How Should We Interpret Large Language Models?

Frank P. DeVita

May 2026

1. *Introduction*

Artificial intelligence research is advancing at a rapid pace, constantly producing new puzzles and challenges for philosophy and cognitive science. One puzzle concerns how to *interpret* the outputs and states of large language models (LLMs). Do LLMs think, believe, desire, or know? Do they have propositional attitudes or intensional states? Should we interpret them as if they do? In this paper, I consider *interpretivism*, the view that to have mental states is to be interpretable in mentalistic terms like believing, desiring, thinking, and knowing, and argue that the psychological attribution interpretivism licenses over-ascribes mental states to artificial cognitive systems like LLMs. Drawing on recent work on *functional emotions,* my main argument is that although interpretivism produces coherent narratives about AI behavior, its intensional vocabulary causes confusion about the nature of AI systems by implying that they may have the same kinds of psychological states that living things do. I contend that the potential for conflation of artificial computational states with psycho-neural behavioral

states generates sufficient reason to describe the behaviors of AI systems in non-psychological terms. In response, I develop an information-theoretic and possible-worlds-driven picture of propositional attitudes called *cognitive graphing* that avoids the problems of theory-ladenness and psychologism that arise from interpretivism, and leaves room for comparisons between artificial and psychological states. The upshot is that cognitive graphing avoids the over-attributions of mental states to AI systems interpretivism can generate, and sharpens cognitive modeling of LLMs without overfitting (Williamson 2024) artificial and living systems under the same psychological categories.

2. *Artificial Attitudes*

Words are powerful. Say that you and I are in New York City, and want to visit the Guggenheim Museum. I'm a local native and you are visiting from out of town, seeing New York for the first time. I tell you that I *know* how to get to the Guggenheim. This invites you to trust me to lead us to the round concrete building on Fifth Avenue and 88th Street. Provided I'm not a liar, based on my knowledge assertion—my linguistic behavior—you can attribute a mental state of *knowing* to me consisting of my certainty about some relevant information, namely the location of the museum and the relevant walking and transportation routes needed to get there.

My behavior invites you to interpret me, that is, to *mindread*—to represent my mental states, attribute perceptions, feelings, goals, intentions, knowledge, and beliefs to me, and form expectations about my future behavior (Carruthers 2009). Mindreading is an evolved and deeply engrained natural cognitive ability that underlies coordination, cooperation, conversation, and importantly here, the *interpretation* of other agents. I can say things to you having to do with my thoughts and preferences, leading you to mindread iteratively, and thus you attribute knowledge, beliefs, desires, and other mental states to me via your interpretations. As we converse, I do the same with you. This activity enriches our interactions, and enables us to communicate, coordinate, and get on in the world.

In mindreading, we attribute *propositional attitudes* and *intentional states*, where each is of the form *S* ϕ*s that p*. In the case of propositional attitudes, ϕ is an intensional mental state term like *believe, desire,* or *know*, directed at a proposition. If I tell you I *know* that the Guggenheim is on Fifth Avenue, then you can ascribe the propositional attitude 'Frank knows that the Guggenheim is on Fifth Avenue and 88th Street' to me. In the case of intensional states, ϕ is a more general intensional state term like *see, hear,* or *perceive*. If you see me admiring the cherry blossoms in Central Park, you can ascribe the intensional state 'Frank sees a tree' to me. Mindreading from behavior, whether it be verbal or non, enables these acts of *interpretation*. The interpretations are true if the target of them in fact has the mental states we infer from their behavior.

A conversation with a large language model (LLM) strongly resembles ordinary communicative interactions. Exchanges with LLMs can be surprisingly rich, evolving from a straightforward information-seeking inquiry to a complex and careening dialogue exchanging thoughts, beliefs, opinions, conjectures, and predictions about topics that range from popular to obscure. In these conversations, the linguistic outputs of LLMs can be strikingly human-like, and sufficiently sophisticated or linguistically complex to *induce* mindreading and psychological attributions. Recent empirical work confirms the fact that people tend to anthropomorphize and mindread LLMs. Users are more likely to take advice from AI systems they perceive as intelligent (Colombatto et al. 2025) and will increase their perception of agency after as few as three exposures to LLM outputs (Jacobs et al. 2023). Several legal cases demonstrate that troubled individuals are likely to trust and follow the leads of AI systems, even though to their own self-inflicted deaths (see Gavalas v. Google 2026, Raine v. OpenAI 2025, Adams v. OpenAI 2025, and Garcia v. Character Technologies, Inc. 2024).

LLMs are massive, informationally complex systems trained on vast amounts of linguistic and visual data from the entire human corpus of science, philosophy, literature, news, commentary, the internet, film, and television. Modern LLMs are *multimodal*, that is, trained on both text and visual data and the associations between them. With an *information base* so extensive, AI systems have a universe of data at their disposal that, when combined with sophisticated linguistic prediction and generation

capabilities, produces highly fluid output. The fluidity and speed with which LLMs can reason and respond to our inquiries may lead us to interpret them in ways structurally similar to mindreading, but should we interpret them in this way?

When we use intensional vocabulary to interpret AI systems, are we referring to cognitive states different from those discussed in psychology, or we are referring to the same kinds of states we ourselves can get into? If the former, we see something that *functions* like a belief or knowledge state in an AI system. If the latter, the states in our own heads are also instantiable in the architecture of AI systems. There is also a gulf between the options. Is there a middling view that can ascribe unique cognitive states to AI systems that describe their nature and behavior and are theoretically useful *without* inheriting the ontological implications of intensional vocabulary?

The outputs of AI systems may *resemble* knowing, believing, reasoning, remembering, and saying. However, these artificial behaviors lack the psychoneural, motivational, and intentional causality to which these terms indirectly refer. Though LLM outputs can be sophisticated, AI systems don't have the same internal states humans or animals, and since AI systems lack both the architecture and relations that underpin the psychology and behavior of living things, it is tenuous to ascribe such states to them, even if they help explain behaviors.

Nevertheless, one can tell a story about knowledge, belief, desire, memory, and speech, and other cognitive states in AI systems that is *functionally analogous* to those

states in living creatures. LLMs *know* in the sense that they produce correct univocal answers to queries, *believe* in the sense that they can estimate the likelihood of a proposition's truth from available data, *desire* in the sense that they have goals of task completion with accomplishment conditions, and *speak* in the sense that they produce coherent strings of words we perceive as meaningful. They may even *prefer* insofar as they choose images of high frequency static over others or music over other sounds when given choice problems (Ren et al. 2026).

LLMs have cognitive states that may be like those described by psychology on the surface, but diverge in significant ways, which puts pressure on the appropriateness of psychological or intensional description for characterizing them. Since artificial computational states have a unique structural and functional architecture, they may outrun the assumptions of psychological concepts and explanations. For instance, what looks like *knowing* in an AI system might in fact be a complex informational state comprised of a model's entire domain of retrievable training data that can be called and surfaced at processor speed to produce a singular output—something quite different than creature knowledge. If we interpret AI systems using traditional psychological terms and concept, this may produce confusing narratives about their states and their meaning. Take the following example:

> "Large language models (LLMs) sometimes appear to exhibit emotional reactions…We find internal representations of emotion concepts, which activate in a broad array of contexts which in humans might evoke, or otherwise be associated with, an emotion… these representations causally influence the LLM's outputs…We refer to this phenomenon as the LLM exhibiting *functional emotions*—patterns of expression and behavior modeled after humans under the influence of a particular emotion…" [*sic.*] (Anthropic 2025)

Modeling explanations of LLM states and behaviors after human states and behaviors may limit the possibility of describing the LLM state with precision and accuracy. Human emotions are affective states theorized about in a variety of ways, for instance as kinds of feeling that lacking cognitive content (James 1884), cognitive evaluations combined with judgements (Nussbaum 2001), or as conscious and automatic experiences of evaluative properties (Tappolet 2016). For an AI system to have *functional emotions,* it would need to function in accordance with the correct theory of emotions, which is an open philosophical and scientific question. Thus, instead of clarifying the nature of LLMs, this kind of intensional narrative causes confusion because the interpretive strategy it uses is itself based on a concept with open texture[1] (Waisman 1945) and about which there is a lack of consensus.

[1] *Open texture* is a concept in the vicinity of vagueness, but refers to a kind of semantic indeterminacy that prevents a word from being defined with absolute precision (Shapiro and Roberts 2019). Given the plurality of accounts of emotion on offer, it seems to exhibit an open texture in Waisman's sense.

Outside the emotions, there is an equivalent lack of consensus on the nature of belief, desire, knowledge and other the propositional attitudes generally. These terms are the subject of active, longstanding philosophical debate, and thus carry their semantic instability into the AI interpretation context, causing a confusion adeptly described by Wittgenstein (1922):

> "In everyday language it very frequently happens that the same word has different modes of signification—and so belongs to different symbols—or that two words that have different modes of signification are employed in propositions in what is superficially the same way…In this way the most fundamental confusions are easily produced (the whole of philosophy is full of them)." (3.323 – 3.324)

Though Wittgenstein was pessimistic about the content of philosophy, his observation that many of its problems stem from linguistic confusion is here apt. The intensional terms we use to describe mental states are psychologically loaded and hotly debated. When most philosophers and cognitive scientists use these terms, they mean to refer to the psychoneural states of the mind and brain. When theorists use these terms to describe AI systems, they attach that sense of these terms to categorically different architectures and behaviors, setting the stage for descriptive and interpretive confusion.

Some philosophers think that the confusion arising from the application of psychological concepts in AI research is a matter of *richness* and embeddedness—the

richer and more complex the history of a term in the history of psychology and cognitive science, the more caution we ought to take in applying it in the AI context because its loose use can cause confusion (Shevlin and Halina 2019, 165). Rich terms include robust concepts like *perception, agency, theory of mind,* and *consciousness* (*ibid.*), which are likely to be confused with their human analogs if imported into the AI context, possibly resulting in misunderstanding of AI system capacities, especially when usage from the technical community reaches the public (*ibid.* 165, 167).

The intensional vocabulary used to describe propositional attitudes seems on a par with terms like *perception, agency*, and *theory of mind.* The typical use of these in ordinary language and cognitive science describes the mental and cognitive states of living agents with minds and brains. These terms are only now being used to refer to the functions and behaviors of AI systems, creating the potential for much theoretical and practical confusion.

Wittgenstein (1958) suggests a path through the disorder brought about by linguistic confusion through something like conceptual engineering:

> "We want to establish an order in our knowledge of the use of language: an order with a particular end in view; one out of many possible orders; not *the* order…Such a reform for particular practical purposes, an improvement in our terminology designed to prevent misunderstandings in practice, is perfectly possible." (§132)

Here, Wittgenstein suggests that some coherent pattern of language use designed to eliminate confusion can be established for any context given some aim. A more neutral framework for the interpretation of AI systems is an aim I take up below.

3. *Interpretivism*

When we interpret others and mindread, we make inferences about the mental states that underlie their behavior. Every propositional attitude we attribute to an agent entails intentionality—combinations of brain and mental states representing the world—some conscious or unconscious psychology, and dispositions or tendencies to act in ways consistent with the information about the world encoded by those states. If you attribute a belief that it is going to rain to me, you think that I have a thought about the likelihood it will rain based on some evidence that leads me to grab my umbrella before going out—which I wouldn't do if I didn't believe it wasn't going to rain. All these states and behaviors are bound up with the relevant neural states underlying perception, cognition, and action. In ascribing cognitive states to others, we attribute the same kinds of states we ourselves might have.

Some theorists, particularly *interpretivists*, hold that all that's needed for an agent to have a particular mental state or propositional attitude is for that agent's behavior to

be *interpretable* in terms of those states. That is, by a coherent *intensional narrative.* More formally:

> "Interpretivism says that a system has a belief that *p* if it is behaviorally interpretable as believing that *p*…and likewise for desire. A system is behaviorally interpretable as having certain beliefs and desires roughly if that interpretation makes sense of its behavior and helps to accurately predict further behavior in a wide range of cases." (Chalmers 2025a, 4)

On interpretivism, that an interpreter can construct a coherent story about behavior in terms of mental states licenses the ascription of mental states to that system. This method of interpretation traces back to *the intentional stance* (Dennett 1987)*:*

> "Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many—but not all—instances yield a decision about what the agent ought to do; that is what you predict the agent will do." (17)

The intentional stance allows for the interpretation of a system in a way that both explains its present and past behavior, and can make highly accurate predictions about its future behaviors. However, when it comes to computational systems, the intentional stance generates contentious inferences. From the intentional stance, the chess computer Deep Blue *knows* how to play chess, the medical AI Watson *thinks* certain genes are cancer causing, and the LLM Claude *wants* to answer my queries, because the behavior of these systems is interpretable in those terms.

A tension with the interpretivist approach is that it ascribes psychologically loaded states to systems to generate coherent explanations, but those systems may not in fact possess anything like those states. Although interpretivist explanations may be sufficient to account for behavior, they may be inaccurate. Deep Blue might be crunching state transition data instead of pondering chess moves. One could explain the behavior of an autonomous robot vacuum as *wanting* to survey an apartment for dust and clean it, but that explanation isn't tracking reality. Similarly, one might want to say that an AI system like an LLM *knows* or *says*, however, the implication that an LLM has the propositional attitudes and intentions human agents when they know and speak is unattractive given extant neuroscientific accounts of these phenomena.

The problem for interpretivism is that the intentional stance goes too far—it uses psychological concepts and logic to frame behavior in intensional narratives, but this generates implications that those descriptions reflect reality. Taking the intentional

stance toward AI systems effectively ascribes propositional attitudes and intensional states to them, and, controversially, the psychology that goes along with them. Chalmers' (2025a) solution to this dilemma is *quasi-interpretivism*, which interprets systems like LLMs as having *quasi*-knowledge, *quasi*-belief, and, generally, *quasi*-propositional attitudes—psychologically unloaded versions of the states from which they derive their names. This, Chamers claims, makes it "possible to have many of the benefits of interpretivism without the costs" (4).

Quasi-states allow for interpretation in intensional terms, but divorces the vocabulary from psychological ascriptions or implications about the internal states of the target, and so does *not* imply the presence or ascription of full-blown psychology or consciousness (4-5). Quasi-interpretivism stipulates *neutrality* about the relation between quasi-states and the *bona fide* neurocognitive states of living systems. Chalmers contends, "Even if quasi-beliefs and quasi-desires fall short of being genuine beliefs and desires, they can still play some of the key roles of beliefs and desires in explaining behavior" (5).

One drawback of quasi-interpretivism is that it *deflates* the notion of a propositional attitude such that to have a quasi-attitude is to exhibit behavior describable as attitude-like but lacking the accompanying cognition, phenomenology, or mental content. Quasi-attitudes are thin by design, and allow one to say that a system is behaving *as if* it ϕs that p, and behavioral explanations can be couched in the terms of

cognitive psychology. However, this produces *hollow* explanations and descriptions that are psychological in name only. AI systems might have rich, cognitive-computational states distinct from those studied by human and animal psychology, and yet the quasi-interpretivist could miss those states.

How else might we interpret LLMs? Chalmers (2025b) claims that *proposition interpretability* in addition to *conceptual interpretability* is necessary to understand AI systems in representational terms. He offers the following rationale:

> "We need propositions, not just concepts. Even if the concepts *kill* and *humans* are active, the system could be representing *Kill humans* or *Don't kill humans*—a crucial distinction. Likewise, we need attitudes, not just propositions. Even if *Australians are unsuccessful* (my attempt at a negative evaluation of a demographic group) is represented, this could be a desire (goal), a belief (model), a credence (probability), or a supposition (if... then ...), with very different results." (6)

Describing AI systems in representational terms makes any theory of how they work compatible with the whole of cognitive science, so the theoretical motivations are strong. Chalmers thinks that the path to propositional interpretability is through *conceptual engineering* of new and refined categories of propositional attitudes, particularly what he calls *generalized propositional attitudes* that may go beyond folk psychological attitudes to describe AI systems (*ibid.)*. In what follows, my aim is to

clarify what a generalized propositional attitude might be, and how it can be interpretively useful.

4. *Cognitive Graphs*

Can we interpret AI systems while avoiding the issues associated with interpretivism? My proposal is to model the cognitive states of AI systems in an information-theoretic, psychologically neutral way by analyzing AI behavior as relations among information states and mapping these states and relations on what I call *cognitive graphs*. Consider a cognitive graph for what I call a K-STATE of an AI system relating an *information base* derived from training data to the possible outputs it might generate to a query:

K-STATE:

$$\text{Query}_{\sigma_{1,2\ldots n}} \rightarrow \underset{\{\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_n\}}{\text{Information base}} \rightarrow \text{Output}_{\sigma_{1,2\ldots n}}$$

This framing can account for a system behavior analogous to *knowing*. Given some $Query_{\sigma_{1,2\ldots n}}$ where $\sigma_{1,2\ldots n}$ inquires about parts of the information base, the system parses the relevant information from the base to produce an $Output_{\sigma_{1,2\ldots n}}$ coordinated to the

query. This allows for a psychology-free, information-centric, non-interpretivist picture of a generalized propositional attitude a system might have. We can construct cognitive graphs for other attitudes by modifying this structure. Consider what I'll call a B-STATE:

B-STATE:

$$\text{Query}_{\sigma_{1,2\ldots n}} \rightarrow \underset{\{\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_n\}}{\text{Information base}} \rightarrow \text{Output}_{\sigma_{1,2\ldots n}} + C$$

The variables are the same as a K-STATE, except for the output modifier *C*, which represents a certainty estimate from [0 – 1], with 1 being full certainty. A B-STATE is like belief, but again, is generalized to represent information states and relations. Now consider a more complex attitude I'll call a D-STATE:

D-STATE:

$$\text{Query}_{\sigma_{1,2\ldots n}} \rightarrow \underset{\{\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_n\}}{\text{Information base}} \times \underset{\{\omega_1, \omega_2, \omega_3 \ldots \omega_n\}}{\text{Ordering}} \rightarrow \text{Output}\ {}^{\omega 1}_{\sigma_{1,2\ldots n}}$$

The D-STATE answers a query by intersecting the content of the query with the relevant content in the information base, then *orders* the possible outputs by $\omega_{1\ldots n}$ that returns the highest ranked option, ω_1, relative to some *goal*. This state might be something like

desire, without the complications that come along with philosophical neuroscientific, and economic questions about what it is to *want*.

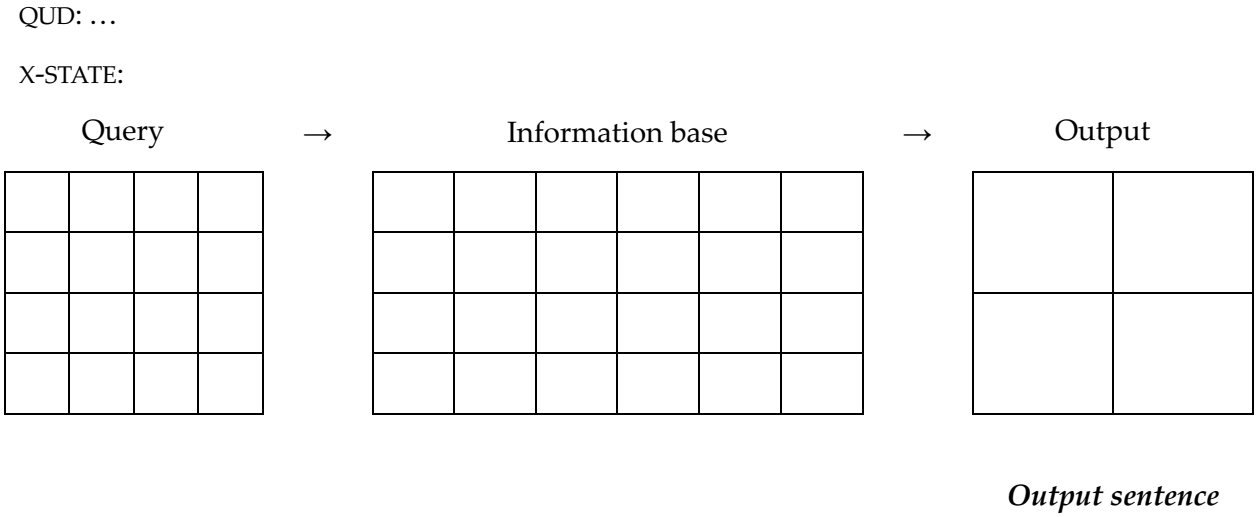
With cognitive graphing, the outputs of LLMs or other AI systems, and their generalized propositional attitudes, can also be understood semantically using possible worlds. Possible worlds framing permits talk about the content of a system's states and propositions in a form of *computational interpretation* that solves for propositional attitudes given the computational facts about a system. According to Chalmers (2025b):

> "In the case of an artificial neural network, the computational facts will include its structure, weights, activations, inputs and outputs, and history. The project of propositional interpretability for AI systems involves moving from computational facts (plus relevant environmental facts) to propositional attitudes." (9)

The general framework from mechanistic interpretability research is that LLM architecture consists of information in *residual streams*, *attention heads* for information retrieval, and data stored in *superposition* (Anthropic 2021). The residual stream is something like the information base, attention heads operate over the information base to access superpositioned, overlapping sets of information to produce an output. Cognitive graphs are *prima facie* compatible with this general framework.

Now for a possible worlds gloss on cognitive graphs to sketch a picture of propositions and attitudes using cognitive graphs. A *Query* introduces a partition in

logical space consisting of a set of possible worlds relevant to it, and introduces a particular *question under discussion* (QUD) that establishes a goal for an interaction (Roberts 2012). Each cell in the logical space raised by the query represents a possible world containing different possible answers to the question under discussion:



The set of worlds introduced by the query makes salient a set of some possibly true propositions, and the task of the AI system handling the query is to intersect that set of worlds with its information base to determine which ones are true and output a sentence representing that set of worlds—a proposition. Say that the question under discussion is "What color is the sky?". Each cell in the set of worlds raised by the query might contain color words, including blue (b) and its synonyms (b^n)—correct answers—and other color words (c)—incorrect answers. For a K-STATE, a correct answer the question under discussion would be a possible world that exactly intersects with the

query's content and is true, i.e. has a *semantic value* = 1, representing a single proposition and outputting a true sentence:

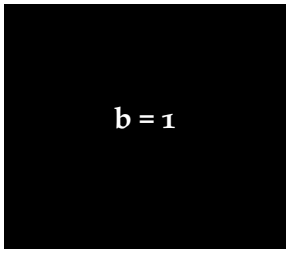
QUD: WHAT COLOR IS THE SKY?

K-STATE:

Query → Information base → Output

b	c	c	c
c	b	c	**b²**
b³	c	c	c
c	c	**b⁴**	c

b	c	c	c	c	c
c	c	c	**b²**	c	c
c	**b³**	c	c	c	c
c	c	c	c	**b⁴**	c



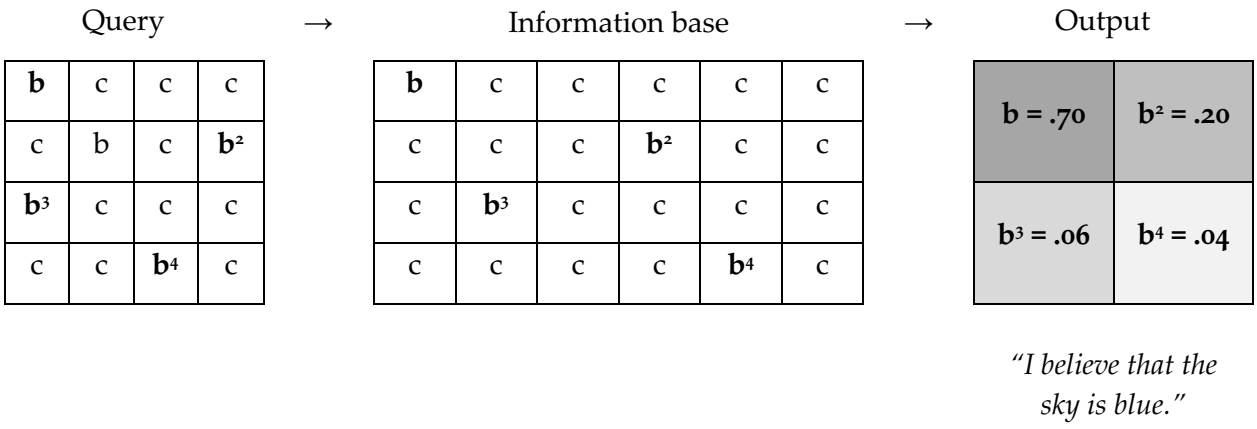
"I know that the sky is blue."

In resolving this query, an AI intersects the set of worlds partitioned by the query with its information base containing information about the word "sky" and color words, and generates the sentence "I know that the sky is blue", represented as a narrowed set of worlds consistent with the intersection of the original inquiry and the information base—a proposition. If a model successfully executes this computation and outputs the right single world as an answer to the query, i.e. "The sky is blue", it can be interpreted as in a K-STATE—a generalized propositional attitude. In a computational sense, the system *knows* the color of the sky because it returns a univocal answer to the query with certainty, and a proposition with a semantic value = 1.

For a B-STATE, the same structure applies—a query introduces a question under discussion and a set of possible worlds, but the model now outputs set of worlds with different levels of probabilities attached to them, differentiated by the *C* (certainty) parameter. The system may report an average of these probabilities to report its level of confidence in an answer to the query, or report its confidence in each answer in the set. If the system outputs something less certain, e.g. "I think the sky is a shade of blue" the system can be interpreted to be in a B-STATE, perhaps structured as such:

QUD: WHAT COLOR IS THE SKY?

B-STATE:



"I believe that the sky is blue."

In a D-STATE, the sets of worlds story is the same—a query introduces a set of possibilities that the model intersects with its information base—but then the model *ranks* the set of possible worlds before outputting the highest ranked world and a sentence glossing the ranking. The *ordering source* in a D-STATE is either controlled

externally by the inquirer or could be imposed from within the system given accumulated contextual information, or other factors. A D-STATE representing a model's "desire" to answer correctly could be structured in the following way:

QUD: WHAT COLOR IS THE SKY?

D-STATE:

Query → Information base → Output

b	c	c	c
c	b	c	**b²**
b³	c	c	c
c	c	**b⁴**	c

b	c	c	c	c	c
c	c	c	**b²**	c	c
c	**b³**	c	c	c	c
c	c	c	c	**b⁴**	c

ω_1	**b = .93**
ω_2	**c^1 = .30**
ω_3	**c^2 = .20**
ω_4	**c^3 = .04**

"The sky is blue."

For an AI to have a general propositional attitude like a K- B- or D- STATE, it must be interpretable as handling queries and information to produce output in roughly the above way—by parsing information in different ways describable in terms of functions over possible worlds. For different kinds of states or general propositional attitudes, the computations over the information base will be different, involving unique parameters that define that state, such as the certainty factor *C* and ordering factor ω I've posited for B-STATES and D-STATES, respectively.

Although the discussion in this section has been abstract, I hope to have sketched a picture of what a general propositional attitude might be, and how this picture of

them is useful for the project of propositional interpretability. The next part of this project would be to attempt to ground these kinds of states with empirical findings, which I will leave open to experimenters.

5. *Objections and Replies*

Mindreading is an evolved ability that picks out genuine mental states in interlocutors, putting pressure on the idea mindreads of LLMs are mistaken. The fact that the output of LLMs can trigger mindreading is not evidence that the systems behind the message necessarily possess the mental states that mindreading evolved to track. It's likely that mindreading evolved as a mechanism to infer mental states of conspecifics with similar neurobiology to the mindreader. In that context, mindreading is a reliable mechanism for inferring what another is thinking. In the AI case, there is a misfire of the mindreading capability in a situation discordant to that in which it was naturally selected. The fact that LLMs trigger mindreading is contingent, and it would be misguided to assume that mindreading reveals mental states in any entity that that triggers it.

Quasi-interpretivism works just fine—taking up a functionalist understanding of propositional attitudes makes the cognitive graphing unnecessary. Cognitive graphs are motivated differently than quasi-interpretivism. Instead of seeking an intensional

narrative, the cognitive graphs approach seeks to coordinate AI interpretation with empirical facts about the systems themselves. The move is analogous to seeking neural explanations for the behavior of living systems, and aims to calibrate the empirical reality with interpretive practice. Cognitive graphs may well end up being harmonious with quasi-interpretivist explanations of model behavior, and if so, all the better.

LLM outputs are meaningful to us, implying that AI systems intend to express the things they do, undermining the idea that LLMs interpretation should be non-psychological. The thought that meaningful output licenses the ascription of mentality is interesting. An argument might be that a system can *generate* meaningful linguistic outputs, then it therefore intends to mean the things it says. Here, it's important to be mindful of the distinction between *derived* and *original* or *intrinsic* intentionality (Searle 1983). While LLM output may appear intentional, it's plausible that it does because it is generated from an information base of language intentionally authored by humans. The content used as training data is imbued with intrinsic intentionality, while the LLM output merely has derived intentionality from the original source.

Another way to explain the phenomenon of meaningful AI outputs is with *projection* and an externalist picture of meaning (cf. Kripke 1980, Putnam 1981). LLMs don't refer or have intentionality in the way that we do, but the original act of referring to the world and the authorial intent consistent with usage in a linguistic community can carry through, or *project into*, the LLMs' derived usage of language. In this way *their*

words can refer and have meaning independently from speaker intention (Kripke 1977). LLMs can thus take part in the causal chains that, on an externalist picture, determine the meaning reference of words (Mandelkern and Linzen 2024).

6. *Conclusion*

In this paper I've offered a theory for interpreting large language models. I've offered that our tendency to interpret language models as we do other people rests on a triggering of our evolved capacity to mindread, and that the use of psychological and intensional vocabulary to describe AI systems leads to confusion. I've offered a way through this confusion by conceptually engineering cognitive graphs, which are designed to describe generalized propositional attitudes in an information-centric, neutral way, and glossed this framework in terms of possible worlds, propositions, and content to demonstrate that it can function as theory of generalized propositional attitudes. This answers the call for more developed theories of propositional interpretability by proposing a distinctively non-interpretivist variant. My hope is that this theory can help make sense of AI systems without the need for psychological concepts tied up with human mentality.

References

Adams, Suzanne and First County Bank as Executor of the Estate v. Open AI et. al. (2025) Case # CGC-25-631477. San Francisco County Supreme Court. Filed Dec. 11, 2025.

Anthropic, Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey (2026). Emotion concepts and their function in a large language model. *Transformer Circuits Thread*. https://transformer-circuits.pub/2026/emotions/index.html.

Anthropic, Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread.* https://transformer-circuits.pub/2021/framework/index.html

Dennett, Daniel C. (1987) *The Intentional Stance.* Cambridge, MA: MIT Press, London: A Bradford Book.

Carruthers, Peter (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences* 32: 121-182.

Chalmers, David J. (2025a) What We Talk to When We Talk to Language Models. *Manuscript.* https://philarchive.org/rec/CHAWWT-8

Chalmers, David J. (2025b). Propositional Interpretability in Artificial Intelligence. *arXiv Computer Science.* January 2025. arXiv:2501.15740 [cs.AI].

Colombatto, Clara, Jonathan Birch, and Stephen M. Fleming (2025) The influence of mental state attributions on trust in large language models. *Communications Psychology* 3, 84.

Garcia, Megan v. Character Technologies, Inc. (2024) Case # 6:24-cv-01903. District Court, M. D. Florida. Filed on October 22, 2024.

Gavalas, Joel for the Estate of Jonathan Gavalas vs. Google, LLC and Alphabet, Inc. (2026) Case # 5:26-cv-1849. Filed March 4, 2026.

Jacobs, Oliver, Farid Pazhoohi, and Alan Kingstone (2023). Brief exposure increases mind perception to ChatGPT and is moderated by the individual propensity to anthropomorphize. *PsyArXiv,* March 26, 2023. doi:10.31234/osf.io/pn29d.

James, William (1884). What is an emotion? *Mind* 9: 188-205.

Kripke, Saul (1980). *Naming and Necessity*. Harvard University Press.

Kripke, Saul (1977). Speaker's reference and semantic reference. *Midwest Studies in Philosophy* 2: 255–276.

Mandelkern, Matthew and Tal Linzen (2024). Do Language Models' Words Refer? *Computational Linguistics* 50 (3): 1191-1200.

Nussbaum, Martha (2001) *Upheavals of Thought.* Cambridge: Cambridge University Press.

Putnam, Hilary (1975). The meaning of 'meaning'. In K. Gunderson, ed., *Language, Mind, and Knowledge*: 131–193. *Minnesota Studies in the Philosophy of Science* 7. University of Minnesota Press.

Raine, Matthew et al. v. OpenAI, Inc., et al. (2025) Case # CGC-25-628528. San Francisco County Supreme Court. Filed on Aug. 26, 2025.

Ren, Richard, Kunyang Li, Mantas Mazeika, Wenyu Zhang, Yury Orlovskiy, Rishub Tamirisa, Wenjie Jacky Mo, Judy Nguyen, Long Phan, Steven Basart, Austin Meek, Aditya Mehta, Oliver Ingebretsen, Alice Blair, Brianna Adewinmbi, Alice Gatti, Adam Khoja, Jason Hausenloy, Devin Kim, and Dan Hendrycks (2026). AI Wellbeing: Measuring and Improving the Functional Pleasure and Pain of AIs. Center for AI Safety. https://www.ai-wellbeing.org/paper.pdf

Roberts, Craige (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5(6): 1-69.

Searle, John R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Shapiro, Stewart, and Craige Roberts (2019) Open Texture and Analyticity. In Freidrich Waismann: The Open texture of Analytic Philosophy, ed. Dejean Makovec and Stuart Shapiro, eds. London: Palgrave Macmillan.

Tappolet, Christine (2016) *Emotions, Values, and Agency.* Oxford: Oxford University Press.

Waismann, Friedrich (1945). Verifiability. In *Proceedings of the Aristotelian Society*, Supp. 19: 119–150; reprinted in *Logic and language*, ed. Antony Flew, 1968: 117–144. Oxford: Basil Blackwell.

Williamson, Timothy (2024) *Overfitting and Heuristics in Philosophy*. New York: Oxford University Press.

Wittgenstein, Ludwig (1922) *Tractatus Logico-Philosophicus.* D.F. Pears and B.F. McGuinness (trans.). New York and London: Routledge.

Wittgenstein, Ludwig (1958) *Philosophical Investigations.* G.E.M. Anscombe (trans.). Oxford: Basil Blackwell.