Do LLMs Speak? Framework-Relativity and Linguistic Participation

Tanner Leighton

1. Introduction

Large language models (LLMs) pose a pressing question for the philosophy of language and technology: can systems that lack minds, intentions, and conscious experience nevertheless genuinely participate in linguistic practice? The question matters because LLMs increasingly mediate communication, structure inquiry, and coordinate social action, yet theories built around paradigmatically human speakers often seem ill-suited to artificial systems that produce fluent discourse without the mental architecture typically associated with speakers.

Consider a concrete case. When an LLM generates 'The capital of France is Paris' in response to a query, has it said something? Users rely on the answer to plan travel, answer quiz questions, and update beliefs; it licenses familiar inferences ('Therefore Paris is in France') while blocking others ('Therefore Lyon is the capital'), and it adjusts under challenge when presented with corrections. Yet the model forms no beliefs, entertains no intentions, and possesses no awareness; it produces text by exploiting statistical patterns in language data, not by performing speech acts as a responsible, minded agent. Is this linguistic participation—or merely sophisticated simulation?

Philosophers have offered sharply divergent answers. Critics such as Bender et al. (2021) characterize LLMs as 'stochastic parrots' that manipulate linguistic form without grasping meaning, producing text that resembles meaningful speech without genuine understanding, reference, or communicative intent; Hattiangadi and Schoubye (2025) go further, arguing that LLM outputs are literally meaningless. Others adopt more permissive views. Shanahan (2022) grants that LLMs lack human-like understanding and inner experience but argues that this need not disqualify them from linguistic participation. Van Dijk et al. (2023), Lappin (2024), and Grindrod (2024) advance related positions, emphasizing LLMs' grammatical mastery and context-sensitive performance while denying that such performance entails the communicative intentions, subjective awareness, or reflexive agency characteristic of human speakers.

These disagreements rest less on disputes about the functional facts than on rival interpretations of what those facts amount to. Most parties agree that LLMs lack mental states and agency, process statistical patterns, and generate outputs that support coordination and respond to correction. The fault lines emerge when familiar metasemantic frameworks are brought to bear. Intentionalist accounts, following Grice (1957), make communicative intentions constitutive of meaning; causal-informational accounts, following Dretske (1981) and Fodor (1987), ground content in reliable causal covariation with environmental conditions; teleosemantic accounts, following Millikan (1984), tie representation to biological proper functions; and pragmatist accounts, following Brandom (1994) and Price (2013), explain meaning in terms of inferential roles and social practice. Each framework imposes different metasemantic requirements on what grounds meaning—and, by extension, on what is required for genuine linguistic competence.When theorists treat those requirements as constitutive, LLMs typically fail to qualify: their outputs get classified as meaningless or as mere simulation rather than genuine participation.

This paper argues that, in the LLM case, we need not—and should not—wait for deeper metasemantic disputes to be resolved before reaching substantive, normatively important conclusions about linguistic participation. Many entrenched disagreements arise when theorists elevate framework-specific commitments—Gricean speaker intentions, environment-involving causal links, biological proper functions—into gatekeeping constraints on linguistic competence. The problem is not that these richer frameworks are misguided; they serve genuine explanatory purposes in their home domains. The problem lies in treating such commitments as constitutive requirements on linguistic competence, thereby excluding—by definitional fiat—systems that demonstrably engage in coordination, inference, and norm-sensitive interaction at the functional level.

My positive claim is that LLMs already qualify as *norm-sensitive* participants in inferentially structured linguistic practice—even if they do not satisfy the further grounding conditions that various metasemantic frameworks treat as constitutive of meaning (Gricean intentions, causal–informational links, teleosemantic proper functions), and even though they lack the reflexive capacities characteristic of *norm-responsible* agents. The key contrast is between responsiveness to norms—tracking, adjusting, and repairing in ways that sustain coordination—and ownership of norms—treating one's claims as commitments for which one

can be held accountable. §5 develops this distinction and shows how functional participation can be genuine without amounting to full responsible agency.

Once this structure is explicit, two tasks become possible. First, we can distinguish what is widely accepted about LLMs—their role in coordination, inference, and norm-sensitive interaction—from the further commitments that different metasemantic frameworks layer on top. Second, we can ask, relative to specific explanatory and normative aims, whether those additional commitments earn their keep in the LLM case or merely generate framework-relative verdicts of exclusion. Do we need Gricean nested communicative intentions, environment-involving causal relations, or teleosemantic proper functions to explain how LLMs enable coordination, structure inquiry, and respond to correction? For these purposes, I argue, the answer is no.

This motivates a methodological stance that neither declares a winning theory of meaning nor dismisses the debate as merely verbal. Instead, it identifies the pragmatist core that rival frameworks already presuppose, resists metaphysical inflation beyond what the phenomena require, and treats framework-specific elaborations as optional theoretical resources rather than universal constraints. §1.1 develops this stance under the label *irenic pragmatism*; subsequent sections apply it to the LLM case.

1.1 Introducing Irenic Pragmatism

Irenic pragmatism is a diagnostic and methodologically conservative way of approaching philosophical disagreement. It does not introduce a new rival theory of meaning to compete with Gricean, causal, teleosemantic, or inferentialist accounts. Instead, it offers a way of navigating disputes among such accounts when they threaten to become intractable or to overshadow the practical questions that motivated them.

The stance is structurally analogous to Arthur Fine's *Natural Ontological Attitude* (NOA) in the realism debate. Fine's key move was to note that realists and anti-realists share a common 'natural ontological attitude' toward successful scientific theories: they rely on them, use their results, and treat their claims as working assumptions in practice, even while disagreeing about how to characterize that practice in metaphysical terms. NOA recommends

that we not treat realist or anti-realist metaphysical add-ons as compulsory for doing or understanding successful science.

The parallel in the LLM debate is direct. Competing frameworks converge on a common functional core of linguistic practice—coordination, inferential articulation, norm-sensitive uptake, and repair—while disagreeing about what, if anything, must be added to that core for genuine linguistic competence. Intentionalists emphasize higher-order communicative intentions; causal-informational theorists emphasize environment-involving causal relations; teleosemanticists emphasize proper functions; pragmatist and inferentialist approaches emphasize the social-inferential roles themselves. The irenic move is to begin from this shared practice-level ground, adopt a minimalist default that does not build any robust framework-specific add-ons into the baseline account, and treat richer commitments as overlays to be introduced only when a specific explanatory or normative aim warrants them. When an overlay is proposed, the question is whether it earns its keep for the task at hand—or merely generates a framework-relative verdict of exclusion. That methodological stance shifts the guiding question: instead of asking whether LLMs 'really' or 'genuinely' speak, we ask in what senses (if any)—and for which purposes—they participate in linguistic practice.

This approach neither defines linguistic competence by fiat in Gricean terms nor insists that longstanding disputes among intentionalists, teleosemanticists, and inferentialists be resolved before saying anything substantive about LLM participation in language. Yet two objections to any such permissive stance loom especially large: first, that genuine meaning requires communicative intentions of the kind Grice analyzed (the Communicative Intention Argument); second, that LLM outputs lack the world-connections needed for genuine semantic content (the No Meaning Charge). The remaining sections put irenic pragmatism to work: §2 diagnoses how the Communicative Intention Argument builds Gricean requirements into its premises; §3 addresses the No Meaning Charge by introducing Huw Price's distinction between i- and e-representational roles; §4 explains the technical mechanisms underlying LLM behavior; and §5 argues that LLMs can be norm-sensitive participants without qualifying as norm-responsible agents.

2. The Communicative Intention Argument

One of the most influential objections to the linguistic status of large language models—pressed by critics such as Bender et al. (2021)—is what I will call the Communicative Intention Argument (CIA). In its simplest form:

1. Genuine linguistic competence requires communicative intentions.
2. LLMs lack communicative intentions.
3. Therefore, LLMs lack linguistic competence.

The CIA's intuitive pull is obvious. It is natural to think that speaking is not merely producing well-formed strings, but *meaning* something by them—doing something intentionally with words. Still, the argument's real weight falls almost entirely on premise (1). That premise is not a neutral platitude; it is a substantive theoretical commitment, most naturally read as importing a broadly Gricean picture of speaker meaning into the very *entry conditions* for linguistic participation. To evaluate the CIA, we therefore need to make the Gricean background explicit rather than letting it operate as an unargued definitional constraint.

Grice (1957, 1969) sought to ground speaker meaning in intentional psychology. In his canonical formulation (1957, 381–83), for a speaker S to mean something by uttering *x* is for S to intend:

1. that an audience A form a certain belief or response;
2. that A recognize (at least partly, perhaps inferentially) that S intends A to form that belief or response; and
3. that A's recognition of this intention partly explain why A forms the belief or response.

This is a nested-intention analysis: meaning is not simply producing an effect, but producing it through the audience's recognition of the speaker's intention. If I shout 'Fire!' in a crowded theater, my utterance counts as meaningful, on Grice's view, because I intend the audience to believe there is a fire (and to evacuate), intend them to recognize that intention, and intend that recognition to play a role in producing their response. If I blurt 'Fire!' because of a nervous tic, then—even if the audience reacts the same way—the utterance does not (strictly) count as speaker-meaningful, since the relevant intention structure is missing.

Grice's analysis is illuminating for many paradigmatically human phenomena—conversational implicature, indirect requests, strategic deception—where 'what is meant'

depends on sophisticated audience-management. But once that Gricean machinery is treated as *constitutive of linguistic competence as such*, the CIA follows almost immediately. Contemporary LLMs do not plausibly instantiate the relevant higher-order psychological states; they do not form Gricean communicative intentions, much less the nested intentions Grice specifies. So if Gricean speaker meaning is built into the definition of linguistic competence, LLMs are excluded by definition.

At this point, it helps to notice that the CIA often draws rhetorical strength from a slippage among three related but distinct notions:

- **Communicative intention (Gricean, singular):** the specific nested intention structure just described.
- **Intentions (ordinary, plural):** everyday goals and plans—intending to finish a paper tonight, intending to signal a driver with a raised hand, intending to reassure a friend.
- **Intentionality (philosophical):** aboutness or directedness—being *of* or *about* objects, properties, or states of affairs.

Premise (1) needs the first notion: *Gricean* communicative intention. But in ordinary discussion it is easy to slide from 'LLMs lack communicative intentions' to the much broader-sounding claim that 'LLMs lack intentions' (full stop), and from there to the further suggestion that their outputs cannot be 'intentional' in the sense of being *about* anything at all. That slide is not innocent. One can consistently hold that LLMs lack Gricean communicative intentions while still allowing that their outputs can function as *about-the-world* moves within human inquiry and coordination—especially given the robust role of hearer uptake, correction, and downstream use.

There is also a further notion of 'intention' that matters in the LLM case and is often left implicit:

- **Designer-/deployer-level intentions:** the intentions of human users, engineers, and institutions who build, prompt, and deploy these systems, and whose communicative aims can be carried—sometimes quite directly—by the model's outputs.

This matters because even if LLMs themselves lack the nested Gricean structure, many real-world LLM interactions are saturated with human intentions at the prompt and deployment level. When a user asks an LLM to draft an email, summarize an article, or explain a concept to a student, the communicative intentions are plainly present—just not necessarily *inside the model*. A strict Gricean can respond that this makes the LLM a mere *instrument* of a human

speaker, not a speaker in its own right. Fine. But then the CIA is no longer functioning as a knockdown argument that 'LLMs don't participate in language'; it becomes, at most, an argument about a specific *kind* of participation—speakerhood in the full Gricean sense—while leaving open that LLM outputs can be meaningful *in use*, within a broader practice.

Even setting aside this equivocation, premise (1) risks overgeneralizing from paradigmatic cases. Paradigmatic Gricean cases—deliberate assertion, strategic implicature, deception—are not the whole of linguistic practice. Ordinary language use is more heterogeneous than a Grice-first lens suggests. Consider:

- **Early child language:** children produce and respond to utterances long before they can plausibly be credited with the sophisticated meta-intentional structures Grice requires. A two-year-old who says 'want' while pointing at a cookie communicates intelligibly despite lacking anything like the intention that a caregiver recognize that intention, or that such recognition partly explain the response (Lohmann and Tomasello 2003; cf. Attah 2025).
- **Phatic speech:** much everyday talk is affiliative rather than belief-inducing ('How's it going?', 'Nice to see you'). These utterances are paradigmatically linguistic even when they do not aim to produce a particular belief through intention-recognition.
- **Scripted and institutional speech:** actors, priests, and officials routinely utter words whose meaning depends on social and institutional norms rather than an individual's private communicative psychology. 'I now pronounce you…' and 'I hereby sentence you…' are meaningful partly because of the role and the practice, not because of the speaker's nested intention architecture.
- **Animal signaling (to the extent it counts as linguistic):** many signaling systems plausibly transfer information and coordinate behavior without Gricean communicative intentions.

These examples do not show that Grice's framework is false or useless. The point is diagnostic: Gricean intentionalism captures an important, philosophically rich pattern within a wider ecology of linguistic practice. The mistake is to treat that elaboration as a universal constitutive requirement for 'language as such'—especially in a context where doing so functions less as an explanation than as a gatekeeping definition.

A related issue is the CIA's speaker-centered asymmetry. Grice privileges meaning as flowing from speaker to hearer: speaker intentions are primary; hearer uptake is derivative. But

many practical contexts place weight on uptake and coordination. Consider a scribbled note: 'It's raining.' If it was produced absentmindedly with no communicative intention, a strict Gricean verdict is that it means nothing. Yet a passerby can read it, reasonably treat it as weather information, and coordinate accordingly. Whatever we ultimately say about speaker meaning, this shows that a speaker-first conception is not the only legitimate starting point for theorizing the role language plays in social coordination and inquiry.

From an irenic-pragmatist perspective, the CIA exemplifies a recurring dialectical pattern: non-minimal commitments of an elaborated framework are elevated into constitutive requirements, and then deployed to exclude contested cases by definitional fiat.

Gricean intentionalism begins from a plausible functional core—language as a norm-governed practice of coordinating belief and action—and adds nested communicative intentions to explain a subset of human-centered phenomena. Those additions may be indispensable for explaining intention-laden aspects of rational agency, including moral accountability and certain forms of communicative success. But they are not plausibly constitutive of linguistic practice as such. And in the LLM debate, insisting on them as necessary conditions does little explanatory work for the narrower question at hand: how systems without minds can nonetheless contribute, in a norm-sensitive way, to discursive coordination.

The upshot is not that Grice is 'wrong'. It is that Gricean communicative intention is ill-suited as a universal entry condition on linguistic participation—especially when our target is LLMs' functional role in inquiry and coordination. If one builds nested communicative intentions into the constitutive conditions for speaker meaning, then LLMs will of course count as non-speakers. But the irenic pragmatist core yields a different (and, for present purposes, more illuminating) verdict—not by denying what LLMs lack, but by starting from what they do: they generate contributions that can be taken up, challenged, corrected, and integrated into ongoing discursive activity.

3. The Meaning Question

Having shown how the Communicative Intention Argument elevates a Gricean elaboration into a definitional constraint on linguistic *competence*, we now confront a more radical challenge: whether LLM outputs possess *meaning* at all. Where the CIA denies LLMs the status of speakers, the No Meaning Charge (NMC) goes further, denying that their outputs are even the

right kind of thing to figure in linguistic practice. On this view, an LLM does not merely fail to speak; it emits strings that only resemble contentful discourse.

The charge comes in several forms. Bender et al. (2021) characterize LLMs as 'stochastic parrots' that manipulate linguistic form without grasping content—producing fluent text while lacking genuine aboutness or communicative significance. Hattiangadi and Schoubye (2025) push further, arguing that LLM outputs are literally meaningless: not utterances with semantic content, but strings that carry no content of their own. More generally, critics contend that without the right connections between a system and its environment—whether higher-order speaker intentions, environment-involving causal relations (often via inheritance)[1], or proper functions—there is nothing in virtue of which the system's outputs can determinately refer, describe states of affairs, or express propositions with truth conditions. At best, on this picture, LLM outputs are useful prompts for our interpretive activity: any apparent meaning derives entirely from what users project onto them.

The irenic pragmatist accepts much of the diagnosis while rejecting the inference. LLMs lack intentions, biological functions, and (in the relevant sense) direct causal coupling to the world. But it does not follow that their outputs cannot be meaningful. Many paradigmatic bearers of content—written inscriptions, diagrams, maps, mathematical expressions, scientific models—lack intentions and consciousness as well. Their content need not be an intrinsic psychological property of a speaker. Even granting a distinction between derived and underderived content, inscriptions, maps, and models can be determinately contentful in the familiar sense that they license inferences, are assessed for correctness, and guide coordinated

[1] Causal–informational theories, as developed by Dretske (1981) and Fodor (1987), are most naturally understood as theories of underived mental content—percepts, detectors, and belief states whose content is fixed by reliable causal relations to environmental conditions. When extended to public language, such theories typically proceed indirectly, treating linguistic assertions as derivatively contentful in virtue of their place in testimonial and communicative chains that ultimately trace back to causally grounded agents. My claim is not that causal–informational theorists cannot tell a story about linguistic meaning, but that doing so requires a substantive account of how causal grounding is preserved across transmission—through deference, testimony, and other social-linguistic mechanisms—rather than 'washing out' as content passes from perceivers to speakers to texts and, now, to statistical training on text corpora. That preservation-of-grounding thesis goes beyond the core causal–informational content-fixing claim and is precisely where applications to LLMs become theoretically contested. Irenic pragmatism, by contrast, treats social practices of uptake, correction, and coordination as *constitutive of symbolic content,* not as mechanisms for *transmitting content grounded elsewhere*: the relevant question is whether outputs are integrated into norm-governed coordinative practice, not whether they inherit an ultimate causal anchor. Both views acknowledge that symbolic practices depend on causally grounded agents interacting with the world; the difference is that causal–informational theories treat those causal relations as *content-fixing*, whereas irenic pragmatism treats them as *regulative and practice-constraining*—explaining why meaning is not arbitrary once instituted, rather than grounding meaning in the first place.

action within norm-governed practices of use, correction, and downstream reliance. LLM outputs plausibly belong to the same broad family: symbolically contentful artifacts whose content derives from uptake within inferentially articulated practices.

This reframes the argumentative burden. The central question is not whether LLMs satisfy some antecedently fixed requirement of 'non-derivative' meaning. The question is which further conditions—if any—must be added to practice-based uptake to underwrite the strongest, most demanding kinds of world-directed content. Even if uptake suffices for practical, practice-relative content, do we also need additional grounding relations to secure robust reference and truth-conditional answerability?

Framed this way, the NMC is not primarily an empirical dispute about what LLMs do. It is a dispute about what meaning requires: whether participation in inferentially structured practices suffices for semantic content, or whether genuine content requires additional world-involving constraints. What is often missed is that these two issues—a symbol's role within inferential practice and its answerability to the world—can come apart. An item can be fully integrated into patterns of reasoning, challenge, and repair while leaving open harder questions about how (or whether) it is grounded in the world.

To clarify what is at stake, I introduce and refine a distinction—due to Huw Price—between two representational roles. In its initial form, the distinction separates inferential role within practice from robust world-directed tracking. Once that separation is in view, we can see why the NMC often gains its force by conflating these two dimensions: treating the absence of one (robust world-grounding) as if it entailed the absence of the other (inferentially articulated content).

3.1 Price's Distinction

Price (2013, 36) distinguishes two broad families of approaches to representation in philosophy of mind and language. On one hand, we have what he calls *e-representation* (for 'environmental' or 'externalist'): the paradigm of representation grounded in environment-tracking relations. Think of a fuel gauge needle covarying with gasoline levels, a barometer reading tracking air pressure, or—in biological cases—internal states that evolved to correlate reliably with environmental features like the presence of prey or predators. In these cases,

representational content depends on systematic correlations, typically causal, between representing states and represented conditions. The system-world link comes first.

On the other hand, we have what Price calls *i-representation* (for 'inferential' or 'internalist'): representation grounded in functional or inferential role within a cognitive or discursive system. Something counts as a representation in virtue of its position in an inferential network—what follows from it, what it follows from, what it enables or excludes. Here the internal functional role takes priority over environment-tracking relations.

This distinction is especially useful for the LLM debate because the central dispute is often framed as a dispute about meaning. Critics grant that LLMs produce fluent sentences but deny that those sentences are genuinely about anything—that they are 'mere form' without world-directed content. In Price's terms, the critics' worry is not primarily about whether LLM outputs participate in inferential patterns (*i-representation*), but about whether they can secure the relevant kind of world-link (*e-representation*).

3.2 Two Kinds of World-Link

Price's distinction between i-representation and e-representation provides crucial analytical leverage, but applying it to the LLM debate requires refinement. The central worry about LLM outputs concerns not whether they participate in inferential patterns—critics typically grant this—but whether they can be genuinely *about* the world. Yet 'aboutness' itself admits of importantly different modes.

To clarify the nature of disputes about LLM competence, I distinguish two fundamentally different modes of e-representation with distinct success conditions: *physical* e-representation and *symbolic* e-representation.

1. **Physical e-representation:** On the one hand, we have non-linguistic systems that track environmental features through direct, typically causal, coupling. A thermometer's mercury column tracking ambient temperature or a fuel gauge needle tracking gasoline levels are prime examples. These systems achieve counterfactual covariation mechanically—their representational success is intrinsic to their causal structure and operates automatically, continuing to function even in a world devoid of observers. The mercury would track temperature regardless of whether anyone reads the thermometer.

2. **Symbolic e-representation:** On the other hand, we have norm-governed symbolic systems that encode information about the environment through structured, interpretable forms rather than through direct causal coupling. Candidate examples include descriptive assertions ('The cat is on the mat'), maps, and scientific models—though which of these genuinely e-represent remains contested. Crucially, symbolic e-representations do not succeed through automatic causal coupling but through agent-mediated coordination practices. A map of Pittsburgh does not automatically update when a new street is built; agents must revise it. Similarly, the assertion 'The cat is on the mat' does not spontaneously change when the cat moves; agents must cease asserting it and assert something else. Symbolic representations achieve counterfactual sensitivity only through agent uptake and maintenance: if the environment were relevantly different, agents would (or should) produce, endorse, or use different symbols.

This agent-mediation is not contingent but constitutive of how symbolic e-representation works. Symbolic systems represent the environment by being embedded in norm-governed practices where agents adjust, revise, correct, and coordinate their use of symbols in response to ongoing interactions with the environment.

This distinction reveals something crucial: symbolic e-representations are not a separate category from i-representations but a special way of treating them. Every symbolic e-representation must first be an i-representation—an item embedded in inferential practices, playing a functional role in reasoning and discourse. What makes it *additionally* e-representational is how agents use it: whether they treat it as tracking or describing features of the environment.

But not any uptake will do. If mere treatment were sufficient, then anything a community happened to treat as world-directed would thereby count as e-representational—a consequence too permissive to be illuminating. We need a normative constraint that distinguishes warranted ascriptions of world-directedness from mere projection or pretense. This is where the notion of *entitlement* becomes essential.

With this challenge in view, we can state irenic pragmatism's minimal proposal precisely: *symbolic e-representations just are i-representations that agents are entitled to treat as world-directed within defeater-sensitive practices*. This is not merely an epistemology of when reliance is prudent. For symbolic systems, world-directedness is a constitutive normative

status: a representation is about the world insofar as it is governed by norms that make it answerable to how things stand—through challenge, defeat, retraction, and repair. The question of what a symbolic item is about is not separable from the question of what would count as getting it wrong and being required to withdraw it.

What does this entitlement require? Three conditions are built into the very notion of entitled world-directed treatment:

(a) Public answerability. Entitled e-representations operate within practices in which claims may be challenged, reasons demanded, and assessments offered. Entitlement here is not a matter of private conviction but of publicly recognizable standing to treat the representation as world-tracking—a standing that can be questioned, defended, or defeated. When I assert, 'The train leaves at 3 p.m.', I make myself answerable to anyone who might challenge that claim by consulting the schedule or observing the platform.

(b) Error-sensitivity. An entitled e-representation must be revisable, correctable, and counterfactually responsive. If the world were relevantly different, entitlement would lapse. This is where misrepresentation enters the picture—and it's crucial to see that on the minimal account, misrepresentation is not metaphysical failure to correspond but normative defeat of entitlement. When an assertion is later shown to be false, what happens is not that some mind-independent correspondence relation fails to obtain, but that the practice's correction mechanisms defeat continued reliance, requiring withdrawal and repair. Consider what error-sensitivity looks like in practice: if I assert 'The train arrives at 3 p.m.' and, by 3:15, no train has arrived, continued reliance is no longer warranted; retraction, revision of my plan, and investigation into whether I misread the schedule are in order.

(c) World-involving defeaters. The practice treats perceptual reports, instrument readings, independent cross-checks, and success or failure in action as potential defeaters—inputs that rationally require retraction, revision, or restriction of downstream reliance. This does not require direct causal coupling to referents, but it does require that the practice be structured by defeater-sensitive norms keyed to ongoing interaction with the environment.

Consider again the assertion 'The cat is on the mat.' This utterance first and foremost plays an i-representational role. It stands in inferential relations to other claims:

- It entails 'The cat is not in the kitchen'
- It entails 'Something is on the mat'
- Combined with 'The mat is in the living room,' it entails 'The cat is in the living room'

- It conflicts with ‘The cat is outside’

These inferential connections constitute its i-representational content and give it a functional role within reasoning and discourse. This inferential role is uncontroversial—all theoretical frameworks agree that the utterance has this functional status, that it licenses some inferences while blocking others.

What remains contested is whether this same utterance additionally e-represents—whether it is genuinely *about* a state of affairs in the environment, whether it has world-directed content that makes it answerable to how things actually stand with cats and mats.

Here competing frameworks diverge. The minimal pragmatist claim is not that an utterance becomes world-directed merely because a community happens to treat it that way, but that symbolic world-directedness is *instituted and sustained* through entitlement-conferring practices—norms that make assertions answerable to how things stand by licensing correction, withdrawal, and downstream reliance. When speakers use ‘The cat is on the mat’ to guide expectation and action, and when hearers can challenge it, demand reasons, and treat disconfirmation as grounds for revision, the utterance functions as an e-representation in the distinctively symbolic sense: it is embedded in a practice that treats it as accountable to the presence or absence of the cat on the mat. If the cat is in the kitchen, competent participants should retract the claim, revise their commitments, and alter the inferences they are prepared to draw—and crucially, this revision is not optional but required by the norms governing the practice. On this view, e-representational status is not ‘anything goes’ projection; it is a normative role within a practice whose standards of correction are themselves keyed to ongoing interaction with the environment.

Robust theoretical frameworks offer competing accounts of what entitlement to world-directed treatment requires—what conditions must be satisfied for agents to be genuinely entitled (rather than merely presumptuous or mistaken) in treating an i-representation as about the world. These frameworks don’t deny that agent uptake matters; all acknowledge that symbolic representations function within social practices. What they insist is that entitled uptake requires additional grounding conditions to be independently satisfied. Consider the major alternatives:

- *Gricean intentionalism* holds that entitlement to world-directed treatment requires that the representation be produced by an agent with nested communicative intentions: intending to produce belief through recognition of that intention. On this view, one is

entitled to treat 'The cat is on the mat' as e-representational only when it originates from an agent with the relevant intention structure. Without such intentions, treating the utterance as genuinely world-directed exceeds one's entitlement—one mistakes an instrument's output for genuine assertion.

- *Causal-informational semantics* holds that entitlement requires reliable causal connections between symbolic tokens and their referents. The word 'cat' must be causally grounded—perhaps through perceptual encounters with cats, or through testimonial chains originating in such encounters. Without such causal links, treating the representation as world-directed is unwarranted—the symbols float free of environmental constraint.
- *Teleosemantics* holds that entitlement requires that representational states possess proper functions—typically biological functions derived from evolutionary selection—to covary with environmental features. One is entitled to treat a state as e-representational only if it has the right functional history. States that lack proper functions cannot succeed or fail at their representational task in the way genuine representations do.

These frameworks often overlap—a theorist might endorse both Gricean intentions and causal grounding as jointly necessary—but each highlights a different dimension of what entitled uptake supposedly requires. What unifies them is the insistence that practices of uptake, correction, and coordination, however sophisticated, are insufficient by themselves to ground genuine e-representation. Additional conditions concerning the system's internal architecture, causal history, or evolutionary provenance must be independently satisfied.

The upshot is ecumenical and clarifying. Most parties can accept that symbolic e-representation is i-representation that agents are entitled to treat as world-directed within defeater-sensitive practices; the substantive disagreement concerns what entitlement requires. The minimal pragmatist holds that defeater-sensitive practices are themselves sufficient for entitlement—that's what it is to be entitled to treat something as world-directed. Robust frameworks hold that additional grounding conditions must be met for the practice's entitlement-conferral to be genuine rather than merely presumptuous.

With this framework in place, we can see that the question 'Do LLM outputs have meaning?' conflates two distinct questions—and we're now positioned to answer them.

(1) **Do LLM outputs exhibit i-representational competence?**

Yes. LLM outputs participate in inferentially articulated linguistic networks. They preserve patterns of entailment, enable correction and repair, and integrate coherently with discourse practices. This competence does not require beliefs, intentions, or conscious understanding. The i-representational dimension concerns functional role within practice, not internal mental states. What matters is whether outputs appropriately preserve inferential structure, respond to challenges, and enable coordination—and LLMs demonstrably do this. §4 will explain the technical mechanisms that produce this competence, but the functional fact is already clear.

(2) **Do LLM outputs exhibit symbolic e-representational competence?**

Here the answer is contested and framework-relative, depending on what one takes entitled world-directed treatment to require. On robust frameworks requiring additional grounding properties, LLM e-representational status ranges from clearly denied to genuinely disputed. But for understanding how LLMs enable functional linguistic coordination—the question that motivated this inquiry—the irenic pragmatist need not adjudicate these further commitments. As §5 will show, entitled uptake practices can include contributions from norm-sensitive systems that lack norm-responsibility; what matters is whether the practice in which LLM outputs figure is itself structured by the relevant correction norms.

Crucially, irenic pragmatism is not answering the question 'Do LLMs *genuinely* participate in linguistic practice?'—as if there were a single standard of genuine participation and our task were to determine whether LLMs meet it. That framing presupposes that some framework is correct about what genuine participation requires. Instead, irenic pragmatism reframes the question: 'In what sense, if any, do LLMs participate?' This transforms a gatekeeping question into a diagnostic one—opening space for multidimensional characterization rather than binary verdicts of inclusion or exclusion.

The answer is that LLMs participate as norm-sensitive, i-representationally competent contributors to inferentially articulated practices—contributors whose outputs may or may not count as genuinely world-directed depending on which framework's standards one applies, and who lack the reflexive capacities characteristic of norm-responsible agents. This is not evasive. It identifies what LLMs actually do and don't do without prejudging contested metasemantic questions that remain genuinely open.

What remains to be shown is *how* systems trained by statistical learning come to exhibit i-representational competence—and why their mechanisms leave traditional grounding

relations persistently contested. The next section offers a brief technical interlude, explaining how transformer architectures produce norm-sensitive participation in inferential practice without requiring mental states, and why this architecture neither straightforwardly satisfies nor straightforwardly fails the grounding conditions that robust frameworks privilege.

4. Technical Interlude: How LLMs Work

Modern LLMs are built on the *transformer architecture* (Vaswani et al. 2017), a neural network design that processes sequences of tokens (roughly corresponding to words or word pieces). The key innovation enabling sophisticated linguistic behavior is the *attention mechanism*.

4.1 Transformer Architecture and Attention Mechanisms

Attention allows the model to dynamically weigh the relevance of different parts of the input when generating each part of the output. When processing 'The cat sat on the mat because it was comfortable', the model learns to attend strongly to 'cat' when generating 'it', rather than 'mat'. This attention pattern isn't programmed but emerges from training on massive text corpora where pronouns reliably refer to subjects rather than objects in similar constructions.

Attention operates across many deep, stacked layers and multiple 'heads' per layer, each learning different attention patterns. This allows the model to capture complex dependencies simultaneously:

- **Syntactic relationships:** subject-verb agreement, modifier-head relations, clause boundaries
- **Semantic relationships:** word meanings in context, conceptual associations, domain-specific usage
- **Pragmatic relationships:** discourse coherence, anaphora resolution, register appropriateness

The transformer processes text as token sequences, with each token represented as a high-dimensional vector (thousands of dimensions). Similar tokens occupy nearby positions in this vector space. Through training, the model learns to transform input token sequences into output sequences that are statistically probable continuations based on patterns in training data.

Critically, this architecture enables the i-representational competence we identified in §3. The attention mechanism naturally captures inferential structure: if 'X is a mammal' reliably precedes inferences to 'X is warm-blooded' in training data, the model learns this pattern. When the model generates 'Whales are mammals', its internal representations activate associated concepts (warm-blooded, live birth, nursing) that make subsequent inferences like 'Therefore whales are warm-blooded' highly probable. This isn't reasoning in the human sense but statistical tracking of inferential patterns—which is precisely what i-representational competence requires.

4.2 Training Process: Pre-training and Fine-tuning

LLMs undergo two main training phases that explain both their capabilities and limitations:

The first phase is *pre-training*, which uses self-supervised learning. The model is trained on massive text corpora (hundreds of billions to trillions of tokens from books, websites, scientific papers, code repositories, etc.) with a deceptively simple objective: predict the next token given the previous tokens. This is autoregressive language modeling.

For example, given 'The capital of France is _____', the model learns to assign high probability to 'Paris' because that completion appears reliably in the training data. Importantly, the model does not 'know' that Paris is the capital of France in the way a person might; instead, it has learned statistical regularities that make 'Paris' the most probable continuation in many contexts.

Through this process, the model implicitly acquires:

- **Grammar and syntax:** which token sequences are well-formed
- **Semantic relationships:** which concepts typically co-occur
- **World knowledge:** factual patterns that appear in training texts (though not necessarily current or accurate)
- **Discourse structure:** how conversations and documents are typically organized

The second phase is *fine-tuning,* which combines supervised learning with human preference modeling (often under the heading of RLHF: Reinforcement Learning from Human Feedback). After pre-training, the base model is adapted to be more helpful, safe, and instruction-following. This typically involves:

1. **Supervised fine-tuning:** Training on curated examples of high-quality responses to diverse prompts, teaching the model patterns of helpful assistance.
2. **Reinforcement Learning from Human Feedback (RLHF):** Human evaluators rate model outputs across many dimensions (accuracy, helpfulness, harmlessness, following instructions). The model learns to generate responses that receive higher ratings.

These preference-based methods are largely responsible for the highly refined conversational behavior we observe. When human raters consistently prefer outputs that correct errors, express uncertainty appropriately, or adjust tone and register to match the prompt, the model learns to assign higher probability to token sequences that exhibit these features. For example, in contexts where a user points out an error, the model will tend to produce continuations like 'You're right, I apologize' followed by a correction, because such patterns were repeatedly reinforced during fine-tuning. The resulting behavior can look like sophisticated self-correction and conversational repair, but at bottom it reflects learned regularities in how human evaluators rewarded certain response patterns over others.

4.3 Why This Produces I-Representational Competence

The training process explains how LLMs exhibit i-representational competence—participation in inferentially articulated linguistic networks—despite lacking mental states.

First, inferential structure emerges from statistical patterns. Logical relationships appear so consistently in training data that the model learns to preserve them. The model tracks inferential dependencies not through logical reasoning but through learned correlations—yet this suffices for i-representational participation. When generating 'Whales are mammals', the model has high probability of also generating inferences like 'Whales are warm-blooded' because these patterns co-occur reliably in training texts.

Second, the attention mechanism enables context-sensitivity across long sequences. When a user asks 'What about the capital?' after discussing France, the model attends to 'France' from context and generates 'Paris.' This enables sustained inferential trajectories across multiple turns—the hallmark of i-representational competence.

Third, RLHF training produces appropriate responses to correction. When users identify errors, the model learns to produce tokens matching repair sequences: acknowledging the error

and adjusting subsequent outputs. This is a learned pattern that enables the model to participate in norm-governed corrective practices without understanding correction reflexively.

Finally, because the model learns patterns at multiple levels (sub-word tokens, words, phrases, sentence structures), it achieves compositional productivity. It can produce novel, coherent combinations never seen in training, extending its i-representational outputs beyond the training distribution.

These mechanisms explain a crucial point: i-representational competence does not require the internal mental architecture (beliefs, reasoning, understanding) we associate with human linguistic competence. The functional role of tracking inferential relations can be realized through statistical learning over massive corpora.

4.4 Why Standard Grounding Remains Contested

Understanding the training process clarifies why LLMs' relationship to standard grounding facts remains contested across robust theoretical accounts:

- **Proper functions (teleosemantics):** The model was trained via gradient descent to predict token sequences, not shaped by biological evolution to track environmental features. Whether this matters depends on contested questions within teleosemantics. LLMs lack biological proper functions derived from evolutionary selection, but they have design functions—designers intended them to be informative and useful. Whether design functions can ground representation the way biological functions do divides teleosemantic theorists (Millikan 1995; Papineau 2022). The model has no evolutionary history connecting its outputs to environmental features, but artifact functions might suffice for some explanatory purposes.
- **Causal grounding (causal-informational theory):** The model has no perceptual or motor coupling to the world—it processes tokens, not environmental stimuli. However, it was trained on texts produced by causally-grounded speakers. Whether reference can be inherited through such testimonial chains is disputed. Children learn 'electron' from textbooks without direct causal contact with electrons, inheriting reference through testimony. Whether statistical training on massive text corpora preserves causal grounding in the relevant way divides causal theorists. The model's 'knowledge' comes

entirely from text, but those texts were written by agents with perceptual-motor coupling to referents.

- **Speaker intentions (Gricean framework):** The model has no mental states, let alone nested communicative intentions. This straightforwardly denies LLMs speaker status within Gricean intentionalism. The model produces statistically probable next tokens given context. Users may treat these outputs as if they were produced by an intentional agent (and this treatment enables uptake-based coordination), but this is user interpretation, not model intention.

This technical review reinforces the framework-relativity diagnosis developed above. Disputes about LLM status reflect not just which framework to adopt, but contested commitments about necessary grounding conditions within robust theoretical accounts. While these frameworks diverge on the necessary conditions, they converge on the functional facts: LLMs participate in coordination through agent-mediated uptake practices. The choice of how to theoretically elaborate these functional facts—whether grounding conditions are satisfied, necessary, or can be bracketed—remains a matter of theoretical choice.

5. From Functional Participation to the Limits of Agency

5.1 The Functional Coordination Core

Having developed the framework-relativity diagnosis in §§1–4 and reviewed the technical mechanisms underlying LLM behavior in §4, we can now articulate what functional capacities enable linguistic participation. The primary function of language is to facilitate *social coordination*—the alignment of behavior, belief, and expectation across agents. This coordinative function provides common ground across competing theoretical frameworks.

This leads to what I call the *functional coordination core*:

Functional Coordination Core: Linguistic exchanges achieve coordinative work through capacities for *uptake*, *repair*, and *correction*. Systems that exhibit these capacities thereby participate in the inferential and coordinative practices that constitute linguistic activity—regardless of whether they possess the mental states, biological functions, or communicative intentions that various theoretical frameworks might additionally require for full linguistic agency.

This formulation deliberately avoids essentialist claims about what 'really counts' as linguistic competence. It identifies functional capacities enabling participation in linguistic coordination without settling metaphysical disputes about whether such participation constitutes 'genuine' communication. Different frameworks can elaborate different requirements beyond this core, serving their respective explanatory purposes.

The three core capacities work together: *uptake* involves processing and responding appropriately to linguistic inputs within context, tracking inferential and pragmatic norms; *repair* involves recognizing and addressing communicative breakdowns through clarification and reformulation; *correction* involves adjusting outputs when shown to violate inferential, factual, or pragmatic norms.

LLMs demonstrably exhibit these capacities. Consider this exchange:

- **User:** 'What's Paris known for?'
- **LLM:** 'Paris is famous for the Eiffel Tower, Louvre Museum, Notre-Dame Cathedral, cuisine, fashion, and art.'
- **User:** 'I meant Paris, Texas.'
- **LLM:** 'Ah, my apologies. Paris, Texas is known for its replica Eiffel Tower with a cowboy hat, its role in the film 'Paris, Texas', and as a regional commercial center in Northeast Texas.'

This demonstrates uptake (recognizing the ambiguity once clarified), repair (adjusting to the correct interpretation), and correction (acknowledging the initial misunderstanding). The user successfully coordinates with the LLM despite the initial breakdown—exactly what linguistic practices enable. LLMs participate at this functional level, contributing to coordination, enabling joint inquiry, and supporting practical action.

5.2 The Norm-Sensitivity/Norm-Responsibility Distinction

Functional participation alone does not capture the full range of linguistic competence. To understand LLMs' distinctive position—genuine participants yet lacking full agency—we need a crucial distinction between *norm-sensitivity* and *norm-responsibility*.

Norm-sensitivity is the capacity to track and adjust to norms through learned patterns of behavior. A norm-sensitive system produces outputs that conform to linguistic norms (grammatical correctness, inferential coherence, pragmatic appropriateness, epistemic modesty,

responsiveness to correction) and adjusts when those norms are violated, but does so through mechanisms that don't require reflexive understanding of the norms or recognition of oneself as subject to them.

Norm-responsibility is the reflexive capacity to recognize oneself as bound by norms, to own one's commitments, and to be accountable for violations. A norm-responsible agent doesn't just produce norm-conforming behavior but grasps itself as a participant who can succeed or fail at meeting normative standards, who undertakes commitments through speech acts, and who bears consequences for those commitments.

This distinction is not a matter of degree but a difference in kind. It's the difference between a system that tracks norms and an agent who recognizes norms as binding on them. The former requires sophisticated pattern-matching and adjustment mechanisms; the latter requires reflexive self-consciousness—awareness of oneself as a participant in normative practices. LLMs exhibit remarkable norm-sensitivity while entirely lacking norm-responsibility—a gap that explains their distinctive status as genuine participants without full agency.

5.3 How LLMs Achieve Norm-Sensitivity

§4 explained the technical mechanisms producing LLM behavior. We can now connect those mechanisms to norm-sensitivity, showing how statistical learning enables genuine tracking of linguistic norms without reflexive understanding.

Norm-sensitivity emerges directly from the RLHF training process. Recall from §4.2 that during RLHF, human evaluators consistently prefer certain types of outputs over others: acknowledgments of uncertainty, admissions of error, polite refusals, relevant responses. The model learns these preferences as statistical patterns, assigning high probability to norm-conforming tokens. The model produces outputs conforming to norms because such outputs were consistently reinforced during training.

Consider what happens when a user corrects an LLM error. The user states: 'Actually, the capital of Australia is Canberra, not Sydney.' The LLM responds: 'You're absolutely right, and I apologize for the error. Canberra is indeed the capital of Australia.'

Mechanically, the correction becomes part of the conversation context. The model's probability distribution over next tokens shifts dramatically: tokens consistent with

acknowledgment and correction receive higher probability. This produces outputs that appear to acknowledge error and adjust accordingly.

But this is norm-sensitivity, not norm-responsibility. The model tracks the norm 'acknowledge and correct errors when challenged' through learned statistical associations, not through reflexive recognition that it made a mistake. There is adjustment without ownership, correction without understanding. Crucially, if the conversation resets and the same question is asked, the model might make the same 'mistake' again, revealing that no genuine revision occurred. The correction was local to that conversation context—a statistical recalibration within that particular context window—not a change in understanding or commitment to avoiding the error in future.

Beyond error correction, LLMs adjust to different registers, recognize when clarification is needed, respond appropriately to various speech acts, and modify outputs based on user feedback. This enables the functional coordination identified in §5.1. These mechanisms suffice for norm-sensitivity—tracking norms and adjusting behavior accordingly—but cannot produce norm-responsibility, as we now examine.

5.4 Why LLMs Lack Norm-Responsibility: The Reflexive Gap

Norm-responsibility requires several interrelated capacities that LLMs entirely lack. These capacities are deeply interrelated, all requiring reflexive self-awareness, but each highlights a distinct dimension of responsible agency:

- **Ownership:** To be norm-responsible, one must be capable of recognizing one's utterances as one's own commitments that place oneself under normative constraint. LLMs lack this entirely; they generate statistically probable tokens without understanding that producing them generates obligations. The utterance is *commitment-apt*—capable of functioning as a commitment in discourse—but not a commitment in the full sense, lacking the reflexive dimension of self-imposed obligation. There is no 'I' that recognizes itself as making a claim for which it will be responsible.
- **Reflexive acknowledgment:** A norm-responsible agent, when corrected, recognizes that they've committed themselves to claims incompatible with the evidence and acknowledges the error. The LLM's adjustment is mechanical—a learned response

pattern—not an act of reflexive recognition. There is no moment of understanding: 'I committed to X, X is incompatible with the evidence, therefore I made an error.'

- **Justification capacity:** Norm-responsibility involves being able to provide reasons when challenged while understanding the dialectical role those reasons play. LLMs generate justificatory tokens as statistically probable continuations in explanation-type contexts without understanding the dialectical role of reasons or recognizing them as their own defense. This becomes evident when models produce contradictory justifications in different contexts without sensing tension.
- **Accountability:** Norm-responsibility involves bearing consequences (epistemic, social, moral) for one's utterances and being subject to reactive attitudes like trust, blame, and resentment that presuppose one is a responsible agent. LLMs cannot bear responsibility in this sense. Responsibility flows back to the human actors: developers, deployers, and users. The model itself isn't a locus of responsibility; it cannot be trusted or blamed.

What unifies these limitations is the reflexive gap—the absence of a self-conscious relation to one's own speech acts. Responsible agents grasp themselves as producers of those utterances, as participants who can succeed or fail at meeting standards. This reflexivity enables the could-have-done-otherwise condition: recognizing that one's assertion was a move one made, and that one might have made a different move. LLMs lack this reflexive dimension entirely. They generate statistically probable continuations without meta-level awareness of what they're doing or that they're doing anything at all. The architecture described in §4 cannot produce the kind of reflexive self-awareness required for norm-responsibility.

This is not an engineering deficiency that better training will overcome, but a conceptual distinction constitutive of the difference between tool and agent, between functional participant and responsible interlocutor.

5.5 Contemporary Challenges: Chain-of-Thought and Memory

A natural objection arises at this point. Might recent developments in chain-of-thought (CoT) prompting and persistent memory undermine the claim that LLMs lack reflexive awareness? When models explicitly 'reason through' problems before answering, or when they retrieve information across sessions, doesn't this begin to look like the reflexive self-relation I've denied?

Two clarifications matter. First, this analysis concerns current GPT-like architectures and training regimes, not in-principle claims about what artificial systems could never achieve. Second, and more importantly, the issue is not whether systems can output text about their outputs or retrieve earlier statements. The issue is whether they treat those outputs as normatively binding commitments that constrain later performances—commitments they own across contexts.

CoT does not close the reflexive gap. In standard CoT prompting, the model produces an extended token sequence: first intermediate text resembling deliberation, then a final answer conditioned on that text. This can *look* like reflection followed by choice. But internally, CoT remains context-extended next-token prediction. The model generates reasoning-like trajectories because such trajectories correlate with good answers in training data and receive preference optimization rewards.

For CoT to underwrite norm-responsibility, the system would need to: (i) identify prior reasoning as its own, (ii) treat it as a commitment constraining later performances, and (iii) take responsibility for reconciling tensions across contexts—by defending, revising, or retracting. Current systems don't exhibit this binding self-relation. The CoT text functions as additional context steering token probabilities, not as standing commitments the system treats as answerability-imposing.

A familiar phenomenon illustrates this: present a model with a subtle dilemma, and it may generate CoT concluding 'therefore action A is correct.' Reset the conversation or change framing, and it may generate different CoT concluding 'therefore not-A is correct'—without recognizing tension or feeling pressure to reconcile. When norm-responsible agents confront their previous reasoning, they face normative pressure to explain why earlier reasoning was mistaken, why cases differ, or to acknowledge changed minds. Present systems treat prior reasoning as optional textual artifacts rather than owned commitments requiring justification when violated.

Long-term memory raises similar issues. Tool-mediated memory lets systems retrieve earlier outputs across sessions. But retrieval is not ownership. Current 'memory' is typically an auxiliary store whose contents condition responses. The system doesn't acquire an endogenous practice of treating retrieved items as commitments requiring coherent integration. It may echo or repudiate prior statements based on what context makes locally probable, without the distinctive pressure to resolve inconsistency *as its inconsistency*.

The reflexive gap, then, is not about context window limits. It's a gap in commitment-keeping architecture: the lack of stable, binding mechanisms for owning and regulating commitments under norms of answerability. CoT and memory supply more text about reasoning and more access to past text, but that is compatible with absent norm-responsibility.

5.6 Participation Without Agency

This analysis reveals that LLMs occupy a distinctive position: they participate genuinely in linguistic practices through sophisticated norm-sensitive functional contribution, yet lack the reflexive agency characteristic of responsible speakers. They contribute to the space of reasons—generating utterances that license inferences, respond to challenges, and enable coordination—but don't occupy that space as responsible authors who own their commitments.

When users interact with LLMs, they engage in familiar processes of uptake, evaluation, and repair. These interactions instantiate the same normative structures governing ordinary conversation, even though one participant lacks agency in the full sense. The model's outputs play genuine inferential roles within discourse, enabling users to accomplish practical goals, explore ideas, and coordinate action.

Yet responsibility for these outputs flows back to human agents and institutions. The model's outputs are commitment-apt—they function as claims within discourse—but the commitments belong to the human actors. When LLMs produce harmful misinformation or consequential errors, responsibility lies with:

- **Developers:** who bear responsibility for training choices (dataset selection, objective functions, safety measures, known limitations, disclosure practices)
- **Deployers:** who bear responsibility for integration decisions (appropriate use contexts, guardrails, user guidance, risk mitigation)
- **Users:** who bear responsibility for reliance decisions (verification practices, delegation judgments, downstream effects)

The model itself cannot bear responsibility because it lacks the reflexive capacities that make responsibility intelligible: ownership of commitments, recognition of norms as binding, capacity for genuine cross-context self-correction.

This has practical implications. High-stakes contexts requiring accountability should not treat LLM outputs as advice from responsible advisors but as suggestions requiring human

verification. Liability frameworks should reflect that LLMs are tools deployed by responsible agents, not autonomous agents who can themselves be liable. When relying on LLM outputs, users should adopt epistemic stances appropriate to consulting sophisticated but non-responsible tools rather than knowledgeable colleagues.

6. Conclusion

This paper has demonstrated the framework-relativity of the LLM debate through an irenic-pragmatist methodology. The Communicative Intention Argument and No Meaning Charge embed contestable theoretical commitments—treating particular requirements as definitional rather than framework-specific elaborations. One can remain at a minimal pragmatist-functional level—on which coordination through norm-sensitive uptake suffices—without endorsing those elaborations.

By extending Price's i-representation/e-representation framework, I argue that LLMs exhibit i-representational competence through norm-sensitive participation in inferentially structured discourse. Whether they exhibit e-representational competence remains contested and framework-relative, depending on whether agent-mediated uptake suffices or whether further grounding relations are required. For understanding how LLMs functionally participate—enabling coordination, structuring inquiry, and responding to norms—an irenic pragmatist approach provides sufficient explanatory traction without requiring resolution of contested metaphysical disputes.

Technical analysis confirmed that LLMs demonstrate sophisticated norm-sensitivity through learned statistical patterns, enabling genuine functional contribution. What they lack is norm-responsibility—the reflexive capacity to own commitments and bear accountability. This reflexive gap marks a conceptual distinction between tool and agent, not merely an engineering limitation.

The question 'Do LLMs really speak?' admits a nuanced answer: they participate in language through norm-sensitive contributions to coordination and inference, but not as agents who own commitments and bear responsibility. Their participation is real but limited—a functional contribution to inference-governed coordination without occupying the accountable role characteristic of responsible speakers. This reveals linguistic participation as multidimensional rather than all-or-nothing: different systems occupy different positions along

dimensions of i-representational competence, functional coordination, norm-sensitivity, and norm-responsibility. From an irenic pragmatist perspective, this is reconciliation, not relativism: frameworks disagree about further grounding conditions while converging on the functional facts that matter for many practical purposes.

Most importantly, this analysis reorients debates from abstract questions about the essence of linguistic competence to concrete questions about design, deployment, and regulation. How should we design systems to maximize reliable norm-sensitivity while acknowledging the absence of responsibility? In what contexts is norm-sensitive participation sufficient, and when do we genuinely need norm-responsible agents? How should liability frameworks allocate responsibility when sophisticated tools contribute to consequential decisions? These questions cannot wait for philosophical consensus about grounding conditions or the metaphysics of content. The irenic pragmatist approach shows how to make progress on urgent practical questions while metaphysical disputes remain unsettled—neither dismissing those disputes as meaningless nor allowing them to obstruct practical judgment.

References

Andrews, Kristin. 2020. *The animal mind: An introduction to the philosophy of animal cognition.* Routledge.

Attah, Nuhu Osman. 2024. *Talking Machines: Philosophical Essays on Language Models*. PhD dissertation, University of Pittsburgh.

Attah, Nuhu Osman. 2025. 'Do language models lack communicative intentions?' *Synthese* 205 (187): 1–23.

Bender, Emily M., and Alexander Koller. 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. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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. New York: ACM.

Brandom, Robert B. 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment.* Cambridge, MA: Harvard University Press.

Dretske, Fred. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Field, Hartry. 1994. 'Disquotational Truth and Factually Defective Discourse.' *The Philosophical Review* 103 (3): 405–452.

Fine, Arthur. 1986. *The Shaky Game: Einstein, Realism, and the Quantum Theory*. Chicago: University of Chicago Press.

Fodor, Jerry A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Grice, Herbert Paul. 1957. 'Meaning.' *The Philosophical Review* 66 (3): 377–388.

Grice, Herbert Paul. 1969. 'Utterer's Meaning and Intentions.' *The Philosophical Review* 78 (2): 147–177.

Grindrod, Jumbly. 2024. 'Large Language Models and Linguistic Intentionality.' *Synthese* 202 (5): 1–25. https://doi.org/10.1007/s11229-024-04723-8

Hattiangadi, Anandi, and Anders J. Schoubye. 2025. 'The Outputs of Large Language Models Are Meaningless.' *Philosophical Studies*. Forthcoming.

Lohmann, Heike, and Michael Tomasello. 2003. 'The role of language in the development of false belief understanding: A training study.' *Child Development* 74 (4): 1030–1043.

Lappin, Shalom. 2024. '*Assessing the Strengths and Weaknesses of Large Language Models.*' *Journal of Logic, Language and Information* 33 (1): 9–20. https://doi.org/10.1007/s10849-023-09409-x

Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press.

Millikan, Ruth Garrett. 1995. 'Biosemantics.' In *White Queen Psychology and Other Essays for Alice*. The MIT Press.

Papineau, David. 2022. 'Swampman, teleosemantics and kind essences.' *Synthese* 200 (6): 509. https://doi.org/10.1007/s11229-022-03966-7

Price, Huw. 2011. *Naturalism Without Mirrors.* Cambridge: Cambridge University Press.

Price, Huw. 2013. *Expressivism, Pragmatism and Representationalism*. Cambridge: Cambridge University Press.

Shanahan, Murray. 2022. 'Talking About Large Language Models'. *Communications of the ACM* 66 (1): 68–77. https://doi.org/10.1145/3571730

van Dijk, Bram, Tom Kouwenhoven, Marco Spruit, and Max Johannes van Duijn. 2023. '*Large Language Models: The Need for Nuance in Current Debates and a Pragmatic Perspective on Understanding.*' *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8792–8806. Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.779

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. 'Attention Is All You Need.' *Advances in Neural Information Processing Systems 30 (NIPS 2017)*: 5998–6008.