those words occupy similar contextual roles. Although the focus is on decoder-only checkpoints, this pattern has been reported for BERT-base – see Ethayarajh 2019 for the monotonic decline and both Ethayarajh 2019 and Li et al. [7] for the embedding anisotropy. [1]

This pattern contrasts the notion of a stable, context-independent character function. Instead, it reveals an implicit "Braunian" dynamic: words are interpreted through their relationships to surrounding tokens. For instance, "cat" and "dog" might converge in the vector space if they appear in highly analogous contexts ("the X sat on the mat"), even though they are distinct lexical items.

3.2 Practical Ramifications

Such context-based embeddings help LLMs achieve human-like performance on tasks such as paraphrasing, summarization, and even professional exams [6]. Their success challenges the assumption that purely stable, dictionary-like meanings are needed to govern correct usage. Indeed, the more context sensitivity the model learns, the better it performs in tasks requiring nuance.

3.3 Hallucination as Context Extrapolation Failure

Suppose a user asks: "Where did I grow up?" Without prior context, a language model may respond: "You grew up in Paris," despite no grounding for that claim. This error exemplifies a context extrapolation failure—the model relationally infers meaning for the indexical "I" using learned priors rather than present discourse. By contrast, if earlier context includes: "I lived in Paris as a child," the same model grounds its inference appropriately. Such examples illustrate how relational semantics can succeed or fail depending on contextual resolution, offering a philosophical lens through which hallucination may be reinterpreted—not as random error, but as inference without referential anchoring.

[1] Whether the same relational semantics appears in encoder or encoder–decoder Transformer Architectures remains an open question; see Section 3 for preliminary evidence from BERT and BLOOM that the anisotropy trend persists independently of objective or model size.

4 Mechanistic Underpinnings of the Transformer Architecture

4.1 Self-Attention as Context Integration

The Transformer Architecture (TA) introduced by Vaswani et al. [13] dispenses with recurrent or convolutional layers. It employs a “masked self-attention” mechanism that computes the relational influence of every other token in the sequence on the token being generated. Queries, Keys, and Values (Q, K, V) are learned projections that capture how tokens should attend to one another. The essential insight is that meaning is aggregated from context: each token’s final representation is a weighted sum of all the “Value” vectors, modulated by learned attention weights.

4.2 Multi-Head Attention and Feed Forward Layers

By splitting the attention mechanism into multiple heads, Transformers can model different aspects of context in parallel. Each head might focus on syntax, coreference, or other semantic cues. After attention, feed-forward layers with ReLU nonlinearities expand and contract dimensionality, further refining each token’s representation. Residual connections and layer normalization help preserve information and stabilize training, ensuring that context influences accumulate effectively across layers.

4.3 Alignment with Context-Dependent Semantics

Taken together, these operations exhibit an emergent property akin to Braun’s relational semantics. Rather than storing a single stable “character” for each word, the model continually recalculates word representations based on the local context. This parallels Braun’s argument that meaning shifts according to relational cues—only here, the “cues” are encoded in the Transformer’s attention weights and feed-forward transformations.

5 Discussion

5.1 Reassessing Philosophical Foundations

The success of LLMs, powered by TAs, compels a reexamination of classical semantic theories. Kaplan’s [5] model of stable character functions may hold in tightly constrained indexical scenarios, but the broad success of GPT-based systems seems to confirm Braun’s [2] insistence that context exerts a profound influence on meaning—even for non-indexical words. As Wolfram [14] suggests, these developments in AI might be the most significant impetus in two millennia for exploring the essence of human language and thought processes.

5.2 Future Directions in Linguistics and AI

Beyond philosophical debates, a deeper understanding of context-dependent semantics could inform strategies for addressing issues such as “hallucination” in LLMs and alignment with human values. By pinpointing how context shapes meaning at each step of inference, researchers may develop more robust and interpretable systems. Additionally, these insights could inform pruning or distillation strategies for LLMs [11], leveraging linguistic theory to optimize model architectures and reduce computational overhead.

6 Conclusion

This paper has argued that the decoder-only Transformer Architecture in LLMs offers a powerful real-world demonstration of how language meaning can emerge from context-dependent relations, consistent with David Braun’s philosophical stance. Empirical data on GPT embeddings—particularly the low self-similarity of the same word in different contexts—reinforce the view that language is best understood relationally. In challenging Kaplan’s classical framework of stable character functions, LLMs serve as computational evidence that context can single-handedly support complex linguistic capabilities.

By bridging philosophical inquiries and technical implementations, the findings underscore the importance of continued interdisciplinary dialogue. If meaning in language is indeed relational at its core, then LLMs are not mere “stochastic parrots,” but sophisticated models that illuminate how context shapes understanding. As these architectures broaden their impact across disciplines—from medicine to mathematics—future research will likely uncover deeper connections between the structures of computational models and the long-studied mysteries of human language and cognition.

References

1. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots. FAccT ’21 Proceedings (2021)
2. Braun, D.: What is character? Journal of Philosophical Logic **24**(3), 227–240 (1995)
3. Downes, S.M., Forber, P., Grzankowski, A.: Llms are not just next token predictors. arXiv preprint arXiv:2408.04666 (2024)
4. Ethayarajh, K.: How contextual are contextualized word representations? arXiv preprint arXiv:1909.00512 (2019)
5. Kaplan, D.: Themes from kaplan. Oxford University Press (1989)
6. Katz, D.M., Bommarito, M.J., Gao, S., Arredondo, P.: Gpt-4 passes the bar exam. Philosophical Transactions of the Royal Society A **382**(20230254) (2024). `https://doi.org/10.1098/rsta.2023.0254`
7. Li, B., et al.: On the sentence embeddings from pre-trained language models. arXiv preprint arXiv:2011.05864 (2020)
8. Lyre, H.: Understanding ai: Semantic grounding in large language models. arXiv preprint arXiv:2402.10992 (2024)

9. Millière, R., Buckner, C.: A philosophical introduction to language models–part i: Continuity with classic debates. arXiv preprint arXiv:2401.03910 (2024)
10. Shanahan, M.: Talking about large language models. Communications of the ACM **67**(2), 68–79 (2024)
11. Sreenivas, S., Nguyen, V., Muralidharan, S., Chochowski, M., Joshi, R.: How to prune and distill llama-3.1 8b to an nvidia llama-3.1-minitron 4b model (2024), `https://developer.nvidia.com/blog/how-to-prune-and-distill-llama-3-1-8b-to-an-nvidia-llama-3-1-minitron-4b-model/`, nVIDIA Developer Blog
12. Titus, L.M.: Does chatgpt have semantic understanding? Cognitive Systems Research **83** (2024)
13. Vaswani, A., et al.: Attention is all you need. NeurIPS 2017 (2017)
14. Wolfram, S.: What is chatgpt doing ... and why does it work? Stephen Wolfram Writings (2023)