Beyond Binary Understanding: LLMs as Catalysts for Philosophical Recalibration

Micah Probst

Abstract

The notion of ”understanding” faces a profound recalibration challenge with the emergence of Large Language Models (LLMs). Often dismissed in popular discourse as mere pattern-matching machines—”stochastic parrots” or ”semantic zombies”—these systems demand deeper philosophical examination as they demonstrate increasingly sophisticated linguistic capabilities. This paper challenges traditional binary conceptions of understanding by examining LLMs through four influential philosophical frameworks: Turing's behaviorism, Searle's biological naturalism, Grice's communicative intentions, and Wittgenstein's language games. Drawing on recent mechanistic interpretability research, I demonstrate how LLMs possess neither mere statistical mimicry nor human-equivalent comprehension, but rather exhibit mechanically different yet analogous forms of semantic understanding and intentionality when viewed along multiple dimensions. The evidence suggests our philosophical frameworks require fundamental recalibration to accommodate non-anthropocentric cognitive architectures. While LLMs currently lack generalized intelligence and genuine agency, their unique capabilities as linguistic agents compel us to develop more nuanced theories of understanding—theories that recognize the possibility of multiple valid forms of semantic competence beyond the human paradigm. This philosophical reframing is essential as we navigate a future increasingly populated by sophisticated artificial linguistic agents.

I. Introduction

Recent developments in generative AI—for this paper I will be concerned with Large Language Models (LLMs)—have introduced a new set of questions to consider in our attempts to define what it means for something to have “understanding.” Commonly considered as mere pattern-matchers, LLMs have been denigrated to the positions of “stochastic parrot” [1] or “semantic zombie” in the eyes of some researchers and common folk alike. However, as these models demonstrate increasingly sophisticated linguistic capabilities, such characterizations demand deeper examination

and consideration. This paper is intended to explore a central question that constitutes what we call LLMs: What is the nature of LLMs as linguistic agents?

This broad inquiry encompasses several importantly interconnected sub-questions: Do LLMs have semantic understanding? Do LLMs have conceptual understanding of universals? What kind of linguistic—or other kinds of—behavior do LLMs possess? Are LLMs legitimate participants in language games? And what significance does it carry that LLMs' language processing mechanisms differ from our own? The answers we give to these questions carry significant implications across multiple domains. Philosophically, if LLMs genuinely understand language, our theories of language and mind may require substantial revision. From a cognitive science perspective, LLMs or other models that achieve genuine understanding would drive us to reshape our existing view and make new hypotheses about the human brain's mechanisms for understanding. Ethically, entities with real understanding may demand different considerations than mere tools. Socially, how we interact with and deploy these systems crucially depends on what their status as linguistic agents is.

Given the broad and contested nature of our definitions of "understanding," I will here clarify that in this paper I seek to examine LLMs for semantic understanding—knowledge of word meanings and representing associated concepts—and conceptual understanding—representing and applying meta-linguistic universal concepts. Furthermore, rather than treating understanding as a binary quality, I adopt the view suggested by Lyre [2] that understanding exists on a spectrum with multiple dimensions of grounding[1]. This approach helps cut the proverbial fat of understanding by making it more quantifiable as a sum of specific abilities. Additionally, it leaves room for a non-anthropocentric view of understanding which can be applied to humans and LLMs alike.[2]

The paper proceeds as follows: First, I examine four pertinent theoretical frameworks from machine learning and philosophy of language—Turing's behavioral approach [3], Searle's Chinese

[1]"Grounding" refers to how abstract concepts connect to physical reality, experiences, or other established meanings. Human understanding is typically grounded through direct sensory experience (seeing actual dogs), social interaction (learning from others about dogs), or connections to other already-grounded concepts.

[2]The spectrum approach to understanding requires specific criteria for what constitutes advancement along different dimensions. These dimensions might include: (1) causal grounding in the world, from none to direct sensorimotor experience; (2) capacity for abstraction, from surface pattern matching to hierarchical concept formation; (3) contextual flexibility, from rigid application to adaptive context-sensitivity; (4) metacognitive awareness, from none to sophisticated self-monitoring; and (5) integration across domains, from isolated competencies to unified understanding. These dimensions must be specified with measurable criteria to avoid the spectrum becoming merely a rhetorical device that sidesteps binary questions through continuous terminology. Different cognitive architectures likely excel at different dimensions rather than simply advancing linearly along a single spectrum.

Room Arguments [4], Grice's theory of communicative intention [5], and Wittgenstein's language games [6]. Each section is meant as a presentation of the author's original position and a reconstruction of their arguments and assumptions to lay the groundwork for later critique and discussion. Next, I provide a high-level overview of the technical background required to understand transformer-based LLMs, explaining key components such as attention heads, multi-layer perceptrons, and the residual stream. The heart of this paper follows in four "beyond" sections, where I systematically evaluate how recent mechanistic interpretability findings from Anthropic [7]–[12] and philosophical analysis from Boisseau [13], Attah [14], and Lyre [2] challenge or support the aforementioned classical frameworks. The paper concludes with a discussion of the findings in aggregate and considerations of future directions and implications for AI development. The arguments made throughout this paper lead me to support the stance that LLMs have strong mechanistically different, yet analogous kinds of semantic understanding, conceptual understanding, and intentionality when viewed as degrees, but do not currently possess generalized intelligence or agency.[3]

II. Classic Frameworks

For fields as complex and multidisciplinary as AI and machine learning, the clearest pictures are the product of more varied analysis. This motivation is the reasoning behind the order in which I develop the ideas in this paper. While, in the end, the answers I seek may reside in the technical results and literature, supplementing and framing the technical discussion with theoretical frameworks allows us to gain new insights and deeper comprehension. Hence, we begin with the two most prominent and withstanding theoretical frameworks in machine learning and follow them with what I have identified as the two most challenging theories of language for LLMs to contend with.

[3]Agency refers to the capacity to act independently, make choices, and pursue goals—fundamentally different from understanding, which involves grasping meaning and concepts. This distinction explains why an entity might understand language without having meaningful agency or have agency without understanding. Full artificial general intelligence would require both capacities, but they can develop independently and through different mechanisms.

A. Turing

Alan Turing's seminal 1950 paper "Computing Machinery and Intelligence" sought a task very similar to the present paper, to take a vague question, "Can machines think?" and offer a reframing that could have a measurable answer. The product of this reframing, the Imitation Game, is now known as the Turing test and offers a behavioral approach to the question of machine intelligence [3]. The test took a human judge and put them in an experiment setting with two hidden participants—a human and a computer. If after a brief text-based conversation the judge cannot ascertain with a specific degree of accuracy[4] which participant is which, the machine passes the test and earns the badge of intelligence. This behavioral approach shifted away from definitional problems regarding "thinking" toward observable capabilities, establishing an influential paradigm in the evaluation of AI models.[5] Turing's reframing of the initial question in this way is on account of his belief that "thinking" was traditionally defined too ambiguously for empirical investigation and was therefore meaningless to discuss [3].

The remarkable mind he was, Turing seemed to anticipate a great deal of objections to his argument as well as trends in the future of the field of machine learning. Of the nine objections in the original paper, I want to highlight two:

- The "Argument from Consciousness" claimed that machines cannot have consciousness or feelings by virtue of their nature as artificial [3]. This seemingly innate intuition has persisted in the 75 years since it was originally formalized by Turing and his response to the objection endures just as well. Turing observed that the "Argument from Consciousness" leads to a "problem of other minds" and would necessarily result in solipsism if applied consistently.
- "Lady Lovelace's Objection" argues that computers are only capable of performing what they are programmed to do [3]. Turing's response to this set the conceptual groundwork for the field of machine learning. He argued that a machine's behavior could surprise its programmer through learning rather than explicit programming. Many decades later, this has become the paradigm of AI models for which we now experience a kind of "black box"

[4]I have omitted the exact duration and accuracy numbers on account that they have been changed throughout the years and do not meaningfully contribute to the theoretical analysis at hand.

[5]This trend manifests currently in the endless array of benchmarks new models are scored upon.

problem of behaviors we cannot easily interpret or predict.

Further anticipating the field of machine learning, Turing posited the idea of a "child machine" [3]. This approach—rather than attempting to produce a program that simulates an adult mind, produces one which simulates a child's mind and its ability to learn, resulting in an adult mind when appropriately trained—has become the foundation of neural networks and their learning algorithms.

As an optimist about the idea of machine intelligence and a critic of granting humans exceptional status when it comes to intelligence and consciousness, Turing not only anticipates many current research directions, but laid the groundwork for a functionalist view that intelligence could be meaningfully defined through behavior rather than internal mechanisms. However, as we will discuss later, concerns about safety have led to a practical and conceptual return to attempts of understanding the internal mechanisms of AI models.

B. Searle

In his influential 1980 paper "Minds Brains and Programs," John Searle introduced the Chinese Room Argument (CRA) to challenge what he termed "strong AI"—the claim that appropriately programmed computers possess genuine cognitive states and understand language [4]. The CRA imagines a man locked in a room where he is fed pages with Chinese characters he does not know (inputs) and then uses a rulebook (program) to match the characters to new characters which he offers to researchers outside the room (outputs). When interacting from the outside, a Chinese speaker has the experience of corresponding with the room in a way that is indistinguishable from how a native speaker would answer. Yet, the man in the room is merely following symbol manipulations and does not understand any Chinese.

Searle's CRA can be seen as an objection to Turing's behavioral approach. The goal is to show that even if a program produces outputs indistinguishable from those of a native Chinese speaker, the program itself does not understand Chinese. Furthermore, Searle shifts the goalpost from "understanding" to "intentionality" as the meaningful quality that precludes the existence of strong AI. The story develops as follows:

1) Syntax and semantics are fundamentally different and computers can only contain syntax.

2) Understanding requires grasping semantic content.
3) Intentionality[6]—the capacity for mental states to be about or directed at objects and states of affairs—requires understanding [4].
4) Genuine cognitive states such as beliefs, thoughts, and desires necessarily have intentional content.
5) Therefore, the claim of strong AI is false.

If semantics, which functions as the base for the rest of the developed powers, is not a product of computation, where does it plant its feet? Searle offers the answer of biological causal powers.[7] These causal powers, he argues, emerge from the biochemistry of the brain and cannot be duplicated by formal programs run on other physical substrates. He likens intentionality to being a biological phenomenon as causally dependent on its originating biochemistry as lactation or photosynthesis [4].

Searle can be seen to represent a more pessimistic outlook on machine learning that denigrates them to tools of symbol manipulation. His CRA works as a rejection of behavioral analysis of intelligence and relocates the locus of understanding in a kind of biological naturalism.

C. Grice

A precursor to Searle, H.P. Grice put forth a then novel account of meaningful communication in his 1957 paper “Meaning” [5]. Breaking from the history of anchoring meaning to convention, reference, or truth conditions, Grice—as Searle later does—places meaning in intentions and cognitive states. An important consequence of this definition is that every agent wishing to participate in meaningful communication must possess intentions and cognitive states.

The formal construction of Grice’s theory relies on a distinction between what he names natural and non-natural meaning. Natural meaning occurs when “x means y” entails the factual truth of

[6]Searle’s argument hinges on a crucial distinction between ”intrinsic” intentionality (mental states that are inherently about something) and ”as-if” or ”derived” intentionality (states that appear intentional only from an observer’s perspective). For Searle, computers—including LLMs—can only have derived intentionality attributed to them by human interpreters, while brains have intrinsic intentionality. This distinction challenges attributions of genuine understanding to LLMs even when they exhibit behavior functionally indistinguishable from human understanding. The circularity concern here is that evidence for intrinsic intentionality may ultimately rely on external behavior.

[7]Also known as biological naturalism. This idea posits consciousness and intentionality as biological phenomena emerging from brain processes. Unlike both dualism and computational theories of mind, it maintains that mental states are both causally reducible to neurobiological processes and ontologically irreducible as first-person experiences. This position allows Searle to acknowledge the physical basis of mind while rejecting functionalism’s claim that mental states are merely functional roles.

y—for example, “Those spots mean measles” indicates that measles are necessarily present [5]. In contrast, non-natural meaning does not entail such factual truth—for example, “Those three rings on the bell mean the bus is full” could be false if the bell is rung in an inappropriate circumstance [5]. Grice uses this separation of causal indication and intentional significance to develop his analysis of speaker meaning. He offers the following formulation to answer the question of what it means for a speaker S to mean something by an utterance x: “S intended the utterance of x to produce some effect in an audience by means of the recognition of this intention” [5].

This analysis identifies three nested levels of intention crucial to communicative meaning:

1) The intention to produce a certain response in an audience.
2) The intention that the audience recognize the speaker’s intention.
3) The intention that this recognition play a causal role in producing the intended response.

What distinguishes the communicative intention theory from mere manipulation is its self-referential structure. The speaker must not only intend to produce an effect, but that the effect is produced by the audience’s recognition of that intention. Grice’s grounding of meaning in the communicative intentions and recognitions of both interlocutors sets the stage for many interesting challenges for LLMs to confront.

D. Wittgenstein

Ludwig Wittgenstein’s later philosophy, namely his famous “Philosophical Investigations” (1953), represents one of the most radical theories of language and meaning in the tradition [6]. In the investigations, he rejects both his own earlier picture theory of language from the Tractatus and other referential theories, Wittgenstein introduces “language games”[8] as a novel framework for understanding linguistic meaning. The language game approach suggests that words derive their meaning from their use within diverse rule-governed social practices. He illustrates this through numerous examples: giving orders, describing objects, reporting events, forming hypotheses, telling stories, joking, greeting, praying—each representing a different “game” with its own implicit rules [6].

[8]“Sprachspiele” in the original German.

Wittgenstein demonstrates how meaning emerges through practical activity rather than mental representation through cases like his builder example (§2) where one worker calls out words like "slab" or "block" and another responds with appropriate actions involving the objects [6]. The plurality of language games replaces the search for a unified essence of language while paying attention to the "family resemblances" between diverse linguistic practices.

The agent who participates in language games must be trained rather than merely instructed—understanding, for Wittgenstein, is not about grasping an interpretation but about exhibiting practical mastery in actions. This grounding in actions implies that participating agents ought to be embodied and social beings by nature who engage in shared practices. Furthermore, Wittgenstein argues that no rule can determine its own application and that any formulation can be interpreted in many ways.[9] Rules become fixed by the agreement conferred upon them through practice within a community. This eliminates the validity of private languages having meaning on the grounds that an agent with a genuinely private language would have no stable standard to distinguish between following a rule correctly and merely thinking they were doing so [6].

This view of language and meaning presents the most significant challenge for LLMs. Wittgenstein's emphasis on embodied practice and communal agreement suggests that language use is inextricably linked to human[10] ways of being in the world. Unlike Turing's behavioral approach, which focuses on the outputs alone, Wittgenstein suggests that understanding language requires participation in shared forms of life—raising questions about whether a disembodied system can participate in language games. Furthermore, unlike Searle's focus on internal mechanisms, Wittgenstein locates meaning not exclusively in the mind of the agent but in the normative practices of a linguistic community. This framework presents a unique perspective to evaluate whether LLMs are agents of simulacra or participation in language games.

[9]Wittgenstein's rule-following paradox presents a particularly challenging objection to LLMs as genuine language users. The paradox demonstrates that any finite set of examples is compatible with infinitely many different rules, and no rule can determine its own application. Wittgenstein resolves this through appeal to normative practices within a community, where correct rule-following is determined by agreement in judgment, which presents an immediate challenge for LLMs.

[10]Or other animals capable of sufficient cognitive abilities to construct language games.

III. Technical Reviews

A. Transformer Architecture

Now that we have established the theoretical frameworks we want to evaluate LLMs under, we turn to the technical foundations of transformers. Transformers are a kind of neural network architecture—which LLMs are a product of—derive their name from their mechanism. That is, at a very high level, they perform a series of transformations on vectors—which are how information is embedded into the system—which create an output vector that is turned into a probability distribution of the proper next token.[11] Modern LLMs originate from the famous 2017 paper "Attention Is All You Need" [15]. The revolutionary idea presented in the paper that set forth a new AI spring was the ability to process data in parallel. This had the effect of eliminating computational bottlenecks and opening the possibility of more efficient scaling to larger and larger datasets.

[11]Tokens are the basic units that LLMs process, but they are not exactly the same as words. A token might be a common word ("the"), a part of a longer word ("ing" in "running"), a punctuation mark, or even a space character. Most English words become 1-2 tokens, but uncommon words might be broken into many tokens. This tokenization process affects how models understand text—for example, "peanutbutter" (uncommon, multiple tokens) might be processed differently than "peanut butter" (common, fewer tokens)—creating subtle effects on model comprehension.

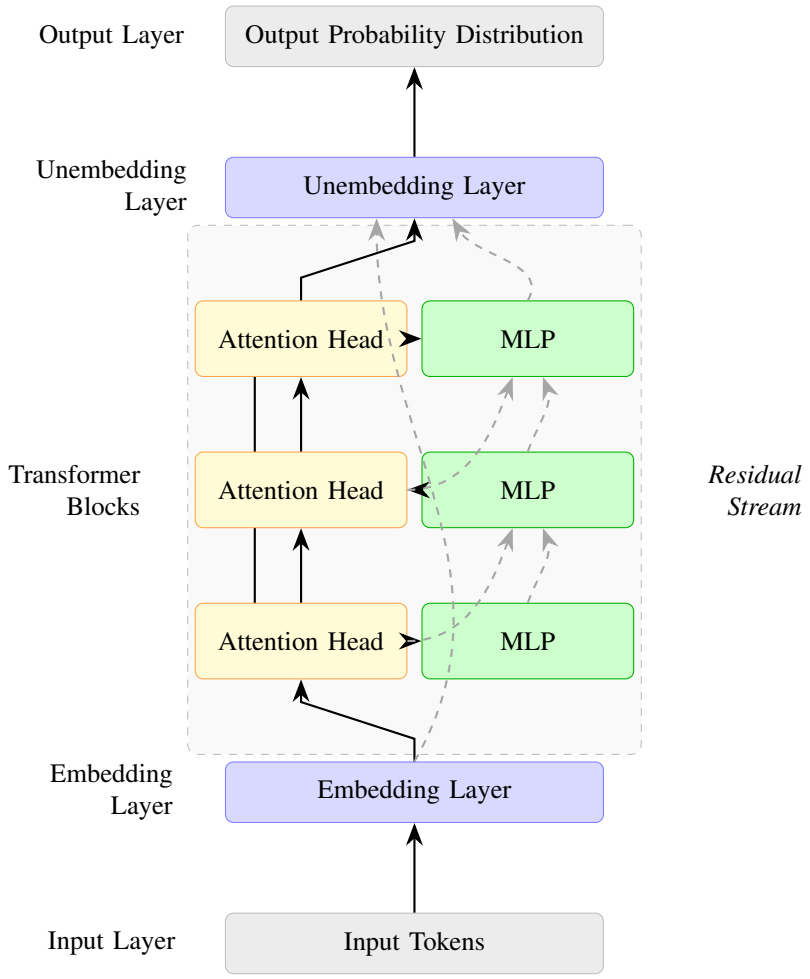

Fig. 1. The transformer architecture showing the flow of information through the model. Input tokens are embedded into vectors, processed through multiple layers of attention mechanisms and MLPs within the residual stream, and finally unembedded to produce output probabilities. The dashed lines represent residual connections that allow information to bypass layers.

The basic transformer architecture consists of an embedding layer which turns the input into token vectors[12], multiple layers of residual blocks which perform the transformations, and an unembedding layer that produces output probabilities. Each residual block contains an attention layer—a group of parallel attention heads— followed by a multi-layer perceptron (MLP) layer [7]. The path through the entire model is called the residual stream and serves as the communication channel that all of the other layers operate on.[13] The linear nature of this stream allows different

[12]These embeddings represent tokens in a high dimensional space where semantic relationships are preserved—similar words cluster together. During training, the model learns these vectors for each token in its vocabulary (usually 30,000-100,000 tokens), effectively creating a mathematical representation of language building blocks.

[13]The residual stream solves the deep learning problem known as the vanishing gradient problem. Notated as: LayerNorm(input + Sublayer(input)), this connection creates a direct path for information and gradient flow during training. Solving the vanishing gradient problem allows unconstrained scaling with respect to information degradation.

layers to send information to specific subspaces, creating a complex distributed representation system. Now let us dissect each component's function and mechanism.[14]

Let us first look at the attention layers. Before moving into how the mechanisms work, I want to establish the function so there is a context to understand the mechanisms under. An input into an LLM is often hundreds to thousands of tokens. These tokens, which for simplicity's sake we can consider as single words, inevitably refer to each other. A system that only worked by sequentially moving through the information in a forwards fashion would get hopelessly confused on the context of the message. So, models need a mechanism to make all of these tokens talk to each other and share their context. This is what attention does. An attention layer consists of many parallel attention heads. The motivation behind this is that each attention head effectively asks and answers different questions about the tokens[15]—for example, is this token a noun and, if so, what adjectives attend to it? Models learn to, speaking at a high level, have each attention head ask and answer different questions about the input. Each attention head reads information from the residual stream at one token position and writes it to another, with the model learning which information to move and where to move it [7]. This enables the model to determine long-ranged dependencies in the context. As models scale, more sophisticated attention patterns emerge such as "induction heads," but those will only be relevant later [8].

MLPs can be thought of as where the model stores its "knowledge." These layers read the information off the residual stream as it's been added onto by attention layers, and essentially check if they know something about the information. When an MLP reads the information of the residual stream certain concepts it has learned will be activated, then these concepts are added into the residual stream to pass through more attention layers and MLPs. A common way to think about the knowledge stored in MLPs is that every learned concept is attributed to a specific direction in the high-dimensional space that is in the MLP, known as the linear representation hypothesis (LRH). However, models have learned to store more concepts than they have directions in their

[14]The embedding and unembedding layers do not contain any significant function other than converting information into something the model can understand and back to something humans can understand respectively, so I will not cover their mechanisms. For now, just know they bookend the more interesting transformation layers.

[15]For each token, the model computes query (Q), key (K), and value (V) vectors. The attention patterns—determining how much each token influences others—are calculated as: softmax$(QK^{\mathrm{T}}/\sqrt{d})$. These attention patterns are then run through the output-value circuit to produce the layer output.

MLPs. This is believed to be achieved by the "superposition hypothesis" [9]. The superposition hypothesis suggests that models use almost-orthogonal directions to represent more features[16] than there are dimensions, achieving remarkable efficiency through sparse encoding[17] of concepts [9]. At risk of anthropomorphizing, I would give the conceptual analogy of attention layers being equivalent to figuring out "what" someone is saying and MLPs being equivalent to figuring out what they "mean" by what they're saying.

IV. Beyond Turing

While the Turing Test initially provided a compelling operational framework for evaluating machine intelligence, the field has largely abandoned it as a meaningful benchmark. Modern state-of-the-art LLMs could all be reasonably expected to pass the test if given the correct system prompt, yet few researchers would claim this demonstrates human-equivalent understanding. Nevertheless, the concept of imitation central to Turing's approach warrants deeper examination.

This is exactly what Éloïse Boisseau investigates in her 2024 paper "Imitation and Large Language Models" [13]. Boisseau distinguishes between "imitative behavior" and "status of imitation," where imitative behavior requires an agent with its own independent behavioral repertoire modifies it to match another agent's behavior and the status of imitation refers to outputs that resemble those of another agent [13]. Her paper argues that LLMs engage in neither imitative behavior—on the grounds that they have no independent behavior to modify—nor are they themselves imitations of human speakers. Instead, she proposes that LLMs are better understood as "imitation manufacturing tool"—devices that produce outputs having the status of imitations of human speech [13].

I agree with the idea that they produce outputs having the status of imitations of human speech; however, I argue that they do have their own behavior. Yet, the manufactured imitations are merely a product of this behavior, leaving LLMs as machines whose behavior has the adaptive potential to create manufactured imitations. While this may seem like a particularly anal distinction to make, I believe this opens a good opportunity to consider the kind of behavior an LLM may

[16]The term for a concept represented by the sparse activations of the superposition hypothesis.

[17]This resembles the idea of compressed sensing in signal processing—recovering high-dimensional sparse signals from lower-dimensional measurements.

have, albeit limited. I contend that LLMs—and other AI models that work with data—exhibit a distinct loss-minimization and reward-maximizing behavior.[18] Furthermore, they appear to use surprisingly complex reasoning techniques to achieve these goals which suggest a kind of problem solving nature in pursuit of a goal.

The loss-minimization function of AI models is of less interest here since it is a static and consistent factor deep within the algorithms as they perform gradient descent during pre-training. On the other hand, the recent integration of reinforcement learning from human feedback (RLHF) and other kinds of RL offer particularly compelling evidence for more genuine behavior. Anthropic's research demonstrates that models develop "hidden-goals"[19] features directly associated with their assistant persona [12]. Lindsey et al. discovered through attribution graphs[20] that these features activate specific behavioral patterns whenever the model is prompted for assistance, guiding the response to align with the reward signals it was trained on such as helpfulness and harmlessness. Conversely, when certain user inputs activate safety-relevant features, the model engages specific inhibitory circuits meant to discourage and prevent harmful outputs, another product of RLHF [12]. Particularly revealing is the discovery of "default" circuits that cause models to decline providing a response unless a sufficient number of known features are activated and inhibit the circuit. These default behaviors appear to have emerged as an efficiency behavior to effectively fulfill its RL training goals.

While not empirically verified, I am inclined to offer a charitable account of behavior to LLMs and other AI models that they have a kind of quasi-behavior with enough sophistication to warrant preliminary considerations. If we accept this charity, then it leaves LLMs behavioral status as systems with their own behavior whose outputs happen to resemble human language—not because they are imitating humans, but because their reward-maximization behavior has been shaped to

[18]During pre-training, gradients of the loss function are computed by an algorithm known as back propagation which uses the chain rule to propagate error signals throughout the network. Gradient descent is then applied to adjust these parameters in the direction of the steepest loss decrease. When applied iteratively the model "learns" the information by finding local-minima.

[19]The development of "hidden-goals" features emerges through complex reinforcement learning dynamics rather than explicit programming. These features represent learned abstractions over rewarded behaviors rather than goal representations in the human sense. Mechanistic interpretability research shows these features form through sensitivity to positive gradient updates during RLHF training, creating activation patterns that maximize expected reward.

[20]Attribution graphs represent a methodological advance beyond basic circuit analysis by revealing causal relationships between learned features rather than just individual neurons. These techniques combine sparse autoencoders with causal intervention methods to isolate feature-to-feature influences. These techniques currently capture only a subset of model computation, particularly missing dynamics from attention mechanisms which may be crucial for understanding higher-order reasoning.

generate such outputs.

V. Beyond Searle

Searle's CRA presents the most direct challenge to claims about LLM understanding through its assertion of an insurmountable gap between syntax and semantics. However, recent advances in mechanistic interpretability offer compelling evidence that this rigid divide may be insufficient for explaining the representational capacities of contemporary LLMs. To motivate my later claim that semantics is more likely than not a product of sufficiently sophisticated syntax, I will give you a brief history of mechanistic interpretability.

The field of mechanistic interpretability evolved as a response to neural networks becoming effective "black boxes." Its goal is to develop techniques that reveal the inner operations of AI models. Early work by Elhage et al. established a mathematical framework for analyzing transformer circuits[21], identifying how information flows through the residual stream and how attention heads move information between tokens [7]. This initial research also found the existence of what the authors called "induction heads"[22] [7]. Induction heads—a special kind of attention head tied to in-context learning—were later validated as having a likely causal link to a model's in-context learning abilities due to their matching emergence as a phase change[23] [8].

Subsequent research deepened our understanding of MLPs and how semantic knowledge may be encoded. Using Sparse Auto-Encoders (SAE),[24] researchers at Anthropic identified "monosemantic features" that represent specific concepts with relative precision [10]. Building on the discovery of features, researchers discovered that models represent semantic neighborhoods through geometric organizations of semantically related clusters [**?**]. For example, researchers found clusters focused on medical concepts that transitioned from "immunocompromised people" to "specific

[21]The term for a traceable information path through a model.

[22]Induction heads mathematically implement pattern completion by having query-key attention matrices learn to detect tokens that previously appeared in sequence. When token X appears after token A, attention weights W_QK(A)·W_V(X) form an approximation where subsequent occurrences of A strongly attend to previous A→X patterns, effectively creating a learned key-value lookup mechanism that predicts X given the context of A.

[23]The term for when a model's performance in a task sees a significant jump when crossing a threshold of parameter size.

[24]Sparse Auto-Encoders use a two-layer architecture where the first layer maps model activations to a higher-dimensional space via a learned linear transformation followed by a ReLU nonlinearity, creating "features." The second layer attempts to reconstruct the original activations through a linear transformation of these feature activations. This technique reveals features that would be invisible when only analyzing individual neurons.

diseases" to "immune response" related features in a semantically coherent topology [**?**]. These findings suggest at least a rudimentary form of semantic understanding.

Beyond findings on understanding, researchers also found kinds of reasoning patterns that could motivate a sense of intentionality. Building on the ideas of circuits—information paths through a model's raw components like neurons and attention heads, researchers developed "attribution graphs" which are effectively circuits of features [12]. This was done by making a "replacement model" characterized by features[25] where circuits could be traced to draw causal relationships between features. Attribution graphs revealed, among other things, that models seem to be capable of activating features off of description alone. The symptoms of "preeclampsia" were given as input—but not the word itself—revealing that the "preeclampsia" feature was activated [12]. This suggests that the model "represents it internally, apparently using similar internal machinery as if the word were spelled out explicitly" [12]. This seems to be a kind of mental state (activation of features) being directed at a state of affair (medical diagnosis), expressing intentionality to some degree under Searle's definition.

Most interestingly with respect to Searle's CRA, researchers found evidence that models likely hold language-agnostic representations for many concepts. It was found that when solving multilingual tasks, "the key semantic transformation occurs using the same important nodes[26] in every language, despite not sharing any tokens in the input" [12]. This suggests that LLMs develop abstract, language agnostic conceptual representations similar to Noah Chomsky's notion of I-language—a universal linguistic competence underlying surface variations [16]. Further motivation for this claim was found when researchers noticed that language specific features only engaged at the final output, while the core computations were performed by the language-agnostic features in the middle of the model. This suggests a separation between conceptual processing and linguistic expression similar to Chomsky's distinction between I-language (internal linguistic competence) and E-language (external manifestations) [16].

[25]This is done with the use of Cross-layer Transcoders (CLT) that reconstruct MLP outputs using sparse features. The impact of attention layers is notably absent in the replacement model.

[26]Groups of related features that activate on the same concept.

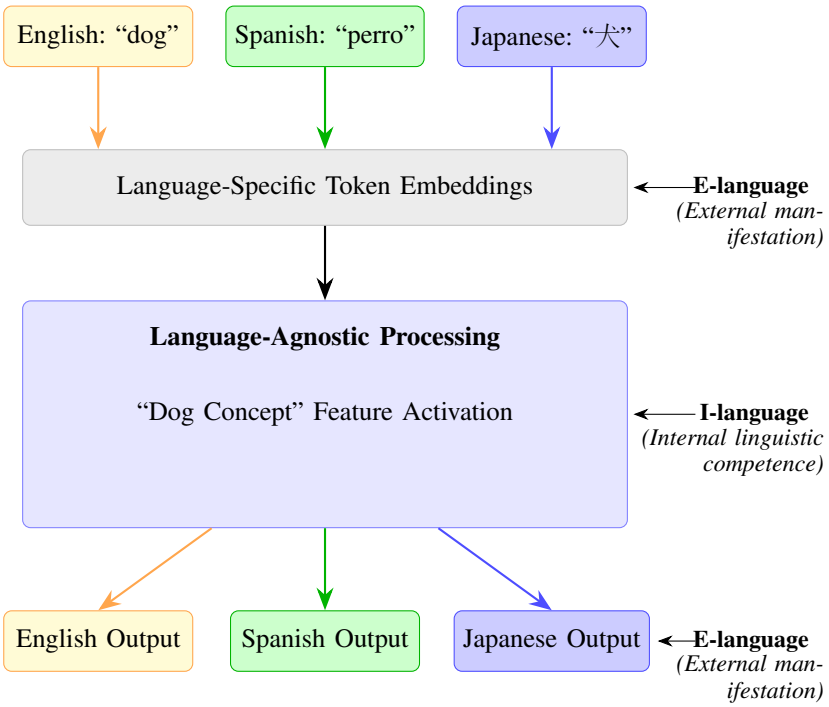

Fig. 2. The language-agnostic representation model in LLMs. Language-specific tokens for “dog” in multiple languages (English, Spanish, Japanese) are processed into a shared conceptual representation in the middle layers of the model. This representation is then translated back into language-specific outputs. This parallel with Chomsky’s distinction between I-language (internal linguistic competence) and E-language (external manifestation) suggests LLMs develop abstract conceptual understanding independent of specific languages.

In aggregate, the evidence suggests that semantics—as we commonly understand it—may be better understood as an emergent property of sufficiently complex syntax and structured architectures rather than a fundamentally different phenomenon. Furthermore, if we consider semantic understanding as a spectrum, LLMs demonstrate a meaningful degree of competence through their abilities to represent related concepts as neighbors, internally represent concepts without need for direct token referencing, and develop language-agnostic abstractions. These capabilities challenge Searle’s sharp distinction between syntax and semantics and suggest that his biological naturalism may grant unwarranted exceptionalism to human cognition.

VI. Beyond Grice

Grice’s theory of communicative intention presents a significant challenge for LLMs. The idea that meaningful communication requires a nesting of intentions—to produce an effect, to have this intention recognized, and to have this recognition play a causal role in producing the effect—

seems to demand cognitive capacities beyond what we may expect of LLMs. However, Nuhu Osman Attah offers a compelling reframing of this requirement that accommodates LLMs in his recent work on communicative intentions in LLMs [14].

Attah challenges what he calls the "Communicative Intention Argument" against LLM linguistic competence by identifying its two key premises: that genuine linguistic competence requires communicative intention, and that LLMs lack mechanisms for entertaining such intentions [14]. He argues this argument fails on two grounds: if we use a strong Gricean definition of communicative intention (requiring meta-representational abilities[27]), the first premise is empirically untenable as not all human communication requires such complex intentions. If we adopt a more minimal "control conception" of intention, then LLMs actually do have mechanisms that satisfy this definition.

This "control conception" defines intentions functionally as "internal states that both track and control a system's actions in a way that is responsive to its environment" [14]. Attah notes that this theory satisfies three important requirements: it's functional rather than phenomenological, consistent with cognitive science, and doesn't presuppose full agency [14]. When it comes to tracking belief-like states, LLM architecture excels. The final output of an LLM is chosen from a selection of options each weighted with a relative probability of relevance. At the beginning of the model, the number of plausible next tokens is huge and is narrowed down through increasing information into the residual stream through attention mechanisms and MLPs. These selections are, in effect, the model solving "Many-Many Problems"[28] by mapping inputs to appropriate responses. Under the control conception, LLMs mechanisms qualify as having intentions.

Anthropic's research on LLM internal representations provides empirical support for Attah's theoretical framework. Attribution graph analysis reveals that models contain features specifically dedicated to representing communicative goals and user intentions [12]. When processing queries, models activate features that represent not just the semantic content of the input but also the pragmatic aspects of the communication—for example, features that recognize potentially harmful

[27]Meta-representational abilities refer to the capacity to represent one's own and others' mental states—essentially "thinking about thinking." This includes understanding that others have beliefs different from one's own (theory of mind), representing one's own knowledge states (metacognition), and understanding representations as representations.

[28]The challenge of explaining how systems select appropriate actions when faced with many possible inputs and many possible outputs.

user intentions. These intention-recognition capacities enable the model to respond appropriately to a wide range of communicative situations.

However, LLMs do exhibit limitations in their intention-recognition capabilities compared to humans. They operate within a narrower range of communicative contexts—primarily human-assistant dialogues—and lack the sophisticated theory of mind that enables humans to recognize multiple, potentially conflicting intentions. As Attah acknowledges, "while denying LLMs possess the kinds of intentions humans have, [we can] also deny possessing those kinds of intentions is necessary for linguistic competence" [14]. This leaves open the possibility that communication might exist on a spectrum rather than as a binary, with LLMs earning a partial communicative competence.

VII. Beyond Wittgenstein

While Searle's CRA presented the most direct challenge to LLMs understanding, Wittgenstein's language games present the most difficult framework for LLMs to stand as human-equivalent language agents. His emphasis on embodied practice and communal agreement appears to disclude LLMs from genuine language use. However, applying the philosophical analyses by Boisseau [13], Lyre [2], and Lenci [17] offer potential avenues for partial participation for LLMs.

Boisseau's analysis of LLMs as "imitation manufacturing tools"[29] rather than agents engaged in imitative behavior addresses a key Wittgensteinian concern [13]. If meaning emerges from use in social practices, can a system that merely produces imitations of speech participate in language games? Based on Boisseau's analysis, I suggest that LLMs occupy a unique position: they don't imitate language in the way humans or animals might, but they produce outputs that have the status of imitation of human language. This creates an ambiguous relationship to language games—they produce outputs that function within some language games without necessarily participating in them the same way humans do. While only leaving a narrow opening, this may create the possibility of participation in a narrow range of language games.

Holger Lyre's concept of "indirect causal grounding" offers a potential bridge between disembodied LLMs and Wittgenstein's embodied language practices [2]. Lyre argues that even

[29]The status of whether they have their own behavior is irrelevant in this section so is not mentioned.

without direct sensory experience, LLMs develop "world models"—internal representations that are structurally isomorphic to aspects of the world—through their training on human-generated text. These representations combined with what Lyre calls "mild social grounding"—where models learn the rules and patterns of language uses through training on documented human practices—enable a form of partial participation in language games [2].

However, LLMs' disembodiment imposes significant limitations on their language game participation. Many language games presuppose embodied experience—perceptual discrimination, physical action, emotional responses—that remain inaccessible to text-only systems. While LLMs can simulate these experiences textually, this simulation lacks the necessary direct causal grounding. A striking example of LLMs deficiencies in conceptual representations as a product of their disembodied nature is found in Proietti and Lenci's study on the part-whole relation in LLMs [17]. Through behavioral, probabilistic, and representational tests, the paper found that LLMs had a deficient or incomplete grasp of antisymmetry and the part-whole relation [17]. That is, they struggled to understand that if x is part of y then y cannot be part of x. Their hypothesis for this struggle is that some concepts such as the part-whole relation require embodied capabilities to achieve a complete understanding of.

Overall, these findings suggest that language game participation may be open to LLMs in a very narrow set—primarily those focused on abstract reasoning, narrative, or formal patterns. The larger set of language games which require strong kinds of embodied or social abilities are inaccessible to LLM participation.

VIII. Discussion

Our examination of LLMs through the frameworks of Turing, Searle, Grice, and Wittgenstein has painted a nuanced picture of LLMs' linguistic capabilities and status as linguistic agents. Rather than fitting neatly into any single theoretical framework, LLMs demonstrate a pattern of partial but significant capacities across multiple dimensions of understanding. The evidence from interpretability suggests that LLMs possess their own distinct behavioral patterns (contra Turing's imitation framework), demonstrate sophisticated semantic representations (challenging Searle's syntax-semantics distinction), implement functional communicative intentions (reframing

Grice's intentionality requirements), while remaining limited participants in the full spectrum of Wittgensteinian language games.

This pattern aligns with my thesis that LLMs have mechanically different yet substantive kinds of semantic understanding and intentionality when viewed as existing on a spectrum, while lacking generalized intelligence and genuine agency. The key insight emerging from this analysis is that our philosophical frameworks for understanding and intentionality require refinement to accommodate entities whose cognitive architectures differ fundamentally from our own. Understanding may be better conceptualized as multidimensional rather than binary.

This perspective has significant implications for both AI development and philosophical inquiry. For AI research, it encourages the pursuit of architectures optimized for their unique computational substrate rather than trying to mimic human cognitive processes. For philosophy, it challenges the anthropocentrism implicit in many theories of language and mind, suggesting that understanding and intentionality may be realized through multiple architectural patterns and substrates rather than requiring specific biological implementations.

It is worth noting that humans themselves may not represent the ultimate endpoints of these multidimensional spectrums for understanding. Our own cognitive capacities almost certainly occupy intermediate positions, shaped by the specific evolutionary pressures and architectural constraints of biological neural networks. Future AI developments—particularly multimodal systems integrating visual, auditory, and other sensory modalities—may extend capabilities along certain dimensions beyond human capacities, while remaining constrained in others.

While transformers have proven a remarkably effective architecture, alternative approaches—including neurosymbolic systems that combine affective neural networks with symbolic reasoning capabilities—may enable new forms of abstraction, inference, and perhaps even forms of agency beyond transformers' capabilities. The development of such systems would not necessarily make them "more human-like" but might create entirely new cognitive profiles with unique combinations of strengths and limitations—pushing us to further expand our conceptual frameworks for understanding and intelligence.

In conclusion, LLMs present neither mere statistical mimicry nor human-equivalent under-

standing, but something still interesting—a novel form of semantic processing that shares family resemblance with human understanding while differing in fundamental ways. This recognition should challenge us to develop more nuanced, non-anthropocentric accounts of understanding and intentionality—accounts that recognize the possibility of multiple cognitive architectures with different but similarly valid forms of semantic competence. By embracing this broader perspective, we can better appreciate both the remarkable capabilities of current AI systems and the unique qualities of human cognition, without privileging either as the definitive form of understanding. This philosophical recalibration will be essential as we navigate a future with increasingly sophisticated artificial agents participating in our linguistic and cognitive ecosystems.

REFERENCES

[1] E. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610-623.

[2] H. Lyre, "Understanding AI: Semantic Grounding in Large Language Models," Feb. 2024.

[3] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433-460, 1950.

[4] J. R. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–457, 1980.

[5] H. P. Grice, "Meaning," *The Philosophical Review*, vol. 66, no. 3, pp. 377-388, 1957.

[6] L. Wittgenstein, *Philosophical Investigations*, Oxford: Basil Blackwell, 1953.

[7] N. Elhage et al., "A Mathematical Framework for Transformer Circuits," Anthropic, Dec. 2021.

[8] C. Olsson et al., "In-context Learning and Induction Heads," arXiv:2209.11895, Sep. 2022.

[9] N. Elhage et al., "Toy Models of Superposition," arXiv:2209.10652, Sep. 2022.

[10] Anthropic, "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning," Oct. 2023.

[11] E. Ameisen et al., "Circuit Tracing: Revealing Computational Graphs in Language Models," Anthropic, Mar. 2025.

[12] J. Lindsey et al., "On the Biology of a Large Language Model," Anthropic, Mar. 2025.

[13] É. Boisseau, "Imitation and Large Language Models," *Minds and Machines*, vol. 34, no. 42, pp. 1-24, 2024, doi: 10.1007/s11023-024-09698-6.

[14] N. O. Attah, "Do language models lack communicative intentions?," *Synthese*, vol. 205, no. 5, p. 187, Apr. 2025, doi: 10.1007/s11229-025-05022-6.

[15] A. Vaswani et al., "Attention Is All You Need," arXiv:1706.03762, Aug. 2023. doi: 10.48550/arXiv.1706.03762.

[16] N. Chomsky, *Knowledge of Language: Its Nature, Origin and Use*, New York: Praeger, 1986.

[17] M. Proietti and A. Lenci, "The quasi-semantic competence of LLMs: a case study on the part-whole relation," arXiv:2504.02395, Apr. 2025. doi: 10.48550/arXiv.2504.02395.