TURKISH ACADEMIC RESEARCH REVIEW

Why Large Language Models Can Say but Cannot Imply

Büyük Dil Modelleri Neden Söyler ama Kastedemez

Mahmut Sami Özdil[1*]

[1] Dr. Öğr. Üyesi, Necmettin Erbakan Üniversitesi, Ahmet Kelesoğlu İlahiyat Fakültesi, Felsefe ve Din Bilimleri Anabilim Dalı, Konya, Türkiye.
Assist. Prof. Dr., Necmettin Erbakan University, Ahmet Keleşoğlu Faculty of Theology, Department of Philosophy and Religious Studies, Konya, Türkiye.
https://ror.org/013s3zh21, https://orcid.org/0000-0003-3167-9696, mahmutsami.ozdil@erbakan.edu.tr

* Sorumlu yazar | Corresponding author

Araştırma Makalesi
Süreç

Geliş Tarihi: 14.01.2026
Kabul Tarihi: 24.03.2026
Yayım Tarihi: 25.03.2026

Benzerlik
Bu makale, en az iki hakem tarafından incelenmiş ve intihal yazılımı ile taranmıştır.

Değerlendirme
Ön İnceleme: İç hakem (editörler).
İçerik İnceleme: İki dış hakem/Çift taraflı körleme.

Etik Beyan
Bu çalışma, etik kurul izni gerektirmeyen nitelikte olup kullanılan veriler literatür taraması/yayınlanmış kaynaklar üzerinden elde edilmiştir. Çalışmanın hazırlanma sürecinde bilimsel ve etik ilkelere uyulduğu ve yararlanılan tüm çalışmaların kaynakçada belirtildiği beyan olunur. Mahmut Sami Özdil

Yapay Zekâ Kullanımı
Bu çalışmanın hazırlanma sürecinde yapay zekâ tabanlı herhangi bir araç veya uygulama kullanılmamıştır. Çalışmanın tüm içeriği, yazar(lar) tarafından bilimsel araştırma yöntemleri ve akademik etik ilkelere uygun şekilde üretilmiştir. Mahmut Sami Özdil

Etik Bildirim
turkisharr@gmail.com

Çıkar Çatışması
Çıkar çatışması beyan edilmemiştir.

Finansman
Bu araştırmayı desteklemek için dış fon kullanılmamıştır.

Yayıncı
Published by Mehmet ŞAHİN Since 2016-Akdeniz University, Faculty of Theology, Antalya, 07058 Türkiye

Atıf
Özdil, M. S. (2026). Büyük dil modelleri neden söyler ama kastedemez. *Turkish Academic Research Review*, *11*(1), 223-238.
https://doi.org/10.30622/tarr.1863625

Öz
Bu çalışma, Büyük Dil Modelleri'nin (BDM) artan üretken akıcılıklarına ve çok-modlu yeteneklerine rağmen, yapısal olarak hakiki bir dilsel ima (implication) eyleminden aciz olduğunu savunmaktadır. Metodolojik olarak çalışma, analitik dil felsefesi ve yapay zeka etiği literatürünü sentezleyen teorik bir analiz çerçevesine dayanmaktadır. Araştırma yöntemi; John Austin ve John Searle'ün Söz Edimi Kuramı, Paul Grice'ın pragmatik ilkeleri, David Kaplan'ın duyusal işaret (indexicality) analizi ve Ludwig Wittgenstein'ın "yaşam biçimleri" kavramlarının, mevcut BDM mimarilerine uygulanması yoluyla yürütülen eleştirel bir soruşturmadan oluşmaktadır. Argüman beş temel aşamada yapılandırılmıştır. İlk olarak, modellerin konuşma normlarını kasıtlı olarak ihlal etme yetisine sahip olmadıkları, dolayısıyla "söylenenden" fazlasını kastetmek için gereken "niyetli faillik" (intentional agency) koşulunu sağlamadıkları gösterilmektedir. İkinci olarak, yapay zeka çıktılarında kullanılan "Ben", "Burada" ve "Şimdi" gibi duyusal işaret lafızlarının, cisimleşmiş bir perspektife dayanmadığı için "merciî bir boşluğa" (referential hollowing) düştüğü analiz edilmektedir. Üçüncü aşamada, ses ve görüntü işleyebilen çok-modlu sistemlerin kinaye veya empati gibi durumları simüle etseler bile, bu eylemleri doğrulayacak toplumsal dayanaktan ve ontolojik statüden yoksun oldukları vurgulanmaktadır. Dördüncü aşamada, Wittgenstein'ın felsefesinden hareketle, imanın sadece dilsel bir oyun değil, risklerin ve sorumlulukların paylaşıldığı bir "yaşam biçimine" katılımı gerektirdiği; dolayısıyla BDM'lerin bu alana erişemeyeceği savunulmaktadır. Son olarak, ses ve görüntü işleyebilen çok-modlu sistemlerin kinaye veya empati gibi durumları simüle etseler bile, bu eylemleri doğrulayacak toplumsal dayanaktan, ontolojik statüden ve niyetli durumlardan yoksun oldukları sonucuna varılmaktadır. Çalışma, bu sistemleri pragmatik failler olarak ele almanın, insan iletişiminin temelini oluşturan normatif kavramların içinin boşaltılması riskini taşıyan bir "işlevselci ayartma" olduğu sonucuna varmaktadır.

Anahtar Kelimeler
mantık, büyük dil modelleri, dilsel ima, muktezâyı hâl, kasıt.

Öne Çıkanlar

- Büyük Dil Modelleri, üretken akıcılıklarına rağmen "niyetli faillik" eksikliği nedeniyle hakiki imadan yoksundur.
- Konuşma normlarının stratejik ihlali, modellerin sahip olmadığı bir iletişimsel kasıt gerektirir.
- Yapay zeka çıktılarındaki "Ben" ve "Burada" ifadeleri, cisimleşmiş bir özneye dayanmadığı için "merciî bir boşluğa" düşer.
- Çok-modlu kinaye simülasyonları, toplumsal statü ve ortak geçmişten yoksun olduğu için pragmatik açıdan geçersizdir.
- Hakiki ima, algoritmaların erişemeyeceği "risklerin paylaşıldığı bir yaşam biçimine" katılımı zorunlu kılar.

Research Article

History
Received: 14.01.2026
Accepted: 24.03.2026
Date Published: 25.03.2026

Plagiarism Checks
This article has been reviewed by at least two referees and scanned via a plagiarism software.

Peer-Review
Single anonymized-One internal (Editorial Board). Double anonymized-Two external.

Ethical Statement
This study does not require ethical committee approval, and the data used were obtained from literature reviews/published sources. It is hereby declared that scientific and ethical principles were adhered to during the preparation of this study and that all studies used are cited in the references. Mahmut Sami Özdil

Use of Artificial Intelligence
No artificial intelligence-based tools or applications were used in the preparation of this study. All content of the study was produced by the author(s) in accordance with scientific research methods and academic ethical principles. Mahmut Sami Özdil

Complaints
turkisharr@gmail.com

Conflicts of Interest
The author(s) has no conflict of interest to declare.

Grant Support
The author(s) acknowledge that they received no external funding in support of this research.

Published
Published by Mehmet ŞAHİN Since 2016- Akdeniz University, Faculty of Theology, Antalya, 07058 Türkiye

Cite as
Özdil, M. S. (2026). Why large language models can say but cannot imply. *Turkish Academic Research Review*, *11*(1), 223-238.
https://doi.org/10.30622/tarr.1863625

Abstract
This study argues that, despite their increasing fluency and multimodal capabilities, large language models (LLMs) are structurally incapable of genuine linguistic implication. In terms of methodology, the paper employs a theoretical analysis that synthesizes analytic philosophy of language with recent critiques in AI ethics, applying Speech Act Theory, Gricean pragmatic maxims, Kaplan's analysis of indexicality, and Wittgenstein's concept of "forms of life" to the architecture of contemporary LLMs to test the conditions of possibility for indirect communication. The argument proceeds through five distinct analytical stages. First, drawing on speech act theory, it demonstrates that models lack the "intentional agency" required to strategically violate conversational norms, which is a prerequisite for Gricean implicature. Second, it identifies a problem of "referential hollowing," arguing that indexical terms like "I" and "here" in AI outputs fail to anchor to a determinate, embodied perspective, leaving them pragmatically inert. Third, the study investigates the critical function of paralinguistic cues and prosody in determining illocutionary force, noting that symbolic abstractions miss the expressive control essential for implication. Fourth, invoking Wittgenstein's analysis, the paper argues that genuine implication presupposes participation in a shared "form of life" involving risks and commitments, a domain inaccessible to algorithmic systems. Finally, the study addresses multimodal systems and role-playing objections, concluding that simulating the acoustic artifacts of sarcasm or adopting a persona creates only a sophisticated simulation devoid of the social standing and intentional states necessary to validate such acts. The study concludes that treating these systems as pragmatic agents is a "functionalist temptation" that risks hollowing out the normative concepts essential to human discourse.

Keywords
logic, large language models, linguistic ımplication, pragmatics, ıntentionality.

Highlights
- Large Language Models lack genuine linguistic implication, despite their fluency, due to the absence of intentional agency.
- Strategic violation of conversational norms requires a specific communicative intent that LLMs structurally lack.
- Indexical terms like "I" and "here" in AI outputs suffer from "referential hollowing" as they are not anchored to an embodied subject.
- Multimodal simulations of sarcasm remain pragmatically inert due to the absence of social standing and shared history.
- Genuine implication mandates participation in a "form of life" involving shared risks, a domain inaccessible to algorithmic systems.

Introduction

Human communication routinely conveys more than what is explicitly said. Speakers imply, hint, suggest, and allude, relying on shared context, embodied presence, and social expectations to transmit meaning indirectly. This capacity for implication is not a marginal feature of language use but one of its most efficient mechanisms. The central claim of this paper is that, despite their impressive linguistic fluency, large language models are structurally incapable of genuine linguistic implication (Bender et al., 2021, p. 616).

The limitation at issue does not concern surface-level performance or engineering constraints such as context window size. Rather, it arises from deeper structural conditions. Implication depends on intentional agency, indexical grounding, embodied reference, and participation in shared social practices, conditions that large language models do not satisfy (Bender et al., 2021, p. 616; Harnad, 2025, p. 4).[1] While functionalist approaches often measure success by syntactic correctness or statistical likelihood, philosophical critiques have long emphasized that meaning emerges from historically situated, embodied agency. From this perspective, the fluency of generative AI creates a powerful illusion of understanding, masking the absence of the pragmatic micro-structures that make implication possible (Bender et al., 2021, p. 616). Functionally, these models operate as what Rosen and Dale (2024) classify as "proxies for illocution-deficient language generators", systems that produce plausible text specifically because they lack the "internal sense of need" that drives human linguistic acquisition and use (Rosen & Dale, 2024, p. 2871).

This paper focuses on implication as a test case for these limits. Drawing on speech act theory, Gricean pragmatics, and theories of indexical reference, it argues that language models reproduce the *products* of implicature without performing the *acts* of implying. Because they lack intentions, they cannot strategically violate conversational norms; because they lack embodiment, they cannot anchor indexicals like "I" or "here" to a determinate perspective (Bender et al., 2021, p. 617; Grindrod, 2024, p. 70; Harnad, 2025, p. 4). Furthermore, as text-based systems, they are insensitive to the paralinguistic cues (especially prosody) that shape illocutionary force. Finally, drawing on Wittgenstein, the paper argues that implication presupposes participation in "forms of life" where speech carries practical stakes, a condition artificial systems cannot inhabit.

The study advances its argument through five distinct analytical stages. The initial section utilizes speech act theory to scrutinize the intrinsic link between implication and intentional agency. Subsequently, the second section investigates the structural problems of reference and indexical grounding in disembodied systems. The third part addresses the critical function of paralinguistic cues in determining pragmatic force. Building on these foundations, the fourth section contextualizes these limitations within the Wittgensteinian framework of language-games. The fifth section responds to anticipated objections regarding multimodality, role-playing, and behavioral indistinguishability. The study concludes by unifying these arguments to evaluate their wider significance for the philosophy of language and the assessment of AI capabilities.

1. Implication, Intentionality, and the Illusion of Understanding

[1] See also (Gubelmann, 2024, p. 1) where he argues that "agents with intentions need to be autonomous organisms while LLMs are heteronomous mechanisms."

To understand why large language models cannot imply, it is first necessary to clarify what implication involves at the most basic level. Implication is not encoded in sentence structure, nor is it recoverable from semantic content alone (Bender et al., 2021, p. 615). Rather, it is a pragmatic phenomenon arising from the interaction between what is said, how it is said, and what interlocutors mutually take for granted within a shared context. This interaction presupposes agents who act with communicative intentions, who can be held responsible for their utterances, and who situate those utterances within a common practical world (Grice, 1975, p. 49).[2]

Large language models do not meet these conditions. Their limitation does not lie in an inability to generate complex syntax or coherent discourse. On the contrary, such systems routinely produce linguistically sophisticated outputs (Grindrod, 2024, p. 70; Gubelmann, 2024, p. 32). The difficulty is that implication is not a property of linguistic form. It is an achievement of speakers (Searle, 1969, pp. 16, 18). Where there is no intention to mean more than what is said, there is no implication, only the appearance of it. This distinction is vividly captured by Gubelmann's (2024) analogy of a tortoise coincidentally tracing a sentence in the sand: despite the semantic coherence of the trace, "it would be inaccurate to say that the tortoise told the person what to do" because the act lacks the teleological structure of intention (Gubelmann, 2024, p. 2).

This becomes clear when viewed through classical speech act theory. Following Austin and Searle, linguistic communication involves not only locutionary acts, which consist in producing meaningful utterances, but also illocutionary acts, which involve performing an action in saying something (Austin, 1962, p. 99). Requests, warnings, promises, and implications are illocutionary in this sense. They depend on the speaker's intention to bring about a certain uptake in the listener. Large language models can produce locutionary forms with high reliability, but they lack the intentional agency required for genuine illocutionary force (Attah, 2025, p. 4; Rosen & Dale, 2024, p. 2874). Even defenders of LLM capabilities, such as Grindrod (2024), concede that "LLMs that lack such mental states could nevertheless meaningfully use words in a non-derivative sense" (Grindrod, 2024, p. 2). However, divorcing meaning from mental states reduces communication to mere output generation, stripping it of the pragmatic force essential for implication.

Consider a familiar example. When a human speaker utters the sentence "It's cold here," the utterance may function not merely as a description but as an implicit request to close a window. The implied meaning arises because the speaker intends the listener to recognize that intention. Nothing in the sentence itself guarantees this function. The implication is generated by the speaker's aim and by the assumption that the listener will infer that aim under conditions of cooperative interaction. A language model that produces the same sentence does not thereby perform the same act. It selects the utterance because it is statistically appropriate in context, not because it intends to alter the state of the world. Consequently, the model fails the logical test for speakerhood: as Gubelmann (2024) demonstrates, since being a speaker requires the capacity to illocute, and illocution requires intention, an entity lacking intention cannot be a speaker (Gubelmann, 2024, p. 12).

This distinction becomes even more pronounced when considering contemporary multimodal interactions that transcend simple text. For instance, a user might upload a photograph of an obviously cluttered and disorganized workspace to a state-of-the-art multimodal model and ask, "What do you think of my afternoon's

[2] Cf. Austin, 1962, on the distinction between what is said and what is done in speaking.

work?" A sophisticated model, detecting the mismatch between the visual evidence of a mess and the linguistic prompt's positive framing, may output a response such as, "Your afternoon was clearly very productive", perhaps even accompanied by a sarcastic vocal tone in a speech-to-speech interface. While this appears to be a sarcastic implicature, implying that the user was actually unproductive, the model is not strategically violating the maxim of quality to convey a personal attitude. Instead, it is executing a high-probability pattern of sarcastic speech synthesis, mapping visual tokens of disorder to textual and acoustic tokens of ironic praise.

Here lies the boundary between reproducing the product of implication and performing the act of implying. Language models can often supply explicit paraphrases of implicit content when prompted. They may explain that "It's cold here" can function as a request, or that praising a mess constitutes irony. But this explanatory capacity should not be confused with pragmatic agency. Explaining an implicature is not the same as generating one with communicative intent (Rosen & Dale, 2024, p. 2874).

The literature on large language models captures this limitation under the label of the "stochastic parrot." According to this diagnosis, language models operate by detecting and reproducing probabilistic regularities in linguistic data rather than by engaging in goal-directed communication. The fluency of their outputs can therefore give rise to an illusion of understanding. Observers attribute intentions, attitudes, or communicative goals to systems that, in fact, possess none (Attah, 2025, p. 4; Bender et al., 2021, pp. 616–617). This phenomenon is described even by critics of the 'no-intention' view like Attah (2025), who notes that the endless loop of self-corrections in LLMs suggests "prima facie that its capacity for conversational repair is "hollow": it merely simulates this linguistic capacity rather than enacts any genuine conversational repair" (Attah, 2025, p. 4). Crucially, this lack of genuine intent is not just a philosophical abstraction but a measurable deficit. Rosen and Dale (2024) demonstrate that human illocutionary intent adds approximately "2.71 bits of entropy" to a response compared to an LLM, revealing that machine fluency is statistically "cleaner" precisely because it is stripped of the pragmatic force that characterizes human agency (Rosen & Dale, 2024, p. 2874). This illusion is reinforced by the model's ability to produce meta-linguistic commentary about meaning and implication, which further masks the absence of agency at the level where implication actually occurs (Gubelmann, 2024, p. 32).

Grice's framework provides a precise lens for this issue. Conversational implicature emerges specifically when an interlocutor deliberately deviates from a communicative norm to transmit meaning via indirection. This communicative act is predicated on the mutual expectation that the listener will discern the deviation and infer the underlying intent (Grice, 1975, p. 49). This process presupposes not only shared norms of cooperation but also a speaker who can strategically choose to violate those norms. Large language models cannot do this. They do not form communicative goals, assess the expectations of an interlocutor, or decide to withhold information in order to suggest it indirectly. Any apparent violation of a maxim in model output is a by-product of pattern completion, not a communicative strategy (Attah, 2025, pp. 5–6; Gubelmann, 2024, p. 32).[3]

Because of this, the pragmatic competence displayed by language models remains derivative. It depends on human users supplying the interpretive background that turns statistically generated text into meaningful communication. The model itself engages in no act of implication; instead, users construct meaning by treating

[3] Attah outlines this position (the 'Communicative Intention Argument') as a basis for his subsequent critique.

the generated text as though it emanated from a goal-directed subject. The resulting impression of pragmatic mastery therefore reflects a projection from the human side rather than a capacity realized by the system itself (Bender et al., 2021, p. 616; Rosen & Dale, 2024, p. 2871).

The upshot is that implication cannot be reduced to linguistic fluency or contextual sensitivity understood in purely textual terms. It presupposes intentional action within a shared world. Without that foundation, large language models may successfully mimic the forms of indirect speech, but they remain incapable of implying anything in the full pragmatic sense.

2. Indexical Reference and the Problem of Disembodiment

The role of intentionality in implication is closely connected to the problem of reference, and especially to the use of indexicals. Indexical expressions such as “I,” “here,” and “now” do not merely refer; they locate an utterance within a concrete perspective (Benveniste, 1971, p. 219). Their successful use presupposes a speaker who occupies a determinate position in space, time, and social interaction. As Benveniste (1971) observes, “it is a fact both original and fundamental that these ‘pronominal’ forms do not refer to ‘reality’ or to ‘objective’ positions in space or time but to the utterance, unique each time, that contains them” (Benveniste, 1971, p. 219). For an LLM, which has no unique “instance of discourse,” these forms refer to nothing; they are syntactic placeholders stripped of their pragmatic function. Without such a position, indexical reference collapses into abstraction (Austin, 1962, p. 8; Grice, 1975, pp. 44, 50).

Philosophy of language has long established this dependence. Kaplan famously argued that indexicals are directly referential and that their content is determined by the immediate circumstances of the speech act. Under this view, the reference of 'I' is contingent upon the identity of the speaker, just as 'here' signifies the location of the verbal act, and 'now' designates its temporal occurrence (Benveniste, 1971, p. 219). These dependencies are not linguistic conventions alone but reflections of the embodied situation of a speaker (Austin, 1962, p. 6). Indexicals function precisely because they tie language to the practical standpoint of an agent.

Large language models lack such a standpoint. They do not occupy space or time, nor do they participate in situations that could ground indexical reference. When a model produces an utterance containing “I” or “here,” it does not thereby refer to itself or to a location. According to Benveniste, “I” does not denote a permanent entity like “tree” or “house”; rather, “I refers to the act of individual discourse in which it is pronounced, and by this it designates the speaker” (Benveniste, 1971, p. 226). Without this specific act of utterance, what he calls the “instance of discourse”, the pronoun remains an empty form.

John Perry’s (1979) “messy shopper” thought experiment perfectly illustrates this disconnect. Perry describes a shopper who knows that “the shopper with the torn sack is making a mess” but fails to realize “I am the shopper I was trying to catch” (Perry, 1979, p. 3). Like the oblivious shopper, an LLM may possess the descriptive knowledge of a situation but lacks the essential indexical belief required to anchor that knowledge to a perspective. The apparent reference is borrowed from the conversational frame supplied by the user. As a result, indexical expressions in model output function only parasitically: their referential force is inherited from human interpretation rather than generated by the system (Attah, 2025, p. 18; Grice, 1975, p. 50; Grindrod, 2024, p. 17; Searle, 1969, pp. 89, 99).

Such a reliance on external interpretation substantiates the 'symbol grounding problem' identified by Harnad (1990). His argument establishes that meaning within a formal architecture is not inherent to the mechanism itself but is, instead, derived entirely from the cognitive resources of the interpreter, what he terms being “parasitic on the meanings in our heads” (Harnad, 1990, p. 335). Harnad illustrates this vacuity through the metaphor of a dictionary “merry-go-round”, describing a closed loop where symbols point only to other symbols without ever anchoring to physical reality (Harnad, 1990, p. 339). Although Harnad acknowledges that a system could achieve intrinsic meaning if it possessed “behavioral capacity” to interact with the objects it names (Harnad, 1990, p. 345), LLMs remain pure symbol systems lacking this sensorimotor grounding. Just as a dictionary cannot refer to a user without a user to read it, an LLM cannot refer to itself without a grounded subject to inhabit it.

Consequently, they cannot form what Perry calls “locating beliefs”, beliefs about “where one is, when it is, and who one is”. Perry argues that “such beliefs seem essentially indexical” and cannot be replaced by objective descriptions (Perry, 1979, pp. 4–5). Without a “here” or a “now” to anchor them, the model’s indexicals are semantically adrift.

Crucially, the optimistic conclusions of semantic externalism fail to account for this structural deficit. Mandelkern and Linzen (2024) recently argued that large language models may indeed achieve reference because their training data consists of “strings of text with natural histories”, causal chains that link words to the world regardless of the speaker’s intent (Mandelkern & Linzen, 2024, p. 1191). While this externalist account might explain how a model refers to static entities like “Napoleon” or “gold,” it fails to account for indexicals. As Kaplan showed, the reference of “I” or “now” is not fixed by a causal history but by the immediate context of utterance. An LLM may inherit the causal history of the word “I,” but without a body to occupy the center of the utterance, it cannot anchor that history to a present reality. This theoretical deficit is supported by empirical evidence. Oğuz, Bakman, and Yaldiz (2025) demonstrate that while state-of-the-art models may mimic static patterns, they “struggle” significantly with shifting indexicals like “you”, “here”, and “tomorrow” (Oğuz et al., 2025, p. 23413). Their analysis reveals that models often fail to update the temporal reference of “tomorrow” when the context shifts (e.g., in reported speech), proving that the system processes the word as a lexical token rather than a time-anchored pointer. The machine sees the word “tomorrow”, but it does not inhabit the “now” required to define it.

We may term this phenomenon “referential hollowing”. As Startari (2025) vividly describes, in predictive systems “the surface of language retains the forms of reference but the chain of reference itself disappears,” leaving behind only “the shell of indexicality, emptied of its anchoring” (Startari, 2025, pp. 4, 7). While Startari (2025) draws on Benveniste’s account of subjectivity to discuss "Indexical Collapse" as a mechanism that paradoxically generates *institutional authority*, this paper employs Kaplan’s semantic framework to argue that the lack of indexical grounding specifically precludes the *pragmatic act of implication*. Implication is typically generated by the contrast between what is linguistically encoded in the text and the particular indexical standpoint of the speaker. If the coordinates ("I," "here," "now") are hollow, the tension required to generate indirect meaning cannot exist. Because language models are trained on vast corpora of indexical uses, they can reproduce the grammatical patterns associated with indexical expressions (Attah, 2025, p. 4). Yet this competence remains detached from the referential anchoring necessary for pragmatic inference. The model does not know who “I” refers to, because there is no entity in virtue of which the reference could be fixed (Attah, 2025, p. 18; Searle, 1969, pp. 91–92).

The consequences for implication are substantial. For an implicature to be calculated, the listener must understand the speaker's position relative to the utterance. A speaker may imply urgency, proximity, or personal involvement precisely by exploiting the perspectival character of indexicals. When someone says "I'll take care of it now," the implied commitment, and the potential relief it signals to the listener, depends entirely on recognizing *who* the speaker is and *when* "now" is. A language model cannot generate such implicatures on its own because it lacks the contextual parameters that make the "unsaid" content determinate. Without a situated "I," the bridge between what is said and what is implied dissolves.

Expanding the model's context window or providing additional metadata cannot overcome this limitation. No amount of textual context can substitute for a genuine perspective. Indexical reference is not a matter of having more information but of occupying a position from which information is accessed (Austin, 1962, pp. 10–11). Language models, being disembodied systems, do not occupy such positions. Their outputs therefore remain ungrounded with respect to the very features of context that indexicals are designed to exploit to convey indirect meaning (Attah, 2025, p. 8).

The problem is not merely semantic but deeply pragmatic. As Özdil (2025) demonstrates, even within logical frameworks, reference relies on non-linguistic elements to bridge the gap between symbol and reality (Özdil, 2025, p. 36). Indexicals play a central role in coordinating action, establishing common ground, and signaling commitments. When these expressions are detached from an embodied speaker, they lose their capacity to support the heavy lifting of implication. What remains is a formal simulacrum of reference, grammatically correct and contextually plausible, but pragmatically inert and incapable of bearing the weight of genuine implicature.

Ultimately, this analysis reinforces the conclusion of the previous section. Just as implication requires intentional agency, it also requires a situated perspective from which reference can be fixed. Large language models lack both. Their apparent mastery of indexical language thus contributes to the broader illusion of understanding while concealing the absence of the conditions under which implication genuinely operates.

3. Prosody, Paralinguistic Cues, and Illocutionary Force

Implication depends not only on what is said and on who says it, but also on how it is said. In ordinary human communication, paralinguistic cues, especially prosody, play a decisive role in determining illocutionary force. Intonation, stress, rhythm, and pause often distinguish between a statement and a request, a sincere remark and a warning, or a literal utterance and an instance of irony. These features are not optional embellishments of meaning but integral components of how indirect speech functions. This parallels Kaplan's (1989) analysis of indexicals, where he insists that "a demonstrative without an associated demonstration is incomplete" (Kaplan, 1989b, p. 490). Just as the linguistic rule for 'that' is insufficient to determine a referent without a physical act of indication, the text of an utterance is often insufficient to determine illocutionary force without the 'associated demonstration' of prosody.

Speech act theory already anticipates this point. Although Austin and Searle primarily analyze illocutionary acts at the level of conventional force, their accounts presuppose a background of expressive cues through which that force is recognized and negotiated in practice. The same string of words can perform radically different acts depending on prosodic modulation. An utterance such as "You're coming" may function as a neutral

observation, a question, or a command, depending on how it is spoken. The illocutionary force is not encoded in the sentence alone. Even when revisiting his theory in *Afterthoughts*, Kaplan compares the act of demonstration explicitly to a paralinguistic aid, noting that "the externalization is an aid to communication, like speaking more slowly and loudly" (Kaplan, 1989a, p. 582). While he attempts to minimize its semantic weight, this analogy inadvertently highlights what LLMs miss: the "aids" of prosody and gesture that are essential for guiding human communicative uptake.

This dependence on paralinguistic cues is especially salient in cases of implication. Many implicatures are conveyed through tonal understatement, emphasis, or deliberate flatness. Sarcasm, for example, often relies on a mismatch between semantic content and prosodic delivery. As Caucci and Kreuz (2012) demonstrate, speakers rely on a complex array of "paralinguistic cues," including "movement of the head, eyes, and mouth," to signal sarcastic intent (Caucci & Kreuz, 2012, p. 1). Crucially, they find that these cues are not universal but depend on social proximity: they are "more commonly employed by friends than by strangers" (Caucci & Kreuz, 2012, p. 1). This reliance on "common ground" exposes a fundamental gap for LLMs. Since the model shares no history, no social relation, and (in text-only versions) no physical channel with the user, it lacks the requisite signaling mechanisms to ground the "risky" act of sarcasm. Similarly, hints and indirect requests frequently depend on subtle shifts in stress or timing that signal non-literal intent. These cues guide the listener in recognizing that the speaker means more, or other, than what is explicitly said.

Large language models, however, operate almost exclusively at the level of text. Even when deployed in conversational interfaces, they process and generate symbolic representations that abstract away from prosodic and embodied dimensions of speech. As a result, they cannot generate paralinguistic cues in the sense relevant to illocutionary force. Consequently, they are trapped in a void between the two necessary conditions of reference identified by Kaplan: they lack the body to perform an external "demonstration"(Kaplan, 1989b, p. 490) and the mind to form an internal "directing intention" (Kaplan, 1989a, p. 582). Without either the bodily act or the intentional state, their output remains semantically hollow. At best, they can describe such cues or simulate their effects through explicit markers (for example, by annotating an utterance as sarcastic). But describing a cue is not the same as producing it as part of an expressive act.

Furthermore, the inability to signal pragmatic distinctions undermines the possibility of genuine implication. Without access to prosodic modulation, a system cannot signal many of the pragmatic distinctions that listeners rely on when interpreting indirect meaning. It cannot intentionally soften a request, heighten urgency, or feign neutrality in order to suggest an attitude. Any apparent sensitivity to these distinctions arises from human interpretation rather than from the system's own communicative control.

One might object that advances in multimodal systems could address this problem by incorporating voice, gesture, or affective signals. Yet even here the difficulty persists. Producing prosodic variation is not sufficient for implying; the variation must be generated as an expression of an intention situated within a shared interaction. A system that outputs vocal contours without possessing communicative goals merely adds another layer of simulation. The issue is not the absence of acoustic data but the absence of agency that could deploy such data meaningfully.

The dependence of implication on paralinguistic cues also clarifies why text based evaluations of language models systematically overestimate their pragmatic competence. When users read model outputs, they

supply tone, emphasis, and attitude imaginatively. The written text functions as a prompt for human prosodic reconstruction. This cooperative supplementation creates the appearance of indirect meaning where none has been performed. The illusion is therefore sustained by the reader's interpretive labor rather than by the model's expressive capacities.

The cumulative effect of these observations is to show that illocutionary force cannot be reduced to linguistic form plus contextual inference construed abstractly. It depends on expressive control exercised by an agent within a concrete interaction. Prosody is one of the primary means through which such control is realized. Where this dimension is absent, implication becomes at best a projected interpretation rather than an achieved communicative act.

This conclusion prepares the ground for the final step of the argument. Even if one were to grant intentionality, reference, and paralinguistic expression independently, implication would still presuppose participation in shared practices governed by norms and expectations. The next section therefore turns to Wittgenstein's notion of language-games and forms of life to argue that implication is inseparable from the practical contexts in which speakers live and act.

4. Forms of Life, Language-Games, and the Boundaries of Implication

Genuine linguistic implication presupposes not only individual agency but also participation in a shared world. As Wittgenstein explicitly states, "to imagine a language means to imagine a form of life" (*Lebensform*) (Wittgenstein, 1986, para. 19). The preceding sections have argued that implication requires intentional agency, indexical grounding, and expressive control. These conditions, however, do not operate in isolation. They are embedded within broader patterns of human activity that give linguistic acts their point. To make this dependence explicit, it is necessary to turn to Wittgenstein's account of language-games and forms of life.

According to Wittgenstein, the meaning of an utterance is inseparable from the role it plays within a practice. Speaking is not merely the production of symbols governed by abstract rules, but a form of action situated within shared ways of acting, responding, and evaluating. Language-games are not theoretical constructs imposed from above; they are descriptions of how words are woven into human life. To understand what an utterance means is to grasp what counts as doing something with it.

The distinction between simulating a move and actually making one is clarified by Wittgenstein's famous analogy in *Philosophical Investigations*: "Why can't my right hand give my left hand money? ... My right hand can write a deed of gift and my left hand a receipt. But the further practical consequences would not be those of a gift."(Wittgenstein, 1986, para. 268). The physical mechanics of the transfer are identical to a real gift, yet the act is void because it occurs within a closed loop, isolated from the economy of needs and debts that gives a gift its value.

Implication fits naturally into this picture. To imply is not simply to leave something unsaid, but to rely on shared expectations about how participants in a practice will respond. When a speaker hints, alludes, or speaks indirectly, she presupposes a background of norms that determine what counts as appropriate uptake. These norms are sustained not by linguistic conventions alone but by patterns of social interaction, practical interests, and mutual accountability.

Large language models do not participate in such practices. Like the "right hand" in Wittgenstein's analogy, they operate within a closed circuit. Although they are trained on linguistic artifacts produced within human language-games, they do not themselves inhabit the activities that give those artifacts their normative force. They can generate the textual tokens of an implication (the "deed of gift") but they cannot generate the "practical consequences." Consequently, their utterances cannot be evaluated as successful or unsuccessful moves within a practice. They can be assessed for coherence or relevance, but not for appropriateness in the sense required for implication. This highlights the critical flaw in arguments like Grindrod's (2024), who proposes that "linguistic intentionality relies on a pre-existing meaningful system," and therefore LLMs can "employ this pre-existing system in the same way that ordinary speakers do" (Grindrod, 2024, p. 15). By reducing language use to the mere employment of a system, such views overlook that the system derives its normative force entirely from the shared forms of life that LLMs cannot inhabit. Such a deficit is critical because human communication functions not merely as information exchange but as a mechanism for communal alignment. As Rosen and Dale (2024) observe in their analysis of online discourse, human interlocutors use illocutionary force to "maintain the concepts and ideas" of their community, actively sustaining the normative structure of the group (Rosen & Dale, 2024, p. 2871). This normative function is further evidenced by the dynamics of social distance. As Caucci and Kreuz (2012) found that speakers employ significantly more paralinguistic signals when interacting with friends than with strangers, suggesting that implication is calibrated by the degree of shared history and intimacy (Caucci & Kreuz, 2012, pp. 15–17). LLMs, lacking membership in any such community or relational standing with the user, cannot engage in this maintenance; they generate text that fits the pattern but carries no social weight. They are structurally "strangers" attempting to play a language game that requires the intimacy of "friends."

This limitation clarifies why the apparent pragmatic competence of language models remains superficial. When a human speaker implies something, the success of the implication depends on what is at stake in the interaction. The speaker risks being misunderstood, rebuffed, or held accountable. These risks are essential to the act. An implication that carries no practical consequences is not fully an implication at all. Language models, by contrast, incur no such stakes. Their outputs do not commit them to anything, nor do they alter their standing within a shared activity.

One might object that models increasingly function within social contexts and influence human decisions, thereby acquiring a kind of practical significance. Yet this influence is always mediated by human agents who interpret, endorse, or act on model outputs. The model itself does not occupy a role within the practice that would render it accountable. Any normativity associated with its utterances is parasitic on human participation in the language-game, not generated by the system.

Wittgenstein's emphasis on forms of life also helps explain why implication resists formalization. There is no algorithmic rule for when an utterance counts as a hint rather than a statement, or as irony rather than error. Such distinctions are drawn against a backdrop of shared habits and expectations. They are learned by participating in practices, not by mastering explicit rules. A system that lacks such participation cannot genuinely imply, even if it can reproduce the linguistic surface associated with implication.

Seen in this light, the limitations of large language models are not merely contingent shortcomings of current technology. They reflect a categorical difference between systems that manipulate symbols and agents who live within practices. Implication belongs to the latter domain. It presupposes a form of life in which saying

something is one way of doing something, and in which indirectness has a place because participants share an understanding of what matters.

The conclusion to be drawn is therefore not that large language models fail to approximate some abstract ideal of pragmatic competence, but that they operate in a different register altogether. They can simulate the outward forms of language-games without entering into them. Their apparent ability to imply is a projection sustained by human users who supply the background practices the system itself lacks.

This completes the primary argument of the paper. Implication requires intentional agency, indexical grounding, paralinguistic expression, and participation in shared practices. Large language models lack all four. However fluent their outputs may be, they remain incapable of implying anything in the full pragmatic sense. Before exploring the broader theoretical consequences of this conclusion, however, it is necessary to address potential counter-arguments. The following section considers objections regarding multimodality, role-playing, and behavioral indistinguishability, after which the study examines the wider implications of these findings for AI evaluation and the philosophy of language.

5. Multimodality and the Persistence of Simulation

The preceding sections have argued that large language models are structurally incapable of genuine implication. This section addresses three common objections to this claim. The first objection suggests that multimodal models, integrating vision, audio, and gesture, could overcome the limitations of text-based systems. If implication depends on expressive cues, why should systems equipped with sensory-motor channels not be capable of implying? Recent technical advancements seem to support this view. For instance, Li et al. (2025) propose an "LLM-enhanced Retrieval-Augmented framework" designed specifically for "making machines sound sarcastic" (Li et al., 2025, p. 1), while Qian et al. (2025) introduce "ProsodyLM," a model designed to "understand" prosody by converting vocal cues into discrete "prosody tokens" (Qian et al., 2025, p. 1).

However, these technical achievements inadvertently confirm the philosophical limitation. As Li et al.'s title reveals, the goal is to make machines *sound* sarcastic, not *be* sarcastic. This distinction is not merely semantic but ontological. Similarly, Qian et al.'s method of "tokenization" betrays the fundamental problem: by reducing the continuous, analog nuances of tone and emotion into discrete, computable symbols (tokens), these systems treat prosody as just another form of syntax to be calculated. The system achieves "sarcastic expressivity" not by forming an attitude toward a proposition, but by retrieving acoustic artifacts or predicting probability distributions of "loudness tokens." It generates the *sound* of sarcasm without the *social standing* required to validate it. This technical mimicry leads to what Başarslan and Büyükyılmaz (2025) identify as an "ontological illusion." In their hermeneutic analysis, they argue that generative AI creates a deceptive perception of agency because "a distinction must be drawn between imitating a function and possessing a property" (Başarslan & Büyükyılmaz, 2025, p. 1473). Similarly, Kömürcü (2025) characterizes this phenomenon as "Digital Sophism", arguing that digital agents produce persuasive discourse stripped of truth-seeking intent (Kömürcü, 2025, p. 2). Reinforcing this view, Büyükada (2025) describes the reasoning capabilities of LLMs as a "statistical illusion", likening their output to "baseless castles floating in the air", structures that mimic the form of intelligence without the foundation of genuine cognition (Büyükada, 2025, p. 147). Just as medical AI reduces empathy to a pattern (Sharon, 2025), LLMs reduce reasoning to what Altunya Sayan (2025) describes as "logical thinking devoid of wisdom",

emphasizing that formal validity does not equate to the purposeful deliberation required for genuine implication (Altunya Sayan, 2025, p. 15).

Ultimately, whether characterized as an ontological illusion, digital sophism, or reasoning devoid of wisdom, the conclusion remains the same: multimodal models simulate the *sensory surface* of implication without inhabiting the *moral and practical world* that gives it weight. They generate the acoustic artifacts of a participant but remain, structurally and ethically, strangers to the act of communication.

The difficulty with the multimodal view is that it conflates the availability of data with the possession of agency. Multimodality expands the range of inputs a system can process, but it does not establish a subject who acts within a shared world. Producing vocal contours or facial expressions is not equivalent to deploying them as expressions of intention. Without intentions to guide expressive choices, multimodal behavior remains a simulation. Moreover, embodiment involves occupying a position where actions have consequences. A system that generates signals without standing to be held responsible does not enter the space of pragmatic normativity required for implication.

A second line of argument appeals to role assignment. When a model is instructed to speak as a particular persona (e.g., "as a doctor"), does it not acquire a perspective from which implications can be generated? The answer is that role assignment supplies a narrative frame, not a situated standpoint. This distinction is rigorously analyzed by Perry (1979), who demonstrates that adopting a persona is merely a descriptive replacement, not a functional equivalent of self-reference. He argues that replacing "I" with a name or description destroys the explanatory force of the statement: "If I had said, in the manner of de Gaulle, 'I came to believe that John Perry is making a mess,' I would no longer have explained why I stopped and looked in my own cart" (Perry, 1979, pp. 4–5).

Mollema (2024) explicitly corroborates this insight, warning against the error of equating valid syntax with a speaking subject. Drawing on Italo Calvino's concept of the "literature machine," he argues that "the language model should, like Calvino's writing machine, be regarded as the instantiation of linguistic personality, rather than as an instance of the psychological narrator you yourself are. [...] The functionalist temptation [...] unreflectively presupposes the psychological narrator alongside the visible linguistic personality. But these two are quite separate" (Mollema, 2024, p. 47).

Consequently, to play a role is to simulate a position within discourse, not to occupy one in the world. A model speaking "as a doctor" does not stand in professional relations or bear the responsibilities that give medical speech its pragmatic force. This gap leads to what Sharon (2025) identifies as the "hollowing out" of concepts. In her analysis of empathetic medical chatbots, she argues that such systems redefine "empathy into a mere communication pattern" composed of "key words and expressions" (Sharon, 2025, p. 1). Just as medical AI reduces empathy to a syntactic performance stripped of the "real mental and emotional states" required for care (Sharon, 2025, p. 9), persona-based LLMs reduce implication to a stylistic artifact stripped of the agency required for meaning. Thus, persona-based prompting provides cues for interpretation but does not generate agency.

Finally, one might argue that if a model's output is indistinguishable from that of a human, the distinction between simulation and genuine implication is practically irrelevant. This objection, however, treats pragmatic competence as a behavioral property rather than a normative one. This fallacy is prefigured in Wittgenstein's

analysis of "reading machines" (Wittgenstein, 1986, para. 156), Wittgenstein asks us to consider a machine, or a human novice acting like one, that transforms written signs into sounds. Even if the output is indistinguishable from that of a fluent reader, Wittgenstein argues that the criteria for "reading" are not satisfied merely by the production of correct sounds. He distinguishes between the mechanical "passage of excitations" in a machine and the guided activity of a reading subject (Wittgenstein, 1986, para. 157). LLMs occupy precisely this functional role: they are sophisticated transducers of signs, calculating probable continuations without engaging in the lived activity of meaning. They function as "implication machines" that produce the surface artifacts of indirect speech without the internal agency that characterizes the act.

This behavioralist view overlooks what Startari (2025) identifies as the "paradox of authority": the fact that "authority does not vanish with reference" but is reinforced by the grammatical persistence of indexical markers (Startari, 2025, p. 10). Implication is not merely an output; it is an act performed within practices involving risk and responsibility. As Wittgenstein emphasizes, understanding is expressed in the way actions are embedded in forms of life, not merely in observable regularities. A system that reproduces linguistic patterns without participating in the practices that give them point does not acquire competence, regardless of how convincing its outputs appear. Indistinguishability at the level of simulation does not erase the structural difference in participation. Even proponents of communicative intent in LLMs, such as Attah (2025), concede that "there is an obvious enough difference-maker: it is the absence of mental agency... that makes the demarcation" (Attah, 2025, p. 19).

Conclusion

The argument of this paper is not that large language models are linguistically incompetent, but that their competence is of a fundamentally different kind than that of human speakers. They are engines of explicit text, capable of producing sophisticated locutionary forms but structurally debarred from the illocutionary acts that constitute genuine implication. This limitation is not an engineering problem to be solved by more data, larger context windows, or multimodal sensors. It is an ontological boundary.

Implication requires a subject who intends to mean more than what is said, a body that anchors reference to a shared world, and a voice that modulates force through expressive variation. It presupposes a life lived among others, where speech acts carry risks, commitments, and consequences. Large language models, existing as they do in the timeless, placeless, and stake-free space of statistical distribution, cannot enter this domain. As Wittgenstein's analysis of "reading machines" demonstrates, the production of correct outputs, however indistinguishable from human performance, does not amount to the exercise of agency.

The danger of the current moment lies in what Mollema (2024) identifies as the "functionalist temptation", the tendency to treat the simulation of implication as equivalent to the act itself. To succumb to this temptation is to risk "hollowing out" our normative concepts (Sharon, 2025), reducing empathy to a communication pattern and agency to algorithmic prediction. By insisting on the distinction between saying and implying, we defend not only a theoretical point about pragmatics but the integrity of human communication itself. Meaning is not merely the statistical alignment of tokens; it is a commitment to a shared world.

Bibliography | References

Altunya Sayan, H. (2025). Dijital çağda muhakeme ve hikmetten yoksun mantıksal düşünme üzerine bir soruşturma. In C. Baba (Ed.), *Dijital Çağda Mantık ve Düşüncenin Dönüşümü: Yeni Paradigmalar*. Palet Yayınları.

Attah, N. O. (2025). Do language models lack communicative intentions? *Synthese*, *205*(5), 187. https://doi.org/10.1007/s11229-025-05022-6

Austin, J. L. (1962). *How to do things with words* (M. Sbisá & J. O. Urmson, Eds.). Clarendon Press.

Başarslan, B., & Büyükyılmaz, Y. (2025). Is it possible to engage in dialogue with generative AI technologies? *Beytulhikme An International Journal of Philosophy*, *15*(4), 1473–1496. https://doi.org/10.29228/beytulhikme.86607

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Benveniste, E. (1971). *Problems in general linguistics* (M. E. Meek, Trans.). University of Miami Press.

Büyükada, S. (2025). Mantık, akıl yürütme ve yapay zeka. In C. Baba (Ed.), *Dijital Çağda Mantık ve Düşüncenin Dönüşümü: Yeni Paradigmalar*. Palet Yayınları.

Caucci, G. M., & Kreuz, R. J. (2012). Social and paralinguistic cues to sarcasm. *Humor*, *25*(1), 1–22. https://doi.org/10.1515/humor-2012-0001

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics, Volume 3: Speech Acts* (pp. 41–58). Academic Pr.

Grindrod, J. (2024). Large language models and linguistic intentionality. *Synthese*, *204*(2), 71. https://doi.org/10.1007/s11229-024-04723-8

Gubelmann, R. (2024). Large language models, agency, and why speech acts are beyond them (for now) – a Kantian-cum-pragmatist case. *Philosophy & Technology*, *37*(1), 32. https://doi.org/10.1007/s13347-024-00696-1

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Harnad, S. (2025). Language writ large: LLMs, ChatGPT, meaning, and understanding. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1490698

Kaplan, D. (1989a). Afterthoughts. In J. Almog, J. Perry, & H. Wettstein (Eds.), *Themes From Kaplan* (pp. 565–614). Oxford University Press.

Kaplan, D. (1989b). Demonstratives. In J. Almog, J. Perry, & H. K. Wettstein (Eds.), *Themes from Kaplan* (pp. 483–563). Oxford University Press.

Kömürcü, K. (2025). Dijital sofizm ve mantık. In C. Baba (Ed.), *Dijital Çağda Mantık ve Düşüncenin Dönüşümü: Yeni Paradigmalar*. Palet Yayınları.

Li, Z., Zhang, Y., Gao, X., Nayak, S., & Coler, M. (2025). *Making machines sound sarcastic: LLM-enhanced and retrieval-guided sarcastic speech synthesis* (arXiv:2510.07096; Version 1). arXiv. https://doi.org/10.48550/arXiv.2510.07096

Mandelkern, M., & Linzen, T. (2024). Do language models' words refer? *Computational Linguistics*, *50*(3), 1191–1200. https://doi.org/10.1162/coli_a_00522

Mollema, T. (2024). Social AI and the temptation of equating Wittgenstein's language-user with Calvino's literature machine. *International Review of Literary Studies*, *6*(1), 39–55. https://doi.org/10.53057/irls/2024.6.1.4

Oğuz, M., Bakman, Y. F., & Yaldiz, D. N. (2025). Un-considering contextual information: Assessing LLMs' understanding of indexical elements. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 23410–23427). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.1203

Özdil, M. S. (2025). *Non-linguistic elements and issues of reference in logic* (Y. E. Akbay, Ed.). Palet Publications.

Perry, J. (1979). The problem of the essential indexical. *Noûs*, *13*(1), 3–21. https://doi.org/10.2307/2214792

Qian, K., Fan, X., Ni, J., Shechtman, S., Hasegawa-Johnson, M. A., Gan, C., & Zhang, Y. (2025). *ProsodyLM: Uncovering the emerging prosody processing capabilities in speech language models*. Second Conference on Language Modeling. https://openreview.net/forum?id=uBg8PClMUu#discussion

Rosen, Z. P., & Dale, R. (2024). LLMs don't "do things with words" but their lack of illocution can inform the study of human discourse. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0). https://escholarship.org/uc/item/25k7z0mz

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

Sharon, T. (2025). Technosolutionism and the empathetic medical chatbot. *AI & Society*. https://doi.org/10.1007/s00146-025-02441-4

Startari, A. V. (2025). Indexical collapse: Reference disappears, authority remains in predictive systems. *Elsevier BV*. https://doi.org/10.5281/zenodo.17228239

Wittgenstein, L. (1986). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Basil Blackwell.