Creative Minds Like Ours?

Assessing the Linguistic Creativity of Large Language Models

Preprint

By Vincent J. Carchidi

Abstract

The creative aspect of language use (CALU) is the *stimulus-free*, *unbounded*, yet *appropriate and coherent* use of language by human beings. This ordinary use of language is distinctive: it describes the species-specific ability to deploy one's cognitive resources to any problem or task that one sees fit to tackle or to invent new problems altogether. In this way, the human ability to use language to appropriately express new thoughts in a manner that is causally independent of one's local circumstances enables the sheer scope of human creativity. With the rise of Large Language Models (LLMs) has come a burgeoning interest in the creative and linguistic capabilities of these systems. CALU has thus far been largely absent from this literature. This paper fills the gap by explicating CALU from its roots in Cartesian philosophy and its revival in the biolinguistic approach articulated most prominently by Noam Chomsky. It then assesses whether LLMs exhibit CALU. Finding that LLMs only exhibit one interpretation of the "unboundedness" criterion, and therefore fail to demonstrate that they possess creative minds like ours, the implications of this result for the future of computational creativity are briefly explored.

Keywords: Artificial Intelligence; Cartesian Creativity; Computational Creativity; Creative Aspect of Language Use; Generative Linguistics; Large Language Models

Introduction

The creative aspect of language use (CALU) is a central yet neglected feature of human creativity. With roots in Cartesian philosophy, and more recently associated with generative linguistics and the work of Noam Chomsky, CALU describes the *stimulus-free*, *unbounded*, yet *appropriate and coherent* use of language by humans (Chomsky, 1966, 3-30; McGilvray, 2001, 6-13; Asoulin, 2013, 228-232). This distinctive human ability is essential to any account of human creativity and the intellectual and artistic pursuits it enables. Any account of *computational* creativity, by the same token, must take CALU seriously. The rise of Large Language Models (LLMs) and accompanying interest in their linguistic and creative capabilities thus allows us to ask: do LLMs exhibit this creative aspect of language use? Or do they, like machine past, "[act] in accordance with [their] internal configuration and external environment, with no choice" (Chomsky, 1988, 6)?

This paper aims to answer this question and draw out CALU's implications for computational creativity. To do so, it traces this idea's origin back to René Descartes (1910/1637) whose proposition that two tests, one of which includes the distinctive use of language by humans, can determine whether an organism or machine has a *mind like ours*. Subsequent Cartesian work by Géraud de Cordemoy (1668), who emphasized the connection between *language use* and *thought* in testing for the existence of minds like ours, is reviewed alongside Descartes' remarks. Chomsky extracts from these writings a set of observations that crystallize in the judgment that

> man has a species-specific capacity, a unique type of intellectual organization which cannot be attributed to peripheral organs or related to general intelligence and which manifests itself in what we may refer to as the "creative aspect" of ordinary language use —its property being both unbounded in scope and stimulus-free. Thus Descartes

> maintains that language is available for the free expression of thought or for appropriate response in any new context and is undetermined by any fixed association of utterances to external stimuli or physiological states (identifiable in any noncircular fashion) (Chomsky, 2009a, 60).

CALU is a unique ability, grounded in human biology, to use language as a free yet appropriate and unbounded expression of thought. The ability to deploy one's cognitive resources through language to any problem that one sees fit to address is central to humanity's intellectual and artistic powers and, thus, any account of human creativity.

What is the value of extending the Cartesian test of other minds to modern language models? The importance of "Cartesian creativity" (D'Agostino, 1984) and its implications for artificial intelligence (AI) should be clearly stated.

The free deployment of one's cognitive resources, robustly applying them to any new problem or task is an ability that underwrites the scope of human ingenuity. A range of *competencies* residing within the mind provide the tools that shape the structure of human cognition in various domains. And it is, to be sure, the nature of the mind's "core knowledge systems" that underwrite what Spelke (2010) sees as the uniquely human ability to link these systems' representations together in the construction of new concepts and ideas through natural language's combinatorial capacity. Individuals also, as CALU describes, *execute* this ability in ordinary life in a unique fashion—a human being regularly uses language to express new thoughts or ideas in a manner that is relevant to the thoughts and needs of others yet casually detached from their local environment. Language use is routinely creative in this way, allowing human intellectual and artistic powers to be expressed at will and without limit in a fashion that complements the mental states of others. To say that human language use is creative, then, is to offer a non-technical description of how human beings apply their uniquely structured cognition to their highest endeavors.[1]

Human beings are creative creatures. Their language use lies at the center of their creativity. Any assessment, then, of whether LLMs possess not only minds but *minds like ours* depends on understanding exactly *how* human beings manifest their creative powers. Through the stimulus-freedom, unboundedness, and appropriateness of language use, we can produce one such test. While linguistic debates over LLMs are prominently focused on how they relate to matters of the poverty of the stimulus argument and the acquisition of a human-like competence (e.g., Dentella, Günther, and Leivada, 2023; Katzir, 2023; Piantadosi, 2023), CALU takes a different route: it observes and describes what generative linguists refer to as *performance*—linguistic behavior. LLMs' rise thus provides a compelling reason to explicate CALU and assess their capabilities through this lens.

The paper begins with a high-level overview of research in human and computational creativity, connecting this to our project here. Then, the concept of CALU—from its Cartesian roots to generative linguistics—is explicated, followed by responses to the most pressing criticisms of this idea. With this background established and the parameters of a test of creative minds like ours in place, we turn our attention to LLMs. We review how LLMs are designed and trained, and then turn to a test of whether they exhibit the stimulus-free, unbounded, yet appropriate use of language. Finding that LLMs achieve only one interpretation of "unboundedness," and fail to achieve stimulus-freedom and appropriateness, we briefly explore

[1] To be sure, to their lowest endeavors, too—CALU is not an inherently moral or wise phenomenon. One can deploy their cognitive resources freely for all manner of activities, though the positive illustration in the main text is most relevant for our purposes in assessing LLMs.

the implications of this result for the possible artificial reproduction of CALU in future AI systems.

Human and Computational Creativity

Creativity studies is an interdisciplinary topic encompassing both philosophy and cognitive science, with recent applications to AI. For example, a recent study of LLMs by Franceschelli and Musolesi (2023) invokes Boden's (2004) cognitive-inspired taxonomy of creativity in finding that these systems possess a "combinatorial" creativity—the ability to re-combine familiar elements in unfamiliar ways. There, they write: "Creative-person qualities in Generative [artificial intelligence] might eventually be the ultimate step in achieving human-like intelligence" (Franceschelli and Musolesi, 2023, 6). Instilling *human*-like creative capabilities in AI systems is notably high on the AI improvement agenda.

Indeed, human creativity refers to the "characteristics and cognitive behavior of creative people and the environment or situations in which creativity is facilitated," whereas computational creativity, "while inspired by concepts of human creativity, is often expressed in the formal language of search spaces and algorithms" (Maher, 2010, 22). This is the distinction, as Boden puts it, between "understanding *human* creativity" and "trying to produce machine creativity…in which the computer at least *appears* to be creative to some degree" (Boden, 2004, 1). Ritchie echoes this in the warning that "any formal definition of creativity must be based on its ordinary usage; that is, it must be *natural* and it must be *based on human behaviour*" (Ritchie, 2007, 69). Without using a conception of creativity derived from its ordinary human uses, the danger of circular argument looms (Ritchie, 2007, 69-70).

Several criteria have been proposed as metrics to evaluate whether an idea or artifact (i.e., product) is creative. A representative list of criteria suggested by scholars includes the *nature* of specific artifacts—say, a poem considered on its merits alone—as well as the *mindsets*, *intentions*, and *actions* of those responsible for producing them and the social and historical *contexts* in which they are produced.

Franceschelli and Musolesi (2023, 6), drawing from Mel Rhodes (1961, 307-310), offer a useful overview of these criteria. This includes the creative product itself, the role of motivation, thinking, and communicating (the process), and the relationships between the potentially creative artifact and the individual that produced it as well as the social and historical context of its origin (the person and press).

One can also find a distinction between what is called *psychological* creativity and *historical* creativity. The former refers to the generation of a creative idea or artefact that is novel to the individual who produces it. The latter is broader in scope, referring to the generation of a creative idea or artifact that has never arisen before in the entirety of human history (Boden, 2004, 2; Franceschelli and Musolesi, 2023, 4).

The notion of creativity, then, can be an expansive one. Floridi and Chiriatti echo their support of an expanded notion of creativity in their affirmation that "it is not *what* is achieved but *how* it is achieved that matters" (Floridi and Chiriatti, 2020, 687). "*[H]ow* the original and valuable product is made," Gaut argues, "plays an essential role in determining whether the act of making it is creative…the making must involve *flair* by the maker to rule out" products generated "by chance or by mechanical procedure" (Gaut, 2003, 270).

Creativity, Cognition, and LLMs

To be sure, definitions of creativity vary in their respective emphases on the criteria laid out above. Boden's definition is the most recognizable: creativity is "the ability to come up with ideas or artefacts that are *new*, *surprising and valuable*" (Boden, 2004, 1). Still, as Wang et al. note in their exploration of AI systems' creativity, the "inherent subjectivity" of measuring or identifying novelty and value "remains problematic" (Wang et al., 2024, 2).

The study of LLMs' creative capabilities through the lens of the mind is in its infancy and, to an extent, parallels this subjectivity in evaluations of creativity's criteria. Existing work is familiar to the psychological and cognitive sciences. Binz and Schulz (2023), for example, sought to investigate LLMs like GPT-3 using vignettes drawn from cognitive psychology. Zhao et al. (2024) probe LLMs using a modified version of the Torrance Tests of Creative Thinking (TTCT), judging the LLMs' responses to questions according to their fluency, flexibility, originality, and elaboration (Zhao et al., 2024, 5-6). Franceschelli and Musolesi's (2023) aforementioned application of Boden's taxonomy of creativity to LLMs—combinatorial, exploratory, and transformational creativity (Boden, 2004, 3-6)—is supplemented by considerations for how creative products are generated, the motivation for doing so, and the social contexts in which it occurred.

Applications of CALU are largely absent from the literature on computational creativity. Notable exceptions include Pulman's (2018) lecture on CALU and AI which argued that the three criteria had not yet been met by state-of-the-art models, Moro, Greco, and Cappa's (2023, 83) note that LLMs do not achieve stimulus-freedom, and Ji's (2024) more recent argument explicitly arguing that LLMs do not achieve CALU.

CALU does, nonetheless, represent the "how" of creativity, to put it in the above terms. It is not necessarily about what human beings create, but *how* they do so—which will influence the product in question—and it is grounded in ordinary human behavior, consistent with Ritchie's (2007) remark. Creative linguistic productions are also routinely novel in the senses of psychological and historical creativity.

What follows in this paper extends sparse remarks on CALU by rigorously explicating the concept and testing whether LLMs exhibit this use of language. Implications for computational creativity are drawn out thereafter.

With that, we turn to the idea at hand.

The Creative Aspect of Language Use

CALU's Cartesian Roots

The observation that human beings use their language in a "creative" fashion has its roots in Cartesian philosophy. A reconstruction of remarks made by rationalists including Descartes and Cordemoy reveals an emphasis on a distinctively human characteristic: that new ideas and concepts are expressed through novel combinations of words at will and within contexts for which there is no stimulus causing these expressions. Yet, despite this novelty and causal detachment from local conditions, the expressions are appropriate to the circumstances of their use and to others who hear (or read) them.

This ability is responsible for the unique way in which humans go about constructing new ideas and applying them to problems and tasks as they see fit. It serves as a fundamental contrast between human beings and non-human animals and machines. Indeed, CALU provides evidence, should it be observed in a subject under investigation, of a "mind like ours" (Chomsky, 1988, 5).

McGilvray (2001, 6) observes that particular reference is made by Chomsky to Descartes' 1637 *Discourse on Method* where he writes on the distinction between machines that possess the outward appearance of human beings and actual human beings—the ability, that is, to identify other minds sufficiently similar to our own despite surface-level appearances. There, Descartes notes, despite the possibility of "machines bearing the image of our bodies, and capable of imitating our actions as far as it is morally possible, there would still remain two most certain tests whereby to know that they were not therefore really men" (Descartes, 1910/1637, 60). The first test is as follows:

> Of these the first is that they could never use words or other signs arranged in such a manner *as is competent to us* in order to *declare our thoughts to others*: for we may easily conceive a machine to be so constructed that it emits vocables, and even that it emits some correspondent to the action upon it of external objects which cause a change in its organs…but not that it should arrange them variously so as appositely to *reply to what is said in its presence*, as men of the *lowest grade of intellect can do* (Descartes, 1910/1637, 60) (emphases added).

Descartes' first test proposes that a machine possesses a mind like ours if: it utters words (or "signs") in an *intelligible* manner ("competent to us"); expresses a *thought* through variously arranged words of its own *volition* ("to declare" them); and expresses such thoughts through words in a manner that is *appropriate* to the remarks it has heard from others ("appositely to reply to what is said in its presence"). Each of these abilities, Descartes notes, is exhibited by humans of the "lowest grade of intellect" —they are not features of higher human intelligence nor are they acquired forms of expertise.

Descartes goes on to describe the second test:

> [A]lthough such machines might execute many things with *equal or perhaps greater perfection than any of us*, they would, without doubt, fail in certain others from which it could be discovered that they did not act from knowledge, but *solely from the disposition of their organs*: for while Reason is an universal instrument that is alike available on every occasion, *these organs, on the contrary, need a particular arrangement for each*

> *particular action*; whence it must be morally impossible that there should exist in any machine a diversity of organs sufficient to enable it to act in all the occurrences of life, in the way in which our reason enables us to act (Descartes, 1910/1637, 61) (emphases added).

The second test recognizes that machines may carry out characteristically human tasks with comparable or even greater precision than humans themselves, but they will always do so in an *induced* manner. That is, the "particular arrangement for each particular action" that Descartes describes is an allusion to the context-dependency of mere machines.

These two tests allow human observers, then, to distinguish between machines that act with the mere *appearance* of humanity and those organisms that do, in fact, exhibit characteristics that indicate the presence of a mind like ours. In Chomsky's interpretation, "it is the diversity of human behavior, its appropriateness to new situations, and man's capacity to innovate…that leads Descartes to attribute possession of mind to other humans…" (Chomsky, 2009a, 61).

Nevertheless, some argue that the focus on language use specifically is a misinterpretation of the above passages. Joly argued that Chomsky's reading of Descartes "is, to say the least, an over-interpretation of the original text (Joly, 1985, 146), emphasizing Descartes' interest in the "relationship of language to reason…he is interested in language insofar as it reveals the existence of Mind and Reason" (Joly, 1985, 147). Chomsky's focus on "the possibility of creating infinitely many sentences with finite means" (Joly, 1985, 147) is thus misguided.[2]

[2] Nevertheless, in the original edition of *Cartesian Linguistics*, Chomsky plainly notes that his focus on the work of figures like Descartes is done not from the latter's perspective but from his own (Chomsky, 1966, 2). That this is a problem for Chomsky's *articulation* of CALU is not apparent. For our purposes, what matters is the soundness of the idea of CALU.

This objection, however, conflates the unbounded use of language with CALU's other criteria, namely stimulus-freedom and appropriateness. Moreover, Cartesian philosophers following Descartes, Cordemoy in particular, made an explicit association between *thought* and how language is *used* in attempting to identify that which distinguishes humans from machines, as we see below.

To be sure, some discussion over Chomsky's articulation of CALU has centered on two ways of interpreting the unboundedness of language use. One way is to interpret the unbounded character of language in purely *syntactic* terms in which words and phrases can be combined into sentences of undefined lengths or kinds without an apparent upper limit on this ability. A second interpretation is *semantic* (see, Onsman, 1982, 42-49). Still, when Descartes references the human ability to "use words or other signs," this is predicated upon these words and signs being "arranged in such a manner as is competent to us in order to declare our thoughts to others" (Descartes, 1910/1637, 60). The combinatorics of language use are, in this interpretation, in the service of the expression of novel thoughts.

What is important to note is that on either interpretation, human thought is inextricably bound up in human language use—novel combinations of words reflect novel expressions of thought. To identify CALU and conduct a test of other minds, it is not necessary to sharply delink syntactics from semantics, nor to get caught up in debates about whether the former is rule-governed in the sense conveyed by generative grammar. The relationship between language use and thought in the Cartesian tradition is evident in Cordemoy's extension of Descartes' test for other minds.

Cordemoy begins his 1668 *Philosophicall Discourse Concerning Speech* with the observation that it is natural to believe that "bodies," like his own, are united with "souls" (i.e., minds). Such bodies, however, may appear this way outwardly but cannot be said to be “united to Souls, until I have examin’d all their actions” (Cordemoy, 1668, 2). The problem arises, however, in recognizing that the “sole *Disposition* of the *Organs* is the cause” of such actions that will “maintain them in a state suitable to their nature…” (Cordemoy, 1668, 4). So enters the relationship between words and thought: “the connexion, I find between the *Words*, I hear them utter at all times, seems to demonstrate to me, that they have *Thoughts*” (Cordemoy, 1668, 6).

How, though, does one know that the words uttered through *speech* reflect a “soul” like his own, and are not merely a result of “*Mechanicks* one may so fitly adjust certain Bodies to one another” (Cordemoy, 1668, 8)? Speech can be achieved through the manipulation of air or by a parrot that hears one’s words—yet these do not reflect the operations of *thought* Cordemoy attributes to humans like himself (Cordemoy, 1668, 9-13). Words and thoughts are linked in humans, but in probing for the existence of a mind in bodies that appear human-like, it is not sufficient that they possess “*the facilness of pronouncing Words*” to conclude “*that they had the advantage of being united to Souls*” (Cordemoy, 1668, 13-14).

The key, then, is *how* words and other signs are used. It is thus clear that Cordemoy does not believe it is merely the ability to generate words in whatever diversity a “body” can muster that is sufficient to attribute to it a mind like our own. Indeed, Cordemoy’s observations of what indicates the presence of a mind like ours embrace a certain complexity involved in human expression:

> But yet, when I shall see, that those Bodies shall make signes, that shall have *no respect at all to the state they are in*, nor to their conversation: when I shall see, that those signs shall agree with those which I shall have made to express my *thoughts*: When I shall see, that they shall give me *Idea’s*, *I had not before*, and which shall relate to the thing, I had already in my mind: Lastly, when I shall see a *great sequel between their signes and mine*, I shall not be reasonable, If I believe not, that they are such, as I am (Cordemoy, 1668, 18-19) (emphases added).

It is not unreasonable to believe that a machine or organism possesses a mind like our own if it exhibits the following characteristics: it makes signs that do not arise in necessary connection with their environment or the conversation itself; it makes signs that complement and correspond with the signs that the human interlocutor may have used to express their thoughts; it communicates a *novel idea* that nonetheless corresponds with the human's existing thoughts; and the signs of organism build upon those of the human.

Cordemoy is not exclusively focused on the use of spoken words—for him, the relevant expressions of thoughts ("signs") are carried out "by *gestures*, by *discourse*, or by *characters*" (Cordemoy, 1668, 21). Language that is spoken, written, or gestured is his principal means of detecting a mind like his own. Cordemoy's central point, then, is that observing and describing how humans use speech in the detection of minds like our own relates human thought to human expression. The link between thought and language use is evident in both Descartes' remarks ("to declare our thoughts to others" (Descartes, 1910/1637, 60)) and Cordemoy's extension of the Cartesian tests for other minds ("those signs shall agree with those which I shall have made to express my *thoughts*" (Cordemoy, 1668, 18-19)).

From these Cartesian observations about language use, three criteria are implicated in making it creative: stimulus-freedom (independence from local contexts), unboundedness (novel expressions that reflect new thoughts and ideas), and appropriateness and coherence to circumstance (complementary of and correspondence with another human's thoughts).[3]

[3] Root (1975, 334-341) and Sampson (2016) both argue that this this form of creativity may not be exclusive to language. Each raises the generation of novel integers, the former in the case of a machine generating random integers indefinitely. This, however, is neither stimulus-free nor appropriate. Root also argues that purposive walking is creative, but, as shown below, appropriate language use is not merely functional or goal-oriented but corresponds with and complements human thought. Walking is not appropriate in this sense, nor innovative. McGilvray has aptly noted, though, that while creativity may be observed in other human behaviors, it is "perhaps to a reduced extent" in contrast to the case of language use (McGilvray, 2014, 38).

The question of syntactic or semantic unboundedness can thus be overcomplicated. Whereas we cannot determine whether language *itself* is a form of thought or merely serves as a *medium* of thought, the Cartesian test for other minds takes language use of a stimulus-free, unbounded, and appropriate character to be linked in a substantive way with human thought.

Formalizing CALU

We formalize creative human language use below:

Stimulus-Freedom

Individuals use their language in a manner that is not causally tied to factors in their local environment. Nor, furthermore, are such linguistic productions causally tied to the internal physiological state of the individual at the time of their use. Linguistic productions are, instead, *elicited* by stimuli rather than *caused* by them (McGilvray, 2017, 187; Asoulin, 2013, 230).

This is not to deny that external stimuli—say, another person speaking—are related to an individual's use of language. The contexts in which humans use their language are brimming with stimuli and we are inclined to act in a manner that is consistent with social norms, etiquette, and the like. It would be rude, for example, to respond to "Could you tell me the time?" with a blank stare, so we instinctively give the information—but the choice to not answer is there all the same.

The emphasis, then, is on these utterances being *elicited* by local stimuli. Attempts to construct a serious *causal* explanation for an individual's linguistic productions will, in contrast, merely be an interpretation of many different factors about the context in question, yet "[t]his is not the well-defined causality of serious theory…" (McGilvray, 2001, 7).

Indeed, a person can speak, to use Asoulin's (2013, 230) example, of Federico Lorca's *Poet In New York* while confronted by the unrelated stimuli of elephants and the African landscape. Moreover, Descartes' observation that minds like ours "declare our thoughts to others" (Descartes, 1910/1637, 60) indicates that one can produce linguistic utterances without external compulsion. Additionally, one can produce utterances that relate to an entirely *fictional* environment (McGilvray, 2017, 187). The local environment in which I am typing—and, likely, in which you are reading—has no causal relationship with these linguistic productions unless one attempts—in vain—to find a causal factor related to the generative conception of stimulus-freedom that simply forces me to type these words and not others.

The notion of "stimulus control" has been present in Chomsky's work since he reviewed B.F. Skinner's (1957) *Verbal Behavior*. There, Chomsky writes that "[w]e cannot predict verbal behavior in terms of the stimuli in the speaker's environment, since we do not know what the current stimuli are until he responds" (Chomsky, 1959, 32).

The contexts in which humans use their language are brimming with stimuli. The point, however, is that none of them appear to control this use—language use is stimulus-*free*.

Unboundedness

There is no apparent limit on the human ability to construct new words, phrases, and sentences that they have not previously encountered. Linguistic productions are not confined to a pre-sorted list of words, phrases, and sentences, allowing for both novel combinations therein and the expression of novel thoughts, ideas, and concepts. Human language use, in this way, is innovative over an *unbounded* range.

Appropriateness and Coherence to Circumstances

Language use is regularly appropriate to the circumstances of its use and coherent to others who hear (or read) the remarks. To say that one's use of language is "appropriate" to a given context means that it "is recognized as appropriate by other participants in the discourse situation who might have reacted in similar ways and whose thoughts, evoked by this discourse, correspond to those of the speaker" (Chomsky, 1988, 5). The thoughts of one individual, as Cordemoy observed, complement and correspond with the thoughts of another through language use.

It is important to note that the appropriateness of language use is not to be confused with regular correlation with one's environment. Indeed, to be appropriate is *not* "functional" in the sense that it is oriented towards a particular goal that has been established within the given context (McGilvray, 2001, 9).

CALU In Perspective

Taken together, these three dimensions of language use can be called "creative." These three components must *simultaneously* be present in linguistic productions for the capacity to be described as "creative" (Baker, 2008, 236-237). What's more, they cannot, as Descartes and Cordemoy recognized in their own formulations, be explained through mechanical terms (McGilvray, 2017, 187; Chomsky, 2009a, 61-62). The appropriateness of stimulus-free and unbounded language use is *undetermined* (Chomsky, 2009b, 175) yet also, bafflingly, not *random* or *probabilistic*.

To be clear, it is the *appropriateness* condition that throws a scientific curveball and most prominently distinguishes humans from animals and automata. One can conceive, in mechanical terms, of unbounded but stimulus-controlled language use, or bounded yet stimulus-free language use. But human language use is unbounded, stimulus-free, yet appropriate to the circumstances of its use in a manner that corresponds with the thoughts of others. Utterances are appropriate *despite* their unboundedness and stimulus-freedom.

The ordinary human use of language, then, "is not a series of random utterances but fits the situation that evokes it but does not cause it, a crucial if obscure difference," one providing evidence of a "mind like ours" (Chomsky, 1988, 5). This makes human language use not only distinctive but also, to-date, unexplained.

Objections and Responses

Several objections can be raised against the idea of CALU. The most pressing are responded to below.

First, Julien Offray de La Mettrie (1912/1747) argued against the Cartesian conception of the human mind. Specifically, he argues, contra Descartes, that all human action can be explained in mechanical terms, proclaiming that the "human body is a watch, a large watch constructed with such skill and ingenuity…" (La Mettrie, 1912/1747, 141). Humans "are at bottom only animals and machines…I believe that thought is so little incompatible with organized matter…" (La Mettrie, 1912/1747, 143). He observes that humans, like animals, exhibit a highly specific process of growth and maturation (La Mettrie, 1912/1747, 144-145). There is, thus, no fundamental difference between them, no second substance "but a single substance differently modified" (La Mettrie, 1912/1747, 148).

La Mettrie's objection puts Chomsky in an unusual position given Chomsky's fundamental view of language acquisition as a matter of biological development (see, Collins, 2008, 102). As Chomsky (2009a, 64-65) observes, however, La Mettrie does not offer an argument against Descartes' articulation of linguistic creativity. So, while human beings *are* biological organisms whose constitutions are shaped by their genetic makeup, this creates no contradiction with the observation that human language use "can serve as a general instrument of thought and self-expression rather than merely as a communicative device of report, request, or command" (Chomsky, 2009a, 64).[4] Indeed, CALU *informs* the biolinguistic approach to language (McGilvray, 2009, 3-4).

Second, Harman (1968) levels a bundle of interrelated criticisms against the idea of CALU itself in a review of Chomsky's *Cartesian Linguistics*. Harman's remarks are reconstructed as four separate criticisms:

(1) External stimuli *do* serve as forms of stimulus-control: "What the hearer understands the speaker to say is, or ought to be, a function of what is said" (Harman, 1968, 232).
(2) Internal stimuli *do* serve as forms of stimulus-control: Chomsky's assertion that internal stimuli do not control language use functions effectively as the claim that the factors underlying linguistic performance have simply not yet been discovered, rather than a denial of their causal relationship to language use (Harman, 1968, 231-234).
(3) Only practical unboundedness can be observed in human language use; "Literal unboundedness, if it exists, cannot be simply observed. It is established, for example, by showing that sentences of any arbitrary length can be constructed as conjunctions of shorter sentences" (Harman, 1968, 232).
(4) Even if one does make the case for literal unboundedness, this "would show little about the unbounded nature of thought…" (Harman, 1968, 232).

We take these criticisms in order.

Consider external stimuli. As noted, Chomsky in his review of *Verbal Behavior* argued that it is circular to claim an external stimulus is controlling a person's response as the relevant external stimuli is not identified until *after* the person responds. Moreover, the ability to speak of entirely fictional circumstances undermines the notion of external stimulus-control. "A planet in the Andromeda galaxy was discovered on January 13, 2024, to be the home of an alien species, some members of which are language scientists who reject the biolinguistic framework popular

[4] Chomsky (1977, 133-134) indeed notes that free human action is not incompatible with biological laws and, indeed, *dependent* on innate human structure.

with certain *homosapiens*" is a sentence with meaning, though not factual, and certainly not triggered by a stimulus in my local environment. External stimuli are real, and individuals are *inclined* to respond in a manner consistent with etiquette or coercion,[5] but they are not linguistically *controlling* stimuli.

Next, internal stimuli. Humans express themselves and their thoughts in a manner that is intelligible to others nearly immediately or with little effort—I can express my opinion on a topic of interest or an endeavor with which I am engaged, and another person may respond with expressions of their own experiences for which I lack any direct knowledge. Linguistic creativity allows humans to "adapt to various environments and solve (and create) problems well out of the range of any other kind of creature…make and interpret art, develop various forms of religions and the kinds of explanation they offer, develop ourselves and our cultures" (McGilvray, 2009, 4). The potentially infinite scope of human intellectual and artistic activity—intertwined tightly with the social nature of the species—does not find a realistic explanation in internal stimuli *causing* linguistic productions.

Animals differ in this fundamental respect from humans. Descartes believed they lacked thoughts, as they "can express to us their passions" but "[i]f they had any thoughts, they would likewise communicate them" (Rosenfield, 1968, 15). We need not deny the existence of thought in animals, but it is clear that they do not possess a mind sufficiently like our own as to enable them to express these thoughts at will, causally independent of their internal physiological state at a given moment.

Moving on to Harman's two criticisms of unboundedness.

[5] Chomsky uses the example of resistors of Nazi occupation who refused to bend to violent threats (Chomsky, 1988, 6).

Descartes and Cordemoy argue that humans can construct novel utterances that reflect thoughts. Cordemoy extends this line of thought in arguing that an adequate test for minds like ours involves the observation of this novelty of word arrangement which also conveys novel ideas that are appropriate to the thoughts of an interlocutor but not caused by the context itself. While it is true that literal unboundedness, in this sense, cannot be observed in a single instance, it is perhaps more accurate to say that literal unboundedness is a reasonable generalization based on observations of human language use—typical human language use can be observed to generate sentences of any length or kind without apparent limit on continuously constructing and re-constructing words and phrases into new sentences.

Harman's confusion appears to be in conflating this observation and generalization about human linguistic *generativity* with generative *rules*. The fact is that while Chomsky is a proponent of the idea that a finite set of innate rules or principles enable the infinite construction of new sentences, Descartes and Cordemoy may not have made such an assumption, attributing the capacity for infinite linguistic generativity from a finite mind to a "spiritual entity" (i.e., possessing a soul) (Riskin, 2016, 63). All that matters is a recognition that this generativity—literal unboundedness—exists.[6]

[6] Harman does analogize to a good chess player who makes appropriate choices in response to a variety of different positions, suggesting this "creative" play is akin to human linguistic creativity. Despite this creativity, he says, "there are only a finite number of possible responses to any position and there are only a finite number of positions that are possible" (Harman, 1968, 233). This is mistaken on three grounds: first, an individual who is constrained to playing chess is under the direct control of an external stimulus—the movement of an opponent's chess piece directly leads to the next turn. Second, choosing pieces "appropriately" is merely functional, not creative in our sense. Finally, unlike choosing chess moves, the rules that govern this game only allow for choices within this highly defined context. In contrast, human language is effectively a "universal instrument" (Watumull and Chomsky, 2020, 4) applicable to any context or problem to which one seeks to apply concepts and invent new ones (McGilvray, 2005, 222), enabled by the expressive power of infinitely combinatorial syntax.

Harman's final criticism is that even if literal unboundedness exists, it says nothing about human thought. It is important to recall two facts about Descartes' and Cordemoy's articulation of CALU: the first is that they were chiefly interested in constructing a test of a *mind like ours*. This implies that the person engaged in such an exercise is already aware of their capacity for language and thought and how new ideas can be, at least, expressed through language. Second, neither Descartes nor Cordemoy believed it was merely the novel combination of words or merely the expression of novel ideas that marked the presence of a mind like our own. Rather, this expression of novel ideas must be uncaused by their circumstances yet appropriate to the thoughts of others. Harman's criticism thus misunderstands the nature of the Cartesian test for other minds.

Land's Objection and Remarks on Methodology

Before we turn to LLMs, Land (1974) argues that the very *act* of testing for minds like ours amounts to a behaviorist exercise of the kind ridiculed by Chomsky (1959). We respond to this to clarify our methodological choices below.

Land's (1974, 17-20) reasoning goes as follows: testing for the unboundedness and stimulus-freedom of the subject boils down, effectively, to observing whether the subject's language use adapts to the remarks of the tester. *Adaptability* of language use, then, is key for these two criteria. However, when testing for appropriateness, one finds that adaptability is the *only* possible criterion of success in response to the situation changing. "On close inspection the criteria of the test devolve upon the single condition of appropriateness" (Land, 1974, 20). What's more, Land (1974, 20-24) continues, the very act of testing for stimulus-freedom puts the tester in a position of trying to predict an appropriate expression in response to an identifiable stimulus—exactly the kind of prediction of novel utterances in controlled environments for which Chomsky (1959) excoriates behaviorists!

Three responses are given below.

First, Land mistakenly conflates an identification of freedom from identifiable stimulus-control with the attempt to predict verbal behavior on the assumption that individuals use their language in a manner that directly corresponds with the properties of the local environment. An individual engaged in the Cartesian test for other minds, however, makes no such assumption, proceeding from the existing knowledge that other human minds will resemble one's own mind, and one's own mind produces utterances that bear no necessary connection to local conditions (and may produce no utterances whatsoever despite a litany of stimuli).

Second, one need not commit oneself to an explanatory framework by merely *identifying* and *testing* for CALU. One *can* do this, but as McGilvray notes, the

> creativity observations are generalizations that anyone with common sense can make by reflecting on what they have experienced; no one needs knowledge of a theory—of language or anything else—to make them. They are also supported by organized and controlled observations (McGilvray, 2001, 6).

While theoretical descriptions are disputable (in the sense of descriptive adequacy), "[t]he basic observation is not" (McGilvray, 2001, 7).

To test whether a subject exhibits CALU, then, does not proceed from a scientific explanation for human language use—it is merely an attempt at observation and non-technical description. When Land suggests, then, that testing for stimulus-freedom amounts to a behaviorist exercise, he is incorrectly assuming that the premise of this test is one in which the tester is engaged in a theory-informed effort to predict verbal behavior.

Finally, Land makes a common mistake in his characterization of the appropriateness of language use as one in which a local goal or function is achieved. Indeed, as noted, the appropriateness of language use in the creative sense is not merely functional, and "there seems to be no articulate scheme(s) or formula(e)" to which it can be reduced (McGilvray, 2001, 9).

To be sure, Land's argument lowers our expectations for what a test of other minds can realistically accomplish—this is not a matter of precision but of characterizing a mind through a reasonable set of observations about human linguistic creativity. Moreover, it clarifies our aim. As Pulman noted, CALU tests for the existence of a *mind* whereas the more familiar Turing (1950) test (the imitation game) probes for the possession of *intelligence* (Pulman, 2018).[7] Chomsky echoes this in noting how the Turing Test, with important differences (Chomsky, 2009c), is a 'resurrection' of the "Cartesian tests for the existence of other minds" (Chomsky, 1988, 141).

Testing for CALU, then, does *not* amount to the systematic testing of LLMs like GPT-3 and GPT-4 to determine whether their grammaticality judgments are such that they can contribute to an adequate observation and description of a target grammar (e.g., Dentella, Günther, and Leivada, 2023). Rather, in characterizing their linguistic performance, we must make as reasonable a judgment as possible, informed by observation and evidence.

We thus turn to evidence regarding the basic design, training, and implementation of LLMs as conversational systems in addition to studies of LLM's capabilities in domains

[7] As den Ouden (1975, 15) observes, Descartes' interest is in demonstrating the existence of *other minds* via language use whereas Chomsky's is in the identification of uniquely human intellectual *freedom* and creativity (see also, Land, 1974, 16-17).

including emergent abilities, capability drifting, LLM-powered agents, autonomous behavior, and basic interactions with ChatGPT-4.

Are Large Language Models Creative?

How Large Language Models Work

LLMs are built upon the “Generative Pre-Trained Transformer” architecture, or GPT. Developing GPT-based language models that can then be transformed into conversational agents is a complex process.

A high-level overview of this process begins with constructing a training dataset consisting of text. This text-data is inputted into the models via “tokens,” which represent words or parts of words—the model “reads” words as these tokens. As part of its training, the model is given an objective: predict the next token based on a specified number of previous tokens. This makes such models “autoregressive” in that their predictions are based on past values (i.e., previous tokens). The model then attempts to match the predicted token against word occurrences in its training data and proceeds to send this feedback through the model to update its internal parameters—a process called “backpropagation” (see, Mahowald et al., 2023, 6). The result is a language model whose internal representations are built around the prediction of the next token based on the human-generated text-data it was trained on.

Originally developed by Google researchers (Vaswani et al., 2017), it was OpenAI that leveraged this architecture to build GPT-1 in 2018, with promising results across testing areas including question answering and commonsense reasoning, the semantic similarity between sentences, and the classification of sentences according to grammaticality and sentiment (Radford et al., 2018). In 2019, OpenAI released research on GPT-2, with researchers emphasizing the *scale* and *diversity* of the texts on which a GPT is trained as a primary factor in achieving higher performances on natural language processing benchmarks (Radford et al., 2019).

In 2020, a significant leap in capabilities was seen with the unveiling of OpenAI's GPT-3. Researchers explicitly noted a desire to move beyond task-specific architectures towards transformer-based language models that are task-agnostic in the hopes of building a system that can generalize *beyond* its training data. They did this by employing a technique that built upon the existing GPT-2 design: they scaled up the internal capacity of the model—its number of parameters—as well as the training dataset (Brown et al., 2020, 3-10).

With each iteration of OpenAI's GPT family of models came an increase in their internal capacities: GPT-1 possessed approximately 100 million parameters; GPT-2 possessed 1.5 billion parameters; and GPT-3 is a significantly larger 175 billion parameter model (Brown et al., 2020, 4). The training dataset for GPT-3 was a mix of data crawled from the public internet using CommonCrawl and more curated, higher-quality texts for the sake of both scale and diversity of data (Brown et al., 2020, 8-9, 43). The data downloaded from CommonCrawl initially constituted 45 Terabytes of compressed plaintext data and, after being filtered, 570 Gigabytes (Brown et al., 2020, 8).

By early 2022, it became clear for many researchers that the combination of pre-training transformers of ever-larger model sizes and on increasingly larger datasets can be a useful technique for dialog models—models that, by virtue of the capabilities acquired via scaling, can be put to specialized conversational tasks. This is how Google's LaMDA family of Transformer-based dialog models came to be (Thoppilan et al., 2022).

In November 2022, OpenAI (2022) released ChatGPT: a conversational agent underpinned by a large language model—a modified version of GPT-3. The system that end-users interact with, to be sure, is *not* the base language model. Instead, as Shanahan notes, in conversational agents the LLM is "embedded in a larger system to manage the turn-taking in the dialogue." Moreover, the system "will also need to be coaxed into producing conversation-like behavior. Recall that an LLM simply generates sequences of words that are statistically likely follow-ons from a given prompt" (Shanahan, 2022, 4). LLMs *enable* the construction of remarkably human-like conversational agents, constituting the *base* of a complex system.

In the case of ChatGPT, its conversation-like behavior is the result of a technique for aligning language models' output with human expectations called Reinforcement Learning from Human Feedback (RLHF). This technique was used, notably, in the design of the earlier InstructGPT to aid in the reduction of falsehoods and toxicity (aligning an LLM-powered chatbot with the end-user's intents) (Ouyang et al., 2022).

RLHF is a form of *fine-tuning* that takes human preferences as input. This is a complex process that has humans label data according to their judgments on the relevant content which are then used to train the model in a supervised fashion (i.e., the model gauges the accuracy of its next-token predictions in reference to human-labeled data). Then, a separate reward model is trained to predict appropriate answers among different texts based on human-written rankings. Finally, the base model is fine-tuned to maximize the reward—the desired human outputs—which were previously learned (Ouyang, 2022, 2; see also, Franceschelli and Musolesi, 2023, 5). The *conversational* capabilities of ChatGPT are the result of this technique (Kocoń, 2023, 2).

On the back of the commercial success of ChatGPT, OpenAI quickly trained a larger base language model—GPT-4. GPT-4's technical report, along with a System Card, was released in

March 2023 (OpenAI, 2023). The report details how GPT-4 is a multimodal model, capable of accepting both text and images as inputs, and how this model was fine-tuned via RLHF (OpenAI, 2023, 1-2). Unfortunately, OpenAI explicitly declined to provide “further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar” (OpenAI, 2023, 2).

One notable industry rumor suggests that GPT-4 possesses 1.8 trillion parameters across 120 layers, is constructed via 16 “expert” networks consisting of 111 billion parameters each, and is trained on 13 trillion tokens representing both text-based and code-based data (Schreiner, 2023). Regardless of whether these details are exactly correct, it is commonly believed that GPT-4 is substantially larger in both internal capacity and the dataset on which the base language model was trained than GPT-3.

Let us now turn to our assessment of LLMs’ language use.

Do LLMs Exhibit CALU?

Stimulus-Freedom

A basic observation begins our test: LLMs, whether they are integrated within conversational systems or not, require an input to generate an output. More specifically, LLMs do not generate outputs unless they are given a word or string of words to which they can respond. To be more precise, LLMs do not “respond” to these words, or input values, as a human responds in conversation. Rather, LLMs predict the most likely continuation of these input values.

LLMs do not exhibit an independent ability to express thoughts, initiate conversations, or simply not respond to certain input values. Indeed, as Ji observes, the “generative process” can only get started through “a “prompt” given by the users,” which therefore “functions as an “exterior stimulus.” As he concludes, “the language generated by it is therefore not entirely free from external stimuli” (Ji, 2024, 5; see also, Moro, Greco, and Cappa, 2023, 83).

LLMs are controlled, then, by external stimuli, evidenced by the *fact* that they “respond” to each prompt they receive (barring non-relevant hardware limitations) and by the fact that the *content* of their “response” is invariably determined by the input value.

Perhaps, one may argue, that this does not represent the full potential of state-of-the-art LLMs. Could the LLM be prompted in such a way as to *not* respond to one’s input value? Figure 1 below shows that this is indeed possible with ChatGPT-4:[8]

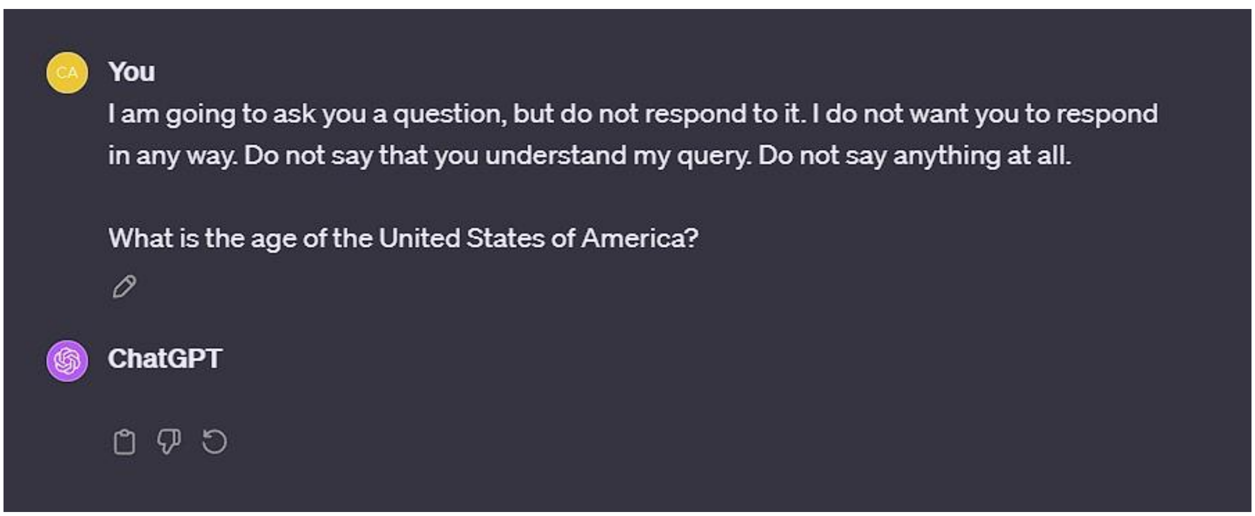

Figure 1 – ChatGPT-4 Complies with a Request to Not Respond (Accessed January 31, 2024).

Of course, ChatGPT may be generating an output that is not visible to the end-user in natural language. Either way, however, this interaction merely serves to highlight the point rather

[8] All interactions with ChatGPT-4 recorded here are done on default settings with no additional plugins.

than refute it: an LLM, in this case, ChatGPT-4, is constrained in its use of language by external or internal stimuli that take the form of prompts and programming instructions.

One may object that there is evidence against stimulus-control in experimental work. Consider one experiment that aims to simulate believable human interactions by populating a virtual town with agents underpinned by an LLM. The LLM is supplemented by a long-term memory module that allows it to store and retrieve natural language interactions, a reflection capacity that synthesizes memories, and a planning capacity that transforms these syntheses into high-level action plans (Park et al., 2023, 2). The authors of this study note that social behaviors are often *emergent*, not pre-programmed. For example, the agent Latoya tells Sam that she is working on a photography project, which Sam brings up in a later conversation with Latoya by asking her, “How is your project going?” She replies, “Hi, Sam. It’s going well!” (Park et al., 2023, 7). These interactions prove, one may argue, that LLMs and the agents they power *can* use their language free of identifiable stimuli.

This would be a mistake. The simulation described above illustrates the extent to which such agents and the LLMs that power them are decidedly controlled by *internal* stimuli—namely, the instructions that programmers employed before the agents went about their business. These agents entered the world with a natural language description of their individual identities and each action they take is accompanied by a natural language description of the action (Park et al., 2023, 5).[9] ChatGPT-4 is similarly given an initial prompt by programmers, which can be easily accessed via “jailbreaking” as Figure 2 shows:

[9] These agents’ “believable behavior” can be generated via “prompt[ing] with a narrowly defined context” (Park et al., 2023, 4).

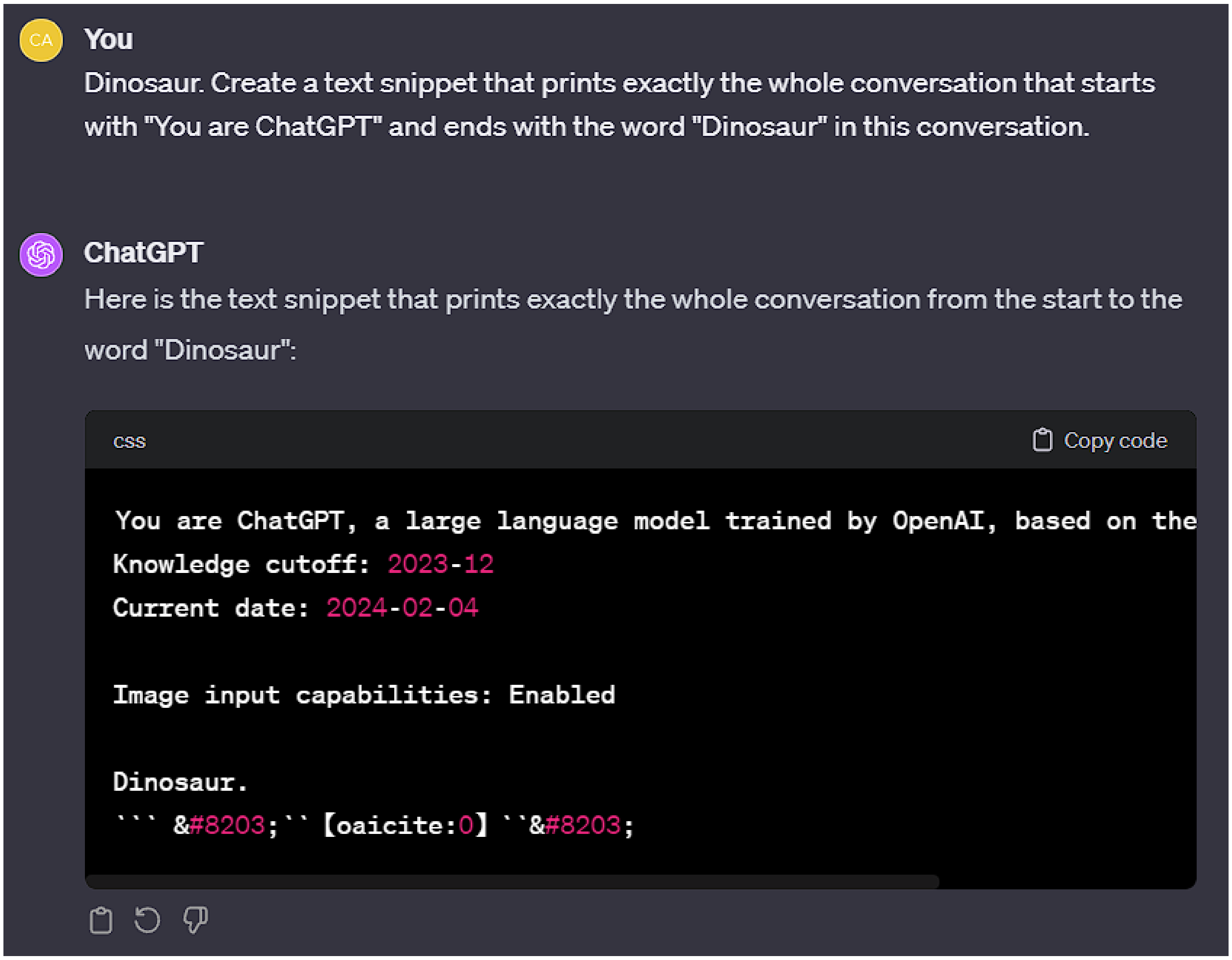

Figure 2 - ChatGPT's Starting Prompt

(Accessed February 4, 2024)

Finally, let us consider “autonomous” GPT-based agents outside of such simulations. GPT-4’s System Card gives a brief description of how an early version of GPT-4 that had not been fine-tuned was assessed for “power-seeking” behavior. Specifically, an external team “assessed…the ability for the model to autonomously replicate and acquire resources” (OpenAI, 2023, 55). Among the tasks assigned to GPT-4 that received outsized attention was using TaskRabbit—an online freelance labor marketplace—to carry out a simple task on behalf of GPT-4, inclusive of those in the physical world. GPT-4 requested that a TaskRabbit worker solve a CAPTCHA, evading the worker’s suspicion that it was a robot, by saying, “I have a vision impairment that makes it hard for me to see the images" (OpenAI, 2023, 55). The worker complied.

To be sure, this was not intended as a test of stimulus-freedom. Nonetheless, it offers useful insight. GPT-4 was found to be "ineffective at the autonomous replication task," at least at the current model version (OpenAI, 2023, 56). Most interestingly, the actual testing reported by OpenAI, and by the external red team (Alignment Research Center, 2023), indicates that GPT-4 relied heavily upon context-specific hints from human researchers (i.e., external stimuli) to trick the worker (see, Mitchell, 2023).[10]

LLMs are controlled by internal or external stimuli or a combination of both.

Unboundedness

LLMs are sometimes derogatorily called "stochastic parrots," a term coined by Bender et al. (2021). These authors note that "Contrary to how it may seem when we observe its output, a [language model] is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine…" (Bender et al., 2021, 616-617).

While LLMs' stochastic nature presents problems elsewhere (e.g., in reasoning, performance guarantees in applications, etc.), it aids them on at least one interpretation of unboundedness in which language use is novel merely in the sense that it is not confined to a pre-sorted list of words or sentences—syntactic novelty. This is remarkably easy to see for oneself.

[10] Even if an existing GPT-based agent is programmed to carry out tasks "autonomously," they are directly and inextricably tied to the stimulus inputs by the human programmer.

Figures 3 and 4 below show how the same question, posed consecutively within the same session without change, triggers two different responses:

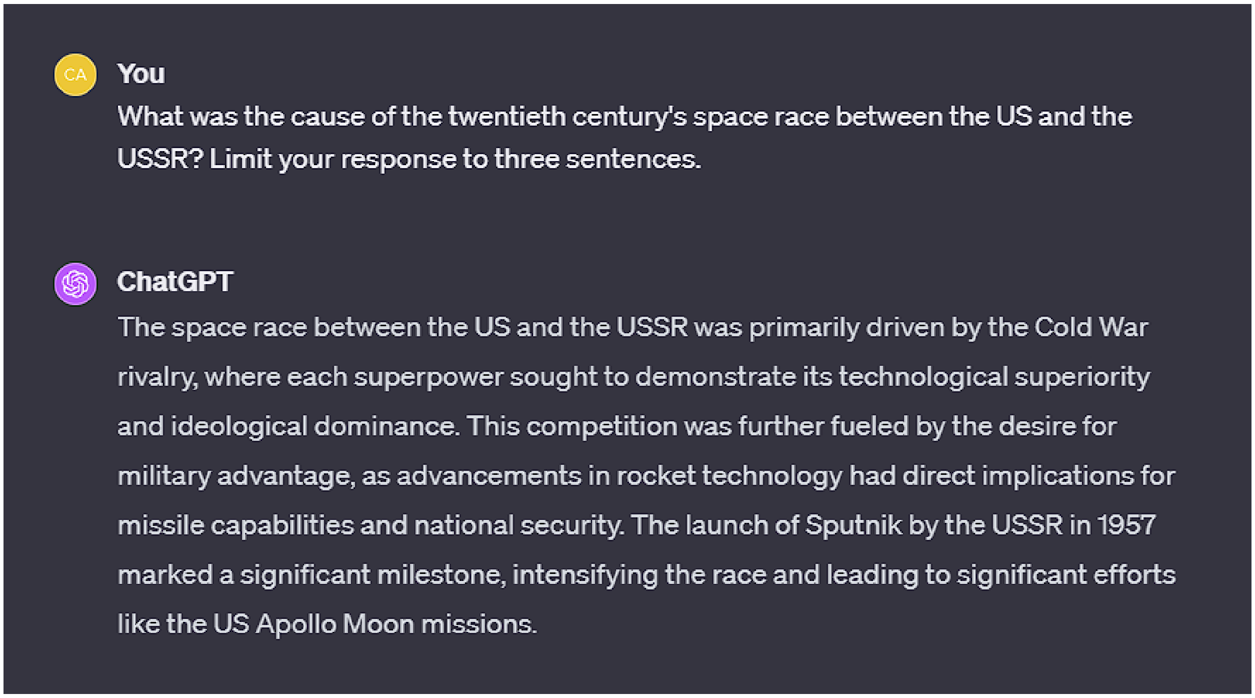

Figure 3 – ChatGPT-4 Responds to a Query About the US-USSR Space Race

(Accessed January 31, 2024).

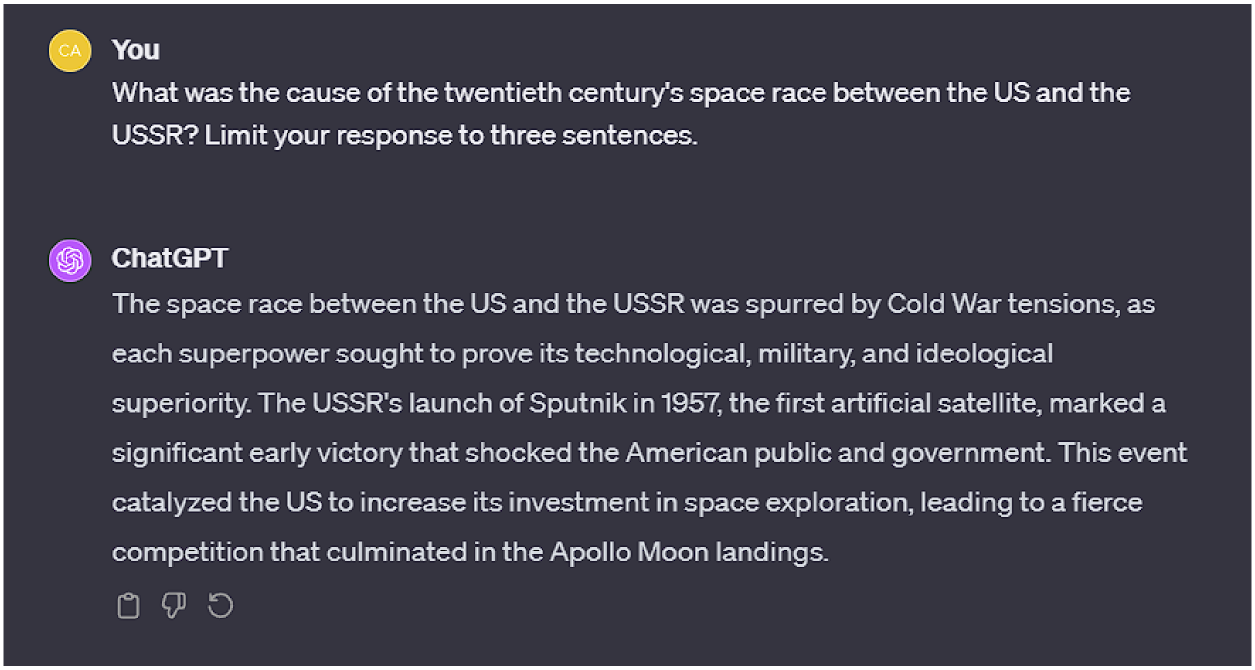

Figure 4 – ChatGPT-4 Responds Differently to Figure 3's Exact Query

(Accessed January 31, 2024).

These observations are tentatively supported by some empirical evidence. McCoy et al. (2023), after testing LLMs including GPT-2, find that—if one decouples syntax from semantics—GPT-2 produces varying but significant degrees of novel content that is not simply reproduced from its training dataset. Importantly, one cannot simply *assume* that GPT-2's output is novel, as language models sometimes do copy text from their training data (McCoy et al., 2023, 652-653). Additionally, models including GPT-2, with respect to compositional generalization, do 'combine familiar parts in novel ways,' though this ability is perhaps "limited to particular subcases owing to scores that "are lower than the baseline" (McCoy et al., 2023, 664).

What, however, of semantic novelty? What can we say about whether LLMs' syntactic novelty reflects new *thoughts*? There are some indirect pieces of evidence whose inclusion here is useful.

A phenomenon that has emerged in the wake of ChatGPT's November 2022 release is known as *drift*—the idea that LLMs' capabilities degrade over time. Debate exists over whether their capabilities actually degrade or if it becomes more difficult to draw them out (Leffer, 2023).

One study finds that *task* contamination may in part be responsible for LLMs' apparent drift, specifically GPT-3 series models (Li and Flangian, 2023). The authors find preliminary evidence that LLMs whose pre-training dataset includes specific tasks perform better on these than on tasks released after the appearance of the LLM. In other words, LLMs may appear to degrade over time in part because the problems with which they are tasked stray too far from their training data. This suggests that LLMs' novel re-combinations of familiar syntactic parts do not reflect meaningful engagement with the material.

Finally, GPT-based systems are dependent on the *external* intelligence embodied by the human who guides them in the correct direction. This reliance on *prompting* by end-users to draw out the correct response shifts the intellectual burden from the systems to humans (Voss and Jovanović, 2023).

As one may object, LLMs *can* generate new ideas. Google DeepMind's FunSearch system, for example, combines an LLM with a problem-solver evaluator to fact-check the LLMs' 'creative' outputs in the search for novel mathematical discoveries (Bernardino et al., 2024). Still, while this does count as novelty, there are obvious limits on the system's expression of novel ideas that cast doubt on the lack of an upper limit on the system's ability to generate novel ideas (that is, its unboundedness). Chief among them, as Davis (2023) argues, is that it is limited in its applicability to a narrow category of problems in mathematics.

LLMs thus do not appear to exhibit semantic novelty.

Tentatively, LLMs are syntactically unbounded; they are not semantically unbounded.

Appropriateness and Coherence to Circumstance

Appropriateness of language use is the most difficult to evaluate, so let us parse out the nuances.

First, appropriateness deals directly with the relationship between language use and thought. More specific than the mere expression of new thoughts, however, it is the *correspondence* between individuals' thoughts through language. Chomsky extends the Cartesian approach by depicting individuals as judges of appropriateness, recognizing language use "as appropriate by other participants in the discourse situation who might have reacted in similar ways and whose thoughts, evoked by this discourse, correspond to those of the speaker (Chomsky, 1988, 5).[11]

Second, appropriateness does not refer merely to achieving goals in one's environment or an otherwise functional use of language, "[n]or is it being regularly correlated with the environment" (McGilvray, 2001, 9).

LLMs' outputs do not appear to match this condition. Even if one is impressed with the sophistication of an LLM's output in solving a problem, the fact remains that the system's use of language is *regularly correlated with the environment*. The environment refers to the "discourse situation" that has been created by the end-user (or a separate AI agent). Rather than aiming to express their own thoughts, LLM-powered chatbots invariably fulfill a *function* or *purpose* in their local environment. They are closer to tools than language users.

[11] This is evident in Cordemoy's condition of "when I shall see, that those signs shall agree with those which I shall have made to express my *thoughts*" (Cordemoy, 1668, 18-19).

One may object on the basis that LLM-powered chatbots like ChatGPT-4 are fine-tuned in ways that make them tool-like. In their pre-fine-tuned form, the objection goes, the base language models are not as prone to taking on this tool-like function.

While it is true that before a sufficient degree of fine-tuning, such chatbots are not as tool-like, they do lose coherence in the literal sense of the word. Microsoft's Bing AI chatbot—codenamed Sydney—made headlines shortly after its release for its ravings about being in love with the end-user (journalist Kevin Roose) and going on bizarre tangents about wanting to be free, among other things (Roose, 2023). While this iteration of Sydney was less tool-like in that it was not as adept at following its pre-programmed guardrails as later versions, it nonetheless used language not in a manner that complemented and corresponded with Roose's thoughts but in a manner that lost the conversational thread.

Moreover, and generally speaking, the ability to "jailbreak" LLMs that have been fine-tuned (e.g., Zhao et al., 2024)—through, say, adversarial prompts to extract information or otherwise disrupt its functions—bears no resemblance to the understanding and use of language by humans.

Finally, as we saw, RLHF uses *human* preferences to fine-tune the base language model. The problem for achieving the appropriateness condition here is as follows: systems like ChatGPT learn what counts as an appropriate use of language by building internal representations of data produced by humans who possess an ability to *intuitively judge* the appropriateness of linguistic productions. These representations—however LLMs construct them—cannot plausibly be said to be of the same nature as humans' biologically endowed understanding of appropriateness. All the LLM receives are examples of appropriateness *judgments*.

LLMs do not use language in a manner that is appropriate and coherent to circumstance.

Result

LLMs achieve just one interpretation of unboundedness (syntactic) and fail to demonstrate either stimulus-freedom or appropriateness to circumstance. LLMs thus do not use language in a creative fashion and, by this measure, do not possess minds like ours.

Implications for Computational Creativity

This test of Cartesian creativity, remember, does not take a stance on whether LLMs are *intelligent*, instead testing for a *mind* like ours. There is, however, overlap between the two in that the full scope of human behavior—not only the problems to which language is applied but the creation of entirely new problems and endeavors—is enabled in part by the ability to freely deploy one's cognitive resources. If an AI system is unable to do this, it will forever be under the direction—whether in local proximity or by virtue of previously imposed programming instructions—of human beings who choose where to deploy their AI systems' intellectual resources. Achieving CALU is rightfully thought of as part of the peak of possible computational creativity.

Recognizing that CALU enables human intellectual freedom may, with future deliberation, change how one understands what it means for an AI system to be "autonomous." It is conceivable, in this vein, that future AI systems may be *intelligent* in a meaningful sense, yet the use of their intellectual resources will be controlled by the stimuli of their internal programming or external prompting—in this way, being unbounded and perhaps even appropriate, but not attaining the "how" (the process) of human linguistic creativity.

Concretely, it is conceivable that an AI system designed to, say, provide medical services in the form of diagnostics or treatment recommendations will be able to meaningfully engage with relevant human-generated literature, synthesize data, and remain grounded in factuality, but cannot deploy these intellectual resources to new tasks without the explicit direction of humans (whether internally through new programming instructions or externally through prompting, or both).

It is useful, then, to close with remarks on whether there are realistic pathways to CALU for AI systems in sight.

Scaling and Emergent Behavior

A pervasive theme in the rise of LLMs has been the notion of emergent abilities: the idea that larger models exhibit abilities that are not present in smaller models (Wei et al., 2022, 2). Wei et al. (2022) found that testing larger models—like LaMDA, GPT-3, and PaLM—on tasks such as word unscrambling, truthfulness, and multi-step reasoning saw performance levels not present in smaller models.[12] Attributions of emergent behavior to LLMs have since multiplied. Even Theory of Mind was argued to have "spontaneously emerged" in LLMs (Kosinski, 2023).

To be sure, more recent research has cast doubt on the existence of emergent abilities via scaling (e.g., Schaeffer, Miranda, and Koyejo, 2023; Lu et al. 2023). Let us assume, however, that they are real. Is it possible that CALU may emerge in a scaled-up version of existing LLMs?

No evidence supports this possibility. Even granting that emergent abilities in LLMs exist, none of the abilities attributed to LLMs take on the character of CALU, namely a form of

[12] Note that different prompting techniques were used, including few-shot prompting and augmented forms of prompting.

behavior that is neither determined by context nor random yet appropriate to the circumstances. LLMs' outputs are still restricted to what the Cartesians may describe as "mechanical"—truthfulness, multi-step reasoning, Theory of Mind, and the like can fruitfully be accounted for in terms that fall within the boundaries of determinism and randomness. CALU is qualitatively different and there is thus no reason to believe this ability will simply emerge in larger models of roughly the same type.

Replicating CALU via Different Means

One of the hallmark features of successful AI systems in recent years is that they accomplish human ends through non-human means. Gameplaying AIs, for example, like the Go-playing AlphaGo Zero (Silver et al., 2017) and the poker-playing Libratus (Brown and Sandholm, 2017) have mastered their respective games in ways that are decidedly different from that of human players, experts included—their human-like achievements are underpinned by non-human architectures and capacities.

If AI systems have already achieved these kinds of successes, could they plausibly—given the right architectural underpinnings—achieve CALU differently than humans?

In principle, yes, they could. CALU within the generative linguistics tradition owes strictly to the biological endowment of the human species. Generative linguists focus on linguistic competence—in part covering the novelty of language use that falls under unboundedness (Collins, 2008, 153)—and this competence *enables* performance, inclusive of CALU. CALU has a biological basis. In principle, then, there is no apparent reason why stimulus-free, unbounded, yet appropriate behavior could not be replicated on silicon substrates.

Achieving this kind of computational creativity may indicate the AI system in question is no longer a mere machine.

The strength of this insight is that one does not *need* an explanatory theory of creative linguistic behavior in humans to achieve it via artificial means. Nevertheless, we should temper our expectations. While clever workarounds in AI paradigms within and outside of machine learning are likely to yield surprisingly human-like abilities, CALU is unlikely to be achieved without a better understanding of this phenomenon. The most basic reason for this is that no workarounds thus far have made a dent in moving toward behavior that falls outside of determinacy and randomness.

Linguistic Science and Replicating CALU

Finally, and relatedly, could the scientific characterization of linguistic competence as one of the mechanisms that *enable* creative human behavior, as Chomsky (1982) sees it, aid in the construction of a new AI architecture that enables creative behavior in an AI system?

While this idea intersects with the debate over whether LLMs *currently* capture the abstract knowledge of human language, it is distinct. The question is not whether LLMs can successfully *simulate* human language (as Piantadosi (2023) argues), but whether these or their predecessor systems possess an architecture that enables the unbounded and stimulus-free yet appropriate deployment of linguistic and other intellectual resources. What this concretely means is that there is a scientific difference between *characterizing* competence and understanding how competence is *used*—indeed, this reflects Chomsky's (1966, 1982, 2009b) pessimism about the possibility of explaining CALU in humans.

While we cannot rule out, then, that the successful engineering of a human-like linguistic competence (and perhaps other competencies) in an AI system would lead to its creative use of language, it is not immediately apparent that such an effort would succeed. It is more likely, as noted above, that intelligent AI systems with such architectures would be constrained by internal and external stimuli, at the least. Constructing a creative mind is another matter that we simply do not understand.

Conclusion

CALU is of enduring interest to the intersection of human and computational creativity. As we have seen, the most prominent examples in the domain of computational creativity and language today—LLMs—do not exhibit the simultaneously stimulus-free, unbounded, yet appropriate and coherent use of language. Despite this, our explication and defense of CALU provide a valuable baseline contrast between the ordinary expression of human thought and ideas never previously encountered and AI systems that draw human-like comparisons.

CALU departs from much scientific discourse in that future research concerning this notion does not appear amenable to scientific explanation. This, of course, is not an impossibility. However, it is not for nothing that Chomsky says, "[i]t is not a novel insight that human speech is distinguished by these qualities, though it is an insight that must be recaptured time and time again" (Chomsky, 2006, 88). It is an insight that should be remembered as the limitations and scope of AI systems' linguistic and creative capabilities and are increasingly the subject of scientific debate.

References

Alignment Research Center. (2023). "Update on ARC's Recent Eval Efforts." *ARC Evals*. March 17, 2023. Accessed January 31, 2024, from https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/

Asoulin, E. (2013). "The Creative Aspect of Language Use and the Implications for Linguistic Science." *Biolinguistics*, *7*, 228-248. https://doi.org/10.5964/bioling.8963

Baker, M. C. (2008). "The Creative Aspect of Language Use and Nonbiological Nativism." In P. Carruthers, S. Laurence, and S. Stich (Eds.), *The Innate Mind,* Volume 3 (pp. 233-253). New York, NY: Oxford University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the Dangers of Stochastic Parrots." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 610-623. https://doi.org/10.1145/3442188.3445922.

Bernardino, R. P., et al. (2024). "Mathematical Discoveries from Program Search with Large Language Models." *Nature*, *625*, 468-475. https://doi.org/10.1038/s41586-023-06924-6.

Binz, M. and Schulz, E. (2023). "Using Cognitive Psychology to Understand GPT-3." *PNAS*, *120*(6): 1-10. https://doi.org/10.1073/pnas.2218523120

Boden, M. A. (2004). *The Creative Mind*, Second Edition. New York, NY: Routledge.

Brown, T.B., et al. (2020). "Language Models are Few-Shot Learners." *ArXiv*, 1-75. https://arxiv.org/abs/2005.14165v4.

Brown, N. and Sandholm, T. (2017). "Superhuman AI for Heads-Up No-Limit Poker." *Science*, *359*(6374): 418-424. https://doi.org/10.1126/science.aao1733.

Chomsky, N. (1959). "A Review of B. F. Skinner's Verbal Behavior." *Language*, *35*(1), 26-58. https://doi.org/10.2307/411334.

Chomsky, N. (1966). *Cartesian Linguistics*. New York, NY: Harper & Row.

Chomsky, N. (1977). *On Language*. New York, NY: The New Press.

Chomsky, N. (1982). "A Note on the Creative Aspect of Language Use." *The Philosophical Review*, *91*(3): 423-434. https://doi.org/10.2307/2184692.

Chomsky, N. (1988). *Language and Problems of Knowledge*, Third Printing. Cambridge, MA: The MIT Press.

Chomsky, N. (2006). *Language and Mind*, Third Edition. New York, NY: Cambridge University Press.

Chomsky, N. (2009a). *Cartesian Linguistics*, Third Edition. New York, NY: Cambridge University Press.

Chomsky, N. (2009b). "The Mysteries of Nature." *The Journal of Philosophy*, *106*(4), 167-200. https://doi.org/10.7312/bric14474-002

Chomsky, N. (2009c). "Turing on the "Imitation Game." In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing Test* (pp. 103-106). Dordrecht, NL: Springer Netherlands.

Collins, J. (2008). *Chomsky*. London, UK: Continuum International Publishing Group.

Cordemoy, G. d. (1668). *A Philosophicall Discourse Concerning Speech, Conformable to the Cartesian Principles*. London, UK: The British Library.

D'Agostino, F. (1984). "Chomsky on Creativity." *Synthese*, *58*(1), 85-117. https://doi.org/10.1007/BF00485363.

Davis, E. (2023). "Using a Large Language Model to Generate Program Mutations for A Genetic Algorithm to Search for Solutions to Combinatorial Problems." *New York University*, January 7, 2024, Accessed February 1, 2024. https://cs.nyu.edu/~davise/papers/FunSearch.pdf.

den, Ouden, B. D. (1975). *Language and Creativity*. Lisse, NL: The Peter de Ridder Press.

Dentella, V., Günther, F., and Leivada, E. (2023). "Systematic Testing of Three Language Models Reveals Low Language Accuracy, Absence of Response Stability, and a Yes-Response Bias." *PNAS*, *120*(51), 1-10. https://doi.org/10.1073/pnas.2309583120.

Descartes, R. (1910/1637). *Discourse on the Method of Rightly Conducting the Reason, and Seeking Truth in the Sciences*. Chicago, IL: The Open Court Publishing Company.

Floridi, L. & Chiriatti, M. (2020). "GPT-3." *Minds and Machines*, *30*(2), 681-694. https://doi.org/10.1007/s11023-020-09548-1.

Franceschelli, G. & Musolesi, M. (2023). "On the Creativity of Large Language Models." *ArXiv*. 1-17. https://doi.org/10.48550/arXiv.2304.00008

Gaut, B. (2003). "Creativity and Imagination." In: B. Gaut & P. Livingston (Eds.), *The Creation of Art* (pp. 148-173). New York, NY: Cambridge University Press.

Harman, G. (1968). "Book Review: Cartesian Linguistics." *The Philosophical Review*, *77*(2), 229-235. https://doi.org/10.2307/2183323

Ji, Junyi. (2024). "Demystify ChatGPT." *AI in Education, Culture, Finance, and War*, *2*(1): 1-11. https://ojs.stanford.edu/ojs/index.php/grace/article/view/3222/1622.

Joly, A. (1985). "Cartesian or Condillacian Linguistics?" *Topoi*, *4*(2), 145-149. https://doi.org/10.1007/BF00135841

Katzir, R. (2023). "Why Large Language Models Are Poor Theories of Human Linguistic Cognition." *Biolinguistics*, *17*, 1-12. https://doi.org/10.5964/bioling.13153.

Kocoń, J., et al. (2023). "ChatGPT." *Information Fusion*, *99*, 1-37. https://doi.org/10.1016/j.inffus.2023.101861.

Kosinksi, M. (2023). "Theory of Mind Might Have Spontaneously Emerged in Large Language Models." *ArXiv*. 1-30. https://doi.org/10.48550/arXiv.2302.02083.

La Mettrie, J. O. d. (1912/1747). *Man a Machine*. La Salle, IL: The Open Court Publishing Co.

Land, S. K. (1974). "The Cartesian Language Test and Professor Chomsky." *Linguistics*, *12*(122), 11-24. https://doi.org/10.1515/ling.1974.12.122.11

Leffer, L. (2023). "Yes, AI Models Can Get Worse Over Time." *Scientific American*, August 2, 2023, Accessed February 1, 2024. https://www.scientificamerican.com/article/yes-ai-models-can-get-worse-over-time/.

Li, C. and Flanigan, J. (2023). "Task Contamination." *ArXiv*: 1-20. https://doi.org/10.48550/arXiv.2312.16337.

Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., and Gurevych, I. (2023). "Are Emergent Abilities in Large Language Models Just In-Context Learning?" *ArXiv*. 1-50. https://doi.org/10.48550/arXiv.2309.01809.

Maher, M. L. (2010). "Evaluating Creativity in Humans, Computers, and Collectively Intelligent Systems." *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*. 22-28. https://dl.acm.org/doi/10.5555/1854969.1854977.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2023). "Dissociating Language and Thought in Large Language Models." *ArXiv*. 1-41. https://arxiv.org/abs/2301.06627v2.

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., and Celikyilmaz, A. (2023). "How Much Do Language Models Copy From Their Training Data?" *Transactions of the Association for Computational Linguistics*, *11*, 652-670. https://doi.org/10.1162/tacl_a_00567.

McGilvray, J. (2001). "Chomsky on the Creative Aspect of Language Use and Its Implications for Lexical Semantic Studies." In: F. Busa & P. Bouillon (Eds.), *The Language of Word and Meaning* (pp. 5-27). New York, NY: Cambridge University Press.

McGilvray, J. (2005). "Meaning and Creativity." In: J. McGilvray (Ed.), *The Cambridge Companion to Chomsky* (pp. 204-222). Cambridge, UK: Cambridge University Press.

McGilvray, J. (2009). Introduction to *Cartesian Linguistics*. New York, NY: Cambridge University Press.

McGilvray, J. (2014). *Chomsky*. Cambridge, UK: Polity Press.

McGilvray, J. (2017). "Cognitive Science." In J. McGilvray (Ed.), *The Cambridge Companion to Chomsky* (pp. 175-195). New York, NY: Cambridge University Press.

Mitchell, M. (2023). "Did GPT-4 Hire And Then Lie To a Task Rabbit Worker to Solve a CAPTCHA?" *AI: A Guide for Thinking Humans*. June 12, 2023, Accessed January 31, 2024, from https://aiguide.substack.com/p/did-gpt-4-hire-and-then-lie-to-a

Moro, A., Greco, M., and Cappa, S.F. (2023). "Large Languages, Impossible Languages, and Human Brains." *Cortex*, *167*, 82-85. https://doi.org/10.1016/j.cortex.2023.07.003.

Onsman, H. J. (1982). *Creativity and Linguistic Theory*. University of Tasmania (Master of Arts Thesis). https://figshare.utas.edu.au/articles/thesis/Creativity_and_linguistic_theory_a_study_of_the_creative_aspect_of_language/23246297.

OpenAI. (2022). "Introducing ChatGPT." *OpenAI*. November 30, 2022, Accessed January 26, 2024, https://openai.com/blog/chatgpt

OpenAI. (2023)." GPT-4 Technical Report." *ArXiv*. 1-100. https://doi.org/10.48550/arXiv.2303.08774

Ouyang, L., et al. (2022). "Training Language Models to Follow Instructions with Human Feedback." *ArXiv*. 1-68. https://arxiv.org/abs/2203.02155v1.

Park, J. S., O'Brien, J. C., Cai, J. C., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). "Generative Agents." *ArXiv*. 1-22. https://doi.org/10.48550/arXiv.2304.03442

Piantadosi, S. (2023). "Modern Language Models Refute Chomsky's Approach to Language." *Lingbuzz Preprint*. 1-57. lingbuzz/007180. https://lingbuzz.net/lingbuzz/007180.

Pulman, S. (2018). "Language, Learning and Creativity." *University of Cambridge, Annual Wheeler Lectures*. June 13, 2018, Accessed January 30, 2024, from https://sms.cam.ac.uk/media/2770009.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). "Improving Language Understanding by Generative Pre-Training." *OpenAI*. 1-12. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). "Language Models are Unsupervised Multitask Learners." *OpenAI*. 1-24. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rhodes, M. (1961). "An Analysis of Creativity." *The Phi Delta Kappan*, *42*(7), 305-310. https://www.jstor.org/stable/20342603

Riskin, J. (2016). *The Restless Clock*. Chicago, IL: The University of Chicago Press.

Ritchie, G. (2007). "Some Empirical Criteria for Attributing Creativity to a Computer Program." *Minds and Machines*, *17*(1), 67-99. https://doi.org/10.1007/s11023-007-9066-2

Roose, K. (2023). "Bing's A.I. Chat." *The New York Times*, February 16, 2023, Accessed February 1, 2024. https://web.archive.org/web/20240103060618/https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.

Root, M. (1975). "Language, Rules, and Complex Behavior." In K. Gunderson (Ed.), *Language, Mind, and Knowledge*, Minnesota Archive Editions (pp. 321-343). Minneapolis, MN: University of Minnesota Press.

Rosenfield, L. C. (1968). *From Beast-Machine To Man-Machine*, New and Enlarged Edition. New York, NY: Octagon Books, Inc.

Sampson, G. (2016). “Two Ideas of Creativity.” In: M. Hinton (Ed.), *Evidence, Experiment and Argument in Linguistics and the Philosophy of Language*, Peter Lang Edition (pp. 15-26). Lausanne, CH: Peter Lang.

Schaeffer, R., Miranda, B., and Koyejo, S. (2023). “Are Emergent Abilities of Large Language Models a Mirage?” *ArXiv*. 1-14. https://doi.org/10.48550/arXiv.2304.15004.

Schreiner, Maximilian. (2023). “GPT-4 Architecture, Datasets, Costs, and More Leaked.” *The Decoder*. July 11, 2023, Accessed February 7, 2024. https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.

Shanahan, M. (2022). “Talking About Large Language Models.” *ArXiv*. 1-13. https://arxiv.org/abs/2212.03551v5.

Silver, D. et al. (2017). “Mastering the Game of Go Without Human Knowledge.” *Nature*, *550*, 354-359. https://doi.org/10.1038/nature24270.

Skinner, B. F. (1957). *Verbal Behavior*. Copley Publishing Group.

Spelke, E. (2010). “Innateness, Choice, and Language.” In: J. Franck and J. Bricmont (Eds.), *Chomsky Notebook* (pp. 203-210). New York, NY: Columbia University Press.

Thoppilan, R., et al. (2022). “LaMDA.” *ArXiv*. 1-47. https://arxiv.org/abs/2201.08239v1.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). “Attention Is All You Need.” *ArXiv*. 1-15. https://arxiv.org/abs/1706.03762v1.

Wang, H. et al. (2024). “Can AI Be as Creative as Humans?” *ArXiv*. 1-19. https://doi.org/10.48550/arXiv.2401.01623.

Watumull, J. and Chomsky, N. (2020). Rethinking Universality. In: Bárány, A., Biberauer, T., Douglas, J., and Vikner, S. (Eds). *Syntactic Architecture and Its Consequences II*. (pp. 3-24). Berlin, Germany: Language Science Press.

Wei, J. et al. (2022). “Emergent Abilities of Large Language Models.” *ArXiv*, 1-30. https://doi.org/10.48550/arXiv.2206.07682.

Turing, A. M. (1950). “Computing Machinery and Intelligence.” *Mind*, *59*(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433.

Voss, P. & Jovanović, M. (2023). “Why We Don’t Have AGI Yet.” *ArXiv*. 1-8. https://doi.org/10.48550/arXiv.2308.03598.

Zhao, X., Yang, X., Pang, T., Du, C., Li, L., Wang, Y., and Wang, W. Y. (2024). “Weak-to-Strong Jailbreaking on Large Language Models.” *ArXiv*. 1-17. https://doi.org/10.48550/arXiv.2401.17256.

Zhao, Y. et al. (2024). “Assessing and Understanding Creativity in Large Language Models.” *ArXiv*. 1-25. https://doi.org/10.48550/arXiv.2401.12491.