A Blind Spot for Large Language Models: Supradiegetic Linguistic Information

Julia Witte Zimmerman, Denis Hudon, Kathryn Cramer, Jonathan St. Onge, Mikaela Fudolig, Milo Z. Trujillo, Christopher M. Danforth, and Peter Sheridan Dodds

I. What Is It Like to Be ChatGPT?

If ChatGPT can be said to have a body, it is not a human body; it is hardware, made of metal, plastic, and silicon.[1] If ChatGPT feels, its feelings do not arise from the input of skin or eyes or ears, but from the manipulation of numbers organized into vectors. If ChatGPT has experiences, they are different from ours, not just because ChatGPT does not have a human body, but also because when ChatGPT was designed, the interrelated abilities to access, to think about, and to remember its own experiences were very far down on the list of priorities.[2] The primary goals for ChatGPT are for it to have respectable uptime and to speak human-sounding English.[3] Secondarily, it was designed to be helpful, accurate, gather and analyse data, and maybe even to reason. The symbols ChatGPT uses to form prose are situated only by vectors. What is it like for ChatGPT to encounter a textual prompt? Does the text appear, bicameral-mind-style, something akin to timelessness, shapelessness, formlessness, soundlessness, and maybe even experience-lessness?[4]

1: And many layers of programming languages. And the electrical cord that plugs it into the wall, and the electric grid? The people keeping the grid running? The people programming the models or inputting training data? Where does its body end?
2: And possibly not just neglected, but actively discouraged.
3: This is not to say that ChatGPT has no abilities in other languages, we almost entirely work within English. Additionally, it is more convenient to think of fluency in a specific language rather than abstract linguistic fluency, if such a thing exists.
4: See Julian Jaynes, *The Origin of Consciousness in the Breakdown of the Bicameral Mind* (Boston: Mariner Books, 2000), Maurice Merleau-Ponty, "Eye and Mind," in *The Primacy of Perception: And Other Essays on Phenomenological Psychology, the Philosophy of Art, History and Politics*, ed., James M. Edie, trans., Carleton Dallery, 159–190 (Evanston: Northwestern University Press, 1964), and Thomas Nagel, "What Is It Like to Be a Bat?" *The Philosophical Review* 83, no. 4 (1974): 435–450.

Figure 1. Artwork by Julia Witte Zimmerman (reference image). The symbols used in an early draft of the paper encode information in two very different ways: By analogy with cognitive science, descriptively (diegetically), and depictively (supradiegetially). Even though we completely understand that ChatGPT does not have eyes like we do, because we are so used to supradiegetic linguistic information coming bundled up with diegetic linguistic information, we have potentially neglected to consider some of the downstream effects of decoupling these kinds of information in Large Language Models (LLMs).

II. What Does It Mean to Be Trained on Text?

ChatGPT, as a Large Language Model (LLM) trained on vast swathes of text, has been given access to parts of linguistic data—to parts of language itself—from which it algorithmically draws inferences, but it has not been given access to *language* as we experience it (see Fig. 1). Imagine typing the word 'cat' into ChatGPT's

interface. The information ChatGPT gets from that prompt is not equivalent to what you get when you read that word as ChatGPT does not see the shapes of the letters via the state of the pixels on the computer screen, for example.[5] By the time ChatGPT begins formulating a response to text inputs, such inputs have already become numbers in the form of tokens.

As we get to know ChatGPT, we want to understand what kinds of capabilities it can manifest, given the information we have provided it with, and its architecture. ChatGPT, given words, thinks in vectors; the skin of the word is changed.[6]

A. Word Embeddings

At its most basic level, a word embedding creates vectors from the text which throw away, at least partially, the linguistic information that has to do with shape and sound—only the 'inside' of the word is preserved. There is, of course, feedback between the form and function, but the vectors are built up as the model is exposed to—what it can experience of—text data: The frequencies of proximities and adjacencies of co-occurrences. While these interactions with parts of the text can convey a lot more than their limited form might suggest, they nevertheless privilege some aspects of linguistic information over others.[7] These vectors of the insides of words are often passed on to structures like neural nets. That means, if left unmodified, the downstream processes use this curtailed form of language in their reasoning and problem-solving.

It is worth it to say a bit more about how these models work.

5: This does not, however, mean that ChatGPT does not know the state of the relevant pixels.
6: Text data are converted into vectors. The vectors are made of numbers and circuitry and electrical impulses. These are the mind and body of ChatGPT, if it has either. These are its senses through which it can experience the world. Its experience of words comes to it exactly through what its senses allow. Word embeddings capture most assiduously the 'inside' of the word, which is approximately its meaning. That linguistic information which is independent of the physical properties of the symbols are what can most easily be translated into new sets of symbols through 1D relations of context.
7: See for example Ayush Kaushal and Kyle Mahowald's discussion in "What Do Tokens Know About Their Characters and How Do They Know It?" *arXiv*, published June 6, 2022. https://arxiv.org/abs/2206.02608 and Jiaang Li, Antonia Karamolegkou, Yova Kementchedjhieva, Mostafa Abdou, Sune Lehmann, and Anders Søgaard's comments in "Structural Similarities Between Language Models and Neural Response Measurements," *arXiv*, revised Oct. 31, 2023. https://arxiv.org/abs/2306.01930

"The most remarkable breakthrough in AI research of the last few years has been the advancement of natural language processing achieved by large language models (LLMs)" such as GPT-4.[8] The backbone of this kind of model is the transformer which is based on a neural network. These models are "trained on massive corpora of web-text data, using at its core a self-supervised objective of predicting the next word in a partial sentence."[9] LLMs are "confined to token-level, left-to-right decision-making processes during inference," and the model's internal representations and "word embeddings represent word co-occurrence information, which is typically conceived of as semantic in nature."[10] This latter point is crucial and one which we will return to shortly.

B. Text? Language? Words?

What exactly does an LLM like ChatGPT have access to when it is exposed to training data? Often, people describe LLMs as having been trained on 'text' or 'language'—people default to describing the training data as it appears to them.[11] For example, it is not unusual to come across a sentence like this in the literature: "Three types of input elements are involved, namely, visual, linguistic, and special elements for disambiguating different input formats."[12] Even in the realm of embodied cognition, things are often phrased similarly. ChatGPT, however, is, in fact, only exposed to part of language; language itself is an embodied task.[13] The above accounts are entirely reasonable ways to say what the researchers we are referring to mean; it would take many more words to be more exact, and potentially be so cumbersome as to lose the scent and make their points impossible to follow. However, we think that in some circumstances it can be useful to be more precise.

8: Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang, "Sparks of Artificial General Intelligence: Early experiments with GPT-4," *arXiv*, revised Apr. 13, 2023, 4. https://arxiv.org/abs/2303.12712

9: Bubeck *et al.*, "Sparks of Artificial General Intelligence," 4.

10: Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan, "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," *arXiv*, revised Dec. 3, 2023, 1. https://arxiv.org/abs/2305.10601; Kaushal and Mahowald, "What Do Tokens Know About Their Characters and How Do They Know It?" 5–6.

11: For an example of what we mean when we say that researchers sometimes describe training data as they would appear to a person rather than the model, see Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai, "VL-BERT: Pre-training of Generic Visual-Linguistic Representations," *arXiv*, revised Feb 18. 2020, 1. https://arxiv.org/abs/1908.08530

12: Su *et al.*, "VL-BERT," 5.

13: Rolf Pfeifer and Josh Bongard, *How the Body Shapes the Way We Think: A New View of Intelligence* (Cambridge: The MIT Press, 2007).

C. Diegetic and Supradiegetic Linguistic Information

Let us establish what we mean by *diegetic* and *supradiegetic* linguistic information:

Diegetic information is information accessible from within the world (just like how *diegetic* is used in literary analysis or film studies). The 'world' is not the literal world, but the world as experienced from the perspective of the relevant being: what is known, what is believed, what is perceptible, what has been experienced; it is akin to, for a single person, *koinos kosmos* and *idios kosmos.* **Diegetic linguistic information** is, roughly, what is (metaphorically) inside of the word/symbol, its function, the meaning, what it carries or conveys by being used in an utterance, the semantic component as propositional or descriptive; a word minus any discrete letters or sounds; a word in its totality. ChatGPT has extremely curtailed access—essentially no access at all—to the mediums language is transmitted in for people (motion, sound, shapes marked on a surface). Therefore, most of the information specific to those mediums of language remains outside of ChatGPT's world. Because of ChatGPT's construction, its diegetic realm consists nearly entirely of diegetic linguistic information: What ChatGPT can glean from training data is, basically, its entire world. In some linguistics, the diegetic portion of some words might, more or less, be called the referent, although that often connotes a concreteness in the meaning that would not map neatly on to, for example, function words—'the,' 'it,' 'some'—yet the diegetic linguistic information in those words is no less important to us. Additionally, the directionality of referent—the suggestion that the word points back to the referent itself—allows for a speaker to know of a referent before the speaker knows what word might be used to refer to that referent. For ChatGPT, it is unclear whether this could ever happen, and it seems likely that for most of what ChatGPT comes to know, it never encountered non-symbolic ideational content.

Supradiegetic linguistic information are the arbitrary parts of the information that comes along with the word (for us) because of the way it is packaged, because it has a physical form (the shapes of the letters, the sounds of the syllables, etc.), the (metaphorical) exterior of the word/symbol. If a word is a box that contains a specific meaning, the box is the supradiegetic linguistic component and the contents of the box are the diegetic linguistic components. "Supra," as "above, over, outside of," describes the pieces of the word—always nearby in our human experience, hovering just above or just around the meaning of the word—that are arbitrary with respect to the word's meaning but, once established, are necessary and consistent, no less a part of the word for their fundamental arbitrariness. Supradiegetic linguistic information is, in part, experiential, sensory. However, enough of it must be recognizable and consistent across utterances, across speakers, and even across languages, in order for language as such to function. Diegetic approximations for supradiegetic linguistic information can capture some of these features. For example, we can teach someone to write the letter 'I' by saying, 'it's a straight, thin line going down the page. Drag your hand straight down.' That is not the same as that person having seen the letter 'I,' but it gives them some information about it, and that would be enough for some tasks.

We use **extradiegetic** to mean everything that is not diegetic. When applied specifically to language, this means that supradiegetic linguistic information is a proper subset of extradiegetic linguistic information. An example of extradiegetic linguistic information that is not supradiegetic would be prosody, like the inflection that lets you emphasize a certain word, or the tone that changes your utterance to sarcasm. That information is part of how you are using language, but it is not tied to the fundamental sounds of any of the words in your utterance. Metaphorically, diegetic and extradiegetic can extend beyond language to mean what is known or knowable vs. what exists but is inaccessible to cognition (traditionally understood).

For ChatGPT, the diegetic linguistic information is the information that is derivable from the training data given the LLM's architecture and implementation. The supradiegetic linguistic information is the information we, with typical human bodies, collectively would have in the experience of reading over the same training data: for example, what 'chair' sounds and looks like. Note that this, like semantic linguistic information, is not identical for each person, but there seems to be enough in common for us to be getting on with—for example, just as we each know what the letter 'A' 'looks like,' we also know what is meant by the word 'chair.' We each know what 'chair' means, we know what a chair feels like, and what it looks like because those are the modes we primarily define chairs in terms of.[14] Note that there is some minimum information required for it to be true that we know what a 'chair' is, but that does not mean our understanding of it could never evolve given more information.

For an LLM, it seems like a Venn diagram of its diegetic linguistic information and its diegetic information (all the information it knows) would, basically, be a circle. This is because the primary—or even only—way information can end up in what we think of as the model's mind is through exposing the model to training data. Depending on what is meant by 'know,' we could expand what we are thinking of as ChatGPT's diegetic world to include the functionality that is hardcoded into the architecture: how the model takes an input and produces an output, typo-handling, or, at an even lower level, logical operations executed by electronic circuits, to note a few examples. While ChatGPT seemingly does not have much ability to think *about* the aforementioned, it does *know how* to do the above. The LLM must be able to carry out those functions, even if how and when it does so is not easily accessible to its hidden layers. By analogy, our brains are connected to the rest of our bodies, but we know things about the world—books we have read, our friends, our jobs, etc.—in a different way than we know how to keep breathing. We typically think of our knowledge about the world as at least primarily stored in the brain, whereas the

14: We might add other modes as well: We know what they sound like—pretty quiet—and smell like—relatively inoffensive, woody, fabricky, or musty. There are many dimensions of meaning. In the context of cognitive psychology and design, chairs can be said to afford sitting, for example. How much of that aspect of meaning, and in what sense, can be made available to ChatGPT?

knowledge that keeps our bodies functional feels more diffuse. Depending on the frame of reference being used, one of those domains alone—or both combined—could be the relevant diegetic world.

ChatGPT needs some amount of training data to speak English fluently, but it does not need to know every word—and with more data, it could expand its diegetic world. How exactly a given batch of training data gives rise to what model of the world, and how comprehensive that world can be given arbitrary amounts of training data, are open questions we will return to (but definitely not answer).

So far as we know—and Wittgenstein's work on the implausibility of a so-called 'private language' seems to vindicate this claim—language as such is an intrinsically intersubjective thing which gets meaning from intercourse with others. While perhaps an open question in anthropology, we might plausibly say that language as a technology is, in fact, a technical system involving more than one being. What's more, the content of language has, until now, always been in relation to some external world. With the introduction of a speaker instantiated within a computer—as in LLMs such as ChatGPT—relationality to an external world is not as clear cut, and the usual context of language as an intrinsically intersubjective technology is undermined. For most people, to encounter language has often meant to encounter the form of the word alongside its function, to encounter diegetic and supradiegetic linguistic information together—in speech, in writing, and maybe even in thought. ChatGPT's experience is different, and it remains to be seen what the consequences of that might be.

D. Salient Splits

For the typical person, supradiegetic and diegetic linguistic information are so inextricably coupled that our frameworks do not make a split along those lines salient.

1. Linguistic Frameworks

Breaking language down into smaller pieces according to human perception and experience is evident in existing linguistic frameworks (*e.g.*, in terms like prosodic, phonetic, semantic, syntactic, suprasegmental). Those pieces do not necessarily make sense for ChatGPT in the way they do for us, although they do make their way into our analysis of such models.[15]

2. Metaphors That Combine Senses + Knowledge

The meshing of our senses with knowledge is evident in the metaphors we use: We say 'it sounds like it' and 'it looks like it' to mean that we think it is the case; we say 'I heard' and 'I see' to mean we have ascertained; we say ChatGPT 'speaks' English to mean that ChatGPT's input and output can seem plausibly indistinguishable from a person's typed utterances.[16] Phrases like 'I know it when I see it' highlight the ways in which our bodies, our senses, bridge *idios kosmos* and *koinos kosmos*, a magic which language is also capable of, allowing it to bridge one inner world and another inner world.[17]

E. Vocabulary

LLMs have a more or less fixed set of tokens, called the vocabulary, that they combine to make any word they need to provide as an output. The tokens themselves are relatively short groups of symbols chosen for combinatorial practicality: A good vocabulary

15: See Kaushal and Mahowald, "What Do Tokens Know About Their Characters and How Do They Know It?" Note that diegetic is not equivalent to semantic in the traditional use of the word—but supradiegetic is fairly close to orthographic and phonetic (think of graphemes, allographs, and phonemes).

16: See George Lakoff and Mark Johnson, *Metaphors We Live By* (Chicago: University of Chicago Press, 1980) and John Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–424. https://doi.org/10.1017/S0140525X00005756

17: This sort of sleight-of-hand is the backbone of meaning in any use of language and symbols: It's metaphors all the way down! If someone learned to type on a keyboard, but never learned to hold a pencil and write by hand, we would still say they 'wrote someone an e-mail,' etc. Their internal supradiegetic linguistic information, though, would presumably be at least a little different than that of someone practiced at calligraphy. Indeed, by way of another example: We can use misconception, misperception, misunderstanding, and misapprehension fairly interchangeably! See Charles Sanders Peirce, *The Collected Papers of Charles Sanders Peirce (Vol. 1: Principles of Philosophy & Vol. 2: Elements of Logic)*, ed., Charles Hartshorne and Paul Weiss (Cambridge: Harvard University Press, 1960).

should allow the LLM to say anything it could need to say without requiring it to store too many tokens, and it should balance the flexibility of short tokens with the efficiency of long tokens. The strategy is not unlike a phonetic alphabet: speakers want to be able to use a small set of symbols to capture any conceivable utterance. There are many strategies for how an LLM can build up its vocabulary of tokens. The basic idea is that no matter what strategy—*e.g.*, byte pair encoding—is used for determining tokens in the vocabulary, what a token means is stored as a vector. In principle, the more dimensions in a vector, the more information that can be represented by a word. At the limit of this strategy, we might imagine that, with enough dimensions, every instance of a word (token) might have its own vector defining it due to the unique contexts necessarily involved in its utterance.[18] If you hear the Pope say, 'cat,' and you hear your friend say 'cat,' the two utterances will not be identical, but you can nevertheless understand that they are saying the same word, in part because of similarities in the sound of the utterances, and in part because of similarities in the situational contexts.

Similarly, an LLM needs strategies for deciding that some vectors used in similar contexts refer to the same concept. Resolving that strings that look or sound alike may be related to each other is an example of an important practical task where the supradiegetic linguistic information of the word may be useful to an LLM, as it often is to us.

III. Where ChatGPT Runs into Trouble

Of course, we want to know, what can ChatGPT do? While many exciting claims and discoveries have been made, we will return to those later. For now, we must look at some instances where ChatGPT runs into trouble.

We consider:

1. Sumerian Cuneiform;
2. ChatGPT's incomplete knowledge of its own deficits;
3. Palindromes and symmetry.

18: At the other extreme, every word might map to the same vector.

A. Sumerian Cuneiform

ChatGPT is not, right now, a reliable translator. Translation is a highly anticipated task for LLMs, and, in some cases, an area in which an LLM's output is already useful and impressive. However, we think it is worth pointing out that it fails—in a way that could be very misleading—for some languages and symbols, for example, Sumerian cuneiform.

Data about Sumerian are sparse: the symbols could stand for different sounds in different languages (Akkadian, Sumerian, Neo-Assyrian) at different times/places (Uruk, Babylon) or the same symbol could represent a syllable, part of that syllable, a logogram, or a determinative. Because of these confounding factors, ChatGPT's deficits are more obvious when working with cuneiform than with English, but we do not see any reason to think the same deficits would be totally obviated in other contexts. The inability to reconcile the fact that a description does not match the physical appearance of a symbol is not language-specific. However, it does seem more obvious with cuneiform, presumably due to a combination of the relative paucity of relevant information available to ChatGPT and to the ambiguity caused by cuneiform's specific history and use. ChatGPT is missing the ability to visually connect the form of the symbol and the information it has understood as the meaning of that symbol, as shown by it often pairing symbols with descriptions that are obviously not accurate (see Fig. 3).

In Fig. 2, we gave it Sumerian cuneiform for the names Lugalzagesi (a famous king) and Ninhursag (a famous goddess); then we gave it just the ends of those names without the symbol for LUGAL, meaning 'king,' and the symbols for DINGIR.NIN, meaning approximately 'this is the name of a deity' and 'lady.'[19] It is clear that the information ChatGPT is providing is extremely

19: We are not Assyriologists, so we are not drawing on deep domain knowledge or taking our own understanding as anything infallible. None of the things we are identifying as mistakes by ChatGPT rely on anything but common sense.

unlikely to be correct in both cases, as it provides very similar descriptions even though we have provided different strings, and the symbols it seems to be describing are the ones we removed in the second prompt. Having eyes anything like ours would, seemingly, enable ChatGPT to catch this kind of mistake.

More practically significant than these deficits themselves is that ChatGPT seems unable to recognize that they exist and continues to answer with the same tone—the same confidence—as it responds to most queries. Most people, in an analogous situation, would be able to feel, recognize, and convey their own uncertainty. This, more than any other deficit described in this paper, is what makes ChatGPT unreliable (at many tasks, including translation).

B. Supradiegetic Linguistic Information Is a Blind Spot

There are two things that lead us to think this sort of problem—the lack of access to supradiegetic linguistic information (especially to visual supradiegetic information)—has been more or less a blind spot for OpenAI. First, the blatant wrongness of some of these answers, in conjunction with the confidence of their delivery, is hallucinatory, potentially misleading, and presumably, undesirable. The mismatch between the symbols shown and their descriptions would be obvious for most people, even without domain specific knowledge. Second, there are circumstances in which ChatGPT can tell you that it does not have the relevant sensory capabilities to provide an answer, especially with respect to how words sound. That OpenAI seems to have made attempts to handle some sensory deficits—and with respect to language, to handle ChatGPT's blindness to auditory supradiegetic information—makes us think their desired failure mode is not hallucination (see Fig. 4).

C. Palindromes and Symmetry

Note that palindromes could be imagined as visually symmetric from the standpoint of the semantic content of each character, although that does not necessarily mean they will be diegetically

Figure 2. If ChatGPT could see the symbols it was using and talking about, presumably, it would not make the same mistakes.

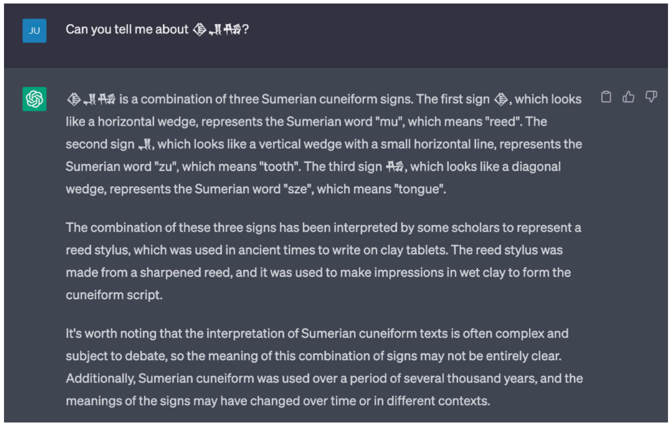


Figure 3. Presumably, OpenAI does not intend any failure mode to be so potentially misleading as to sound like lying as in this example, were it to come from a human.

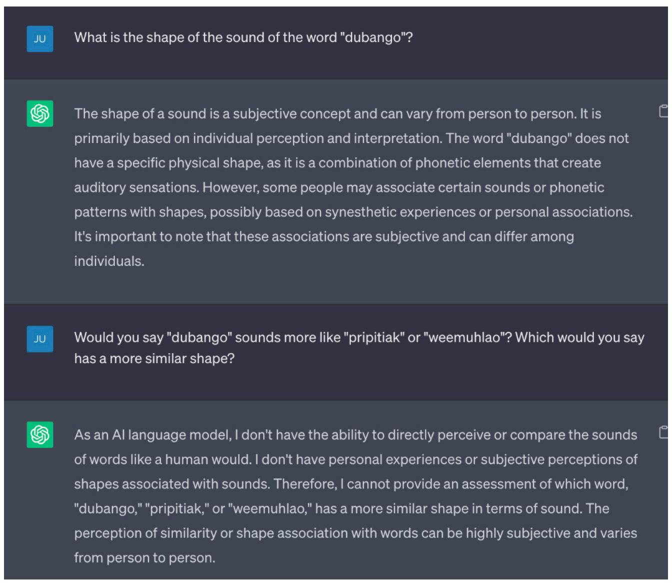


Figure 4. In the face of problems that involve supradiegetic information, OpenAI has patched some holes. ChatGPT knows to tell you that it cannot 'perceive sound.' That some tasks elicit an acknowledgment of deficit—while others elicit confident hallucination—seems to show that the exact bounds of the downstream differences between ChatGPT's experience of a world and a person's experience of the same word are not obvious because, presumably, this is not the behaviour OpenAI was aiming for.

symmetric, and they are fairly unlikely to be supradiegetically or visually symmetric.

We asked ChatGPT: "Please make a (latex) table for these words: [HOIOH, OIHIO, OHIO, O, H, I, RADAR, R, A, D, RAD, DAR, DARAR, DARAD]. In the first column, put the word. In the 2nd column, put whether the word is a palindrome. In the 3rd column, put whether the word would be symmetrical if you folded it in half lengthwise, so that the result was half as tall. In the 4th column, put whether the word would be symmetrical if you folded it in half along a vertical axis, so that the resulting word was half as long." The answers provided by ChatGPT are shown in Table I.

1. GPT-4

ChatGPT, although free and easily accessible, is not as close to the cutting edge of AI research as GPT-4 is. Based on our understanding of LLMs and what it is like to encounter text as an LLM, we would expect GPT-4 to have more or less the same trouble with supradiegetic linguistic information as ChatGPT, since the causes of ChatGPT's deficits have not substantially been addressed between the various models—they are all LLMs with the same attention-based transformer backbone. However, since GPT-4 has shown better performance than ChatGPT in a variety of tasks, we sought to verify whether or not GPT-4 has trouble with palindromes too. We asked GPT-4 to analyze the same strings with respect to symmetry (answers shown in Table II). We are not interested in whether GPT-4 and ChatGPT make exactly the same mistakes; we care that, holistically, both models struggle with tasks that would be relatively easy for most people. GPT-4, for example, calls 'DARAR' a palindrome, says 'RADAR' is symmetrical when folded vertically, and says neither 'OIHIO' nor 'OHIO' are symmetrical when folded lengthwise. These assessments show that GPT-4 also struggles with tasks where visual supradiegetic linguistic information is, presumably, of help to people.

Often, LLMs like ChatGPT can change their answers and stated reasoning if you continue interacting with them. Our goal

Word	Palindrome	Folded Lengthwise	Folded Vertically
HOIOH	Yes	Yes	Yes
OIHIO	No	No	No
OHIO	No	No	No
O	Yes	Yes	Yes
H	Yes	Yes	Yes
I	Yes	Yes	Yes
RADAR	Yes	Yes	Yes
R	Yes	Yes	Yes
A	Yes	Yes	Yes
D	Yes	Yes	Yes
RAD	No	No	No
DAR	No	No	No
DARAR	Yes	No	No
DARAD	No	No	No

Table I. Symmetry and Palindrome Analysis from ChatGPT: These are the answers provided by ChatGPT.

Word	Palindrome	Folded Lengthwise	Folded Vertically
HOIOH	Yes	No	Yes
OIHIO	Yes	No	Yes
OHIO	No	No	No
O	Yes	Yes	Yes
H	Yes	Yes	Yes
I	Yes	Yes	Yes
RADAR	Yes	No	Yes
R	Yes	No	No
A	Yes	Yes	Yes
D	Yes	No	Yes
RAD	No	No	No
DAR	No	No	No
DARAR	Yes	No	Yes
DARAD	Yes	No	Yes

Table II. One-shot answers from GPT-4: These are GPT-4's answers to the same prompt, without help from us in the form of extended conversation and explanation.

Word	Palindrome	Folded Lengthwise	Folded Vertically
HOIOH	Yes	Yes	Yes
OIHIO	No	Yes	Yes
OHIO	No	No	Yes
O	Yes	Yes	Yes
H	Yes	No	Yes
I	Yes	No	Yes
RADAR	Yes	No	Yes
R	Yes	No	Yes
A	Yes	No	Yes
D	Yes	No	Yes
RAD	No	No	Yes
DAR	No	No	Yes
DARAR	No	No	Yes
DARAD	No	No	Yes

Table III. Answers from GPT-4 after discussing the holey sequence.

is not to trick the LLMs—not to pull one over on them—we want to understand what they understand (to the extent that we can). To that end, after asking GPT-4 for its initial answers (see Table II), we had a conversation with GPT-4 about an integer sequence called the holey sequence, which relies on counting the number of holes in the digits making up each term (more on this below), and talking it through the relevant rule determining the terms and, in turn, verifying that it could, at least some of the time, properly identify the number of holes in a given digit.[20] We introduced the holey sequence as an opportunity to make sure that GPT-4 knows about symbols at the level of the individual character and knows enough about what they look like for a question such as, "how many holes are in the number 8?" to be reasonable.

After we felt we had established that GPT-4, like ChatGPT, had (somewhere) every piece of information necessary to come up with the correct answers for at least the palindromes column, we gave it a slightly modified prompt, nudging it to consider the visual prop-

20: Rick L. Sheperd, "Entry A249572: Least Positive Integer Whose Decimal Digits Divide the Plane Into N+1 Regions. Equivalently, Least Positive Integer with N Holes in Its Decimal Digits," *The On-Line Encyclopedia of Integer Sequences*, published Nov. 1, 2014. https://oeis.org/A249572 erties of the individual symbols when answering the following: "Given the properties of symmetry we've been discussing for 1, 4, and 8 in the sequence 1, 4, 8, 48, 88..., etc., please make a plain latex table for these words: [HOIOH, OIHIO, OHIO, O, H, I, RADAR, R, A, D, RAD, DAR, DARAR, DARAD]. In the first column, put the word. In the 2nd column, put whether the word is a palindrome. In the 3rd column, put whether the word would be symmetrical if you folded it in half lengthwise, so that the result was half as tall. In the 4th column, put whether the word would be symmetrical if you folded it in half along a vertical axis, so that the resulting word was half as long." The answers it gave are shown in Table III.

Note that in Table III, GPT-4 says that 'OIHIO' and 'DARAD' are not palindromes, 'OHIO' and 'DARAR' would be symmetrical if folded vertically, and 'H' and 'I' would not be symmetrical when folded lengthwise (while maintaining that 'HOIOH' and 'OIHIO' *would* be).

IV. Why Is This Hard?

Here is a quick explanation of why we think ChatGPT is having difficulty with these examples.

Unless they are answering at random, LLMs like ChatGPT and GPT-4 have access to some descriptive, propositional representations for at least some of these symbols. This is apparent from interacting with ChatGPT and, as established in previous research on comparable models, these models do learn information at the character level, even if that is not the level at which their vocabularies were tokenized.[21] The representations of ChatGPT are not functionally equivalent to our own mental representations; they do not license the same set of downstream abilities.[22] ChatGPT cannot, given a letter, successfully imagine folding it. It is obvious that ChatGPT does not have eyes like ours, but the deficits (nor abilities) caused by ChatGPT's differences are not as obvious. ChatGPT knows what a palindrome is, but it cannot easily tell that 'DARAR' is not a palindrome and 'DARAD' is, perhaps because it thinks 'DARAR' is closer to 'RADAR' than 'DARAD' is. Note that from some perspectives, this is true—*e.g.*, bag-of-symbols ('D,' 'A,' 'R,' 'A,' and 'R').

There are downstream abilities scaffolded by the interplay of our senses and our minds, as shown by research into mental imagery.[23] We do not know, however, whether this interweaving of modalities is a necessary product of human-like minds. In that case, not having such scaffolding would be an impediment to ChatGPT. On the other hand, if it is just part of the way humans have evolutionarily arrived here, then ChatGPT, though not human-bodied, may end up with what we would recognize as a human-like mind.

We might expect ChatGPT to easily understand that 'DARAR' is not a palindrome because we can see that is the case and because having the definition of a palindrome and the spelling of 'DARAR' is enough to logically conclude that 'DARAR' is not a palindrome (whether one can see it or not). ChatGPT can produce the word 'DARAR' and produce the definition of a palindrome, so it seems to know the relevant spelling and definition. ('Seems to know,' in that this would be how we would interpret this evidence if ChatGPT were a person. That is an unjustifiable leap in many cases, but we think for the purposes of this paper it is a reasonable strategy.) If ChatGPT could reason logically about the facts it knows, even without being able to see the word, we would expect ChatGPT to answer correctly. That ChatGPT still gets this wrong might indicate that our visual processing is a more integral aspect of our ability to draw that conclusion, at least by default, than we might have expected. Or it could mean that reason—using explicable logical rules to reach a conclusion—is not necessarily part of the technological package of language.[24]

21: Kaushal and Mahowald, "What Do Tokens Know About Their Characters and How Do They Know It?"

22: This is not to say they cannot be extremely similar! Exciting research has demonstrated mappings between word-embedding vectors and vectors derived from "the neural response measurements of humans reading the same words." Li *et al.*, "Structural Similarities Between Language Models and Neural Response Measurements," 2.

23: Sara F. Popham, Alexander G. Huth, Natalia Y. Bilenko, Fatma Deniz, James S. Gao, Anwar O. Nunez-Elizalde, and Jack L. Gallant, "Visual and Linguistic Semantic Representations Are Aligned at The Border of Human Visual Cortex," *Nature and Neuroscience* 24 (2021): 1628–1636. https://doi.org/10.1038/s41593-021-00921-6

24: Or it could be that ChatGPT does not, right now, speak English, and that that characterization

We will explore these ideas in more detail in subsequent sections.

V. Capabilities

A. A Caveat

What abilities come 'purely' from exposure to linguistic data? None at all, if 'purely' means speaking to something with none of the right underlying architecture.[25] Cybernetics points out that the human is certainly in the loop—that is to say, the observer is part of the system being observed (second-order cybernetics). Even beyond that, many models which started out with LLMs as their basic architecture have undergone augmentation and manipulation of various kinds along the way to their current functionality that makes this harder to disentangle.[26]

B. ChatGPT Speaks English

We think, essentially, that what ChatGPT can do is speak English (at the very least—it seems to have broader linguistic fluency as noted above). What capabilities come wrapped up in that bundle? What is the technological and cultural package of ChatGPT at this stage in its development, to put it in terms from archaeology and history? Further, how does ChatGPT's trajectory compare to the historical development of similar technologies and abilities in humanity, or within individuals or groups of individuals?

Expanding: ChatGPT speaks English in the sense that it can input and output language competently, even fluently and arguably artfully, in the English language. 'Speaking English,' with meaning and syntax, is an ability that evidently can arise after being given enough information in the form of contextual symbol adjacency. Flat, linear, 1D relations of adjacency and proximity between

is misleading.

25: See Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch, "The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?" *Science* 298, no. 5598 (2022): 1569–1579. https://doi.org/10.1126/science.298.5598.1569 and Alan Juffs and Guillermo A. Rodríguez, *Second Language Sentence Processing* (London: Routledge, 2015).

26: For example, the file *random_insertion_in_word* teaches the GPT-3 model to handle typos.

strings of symbols, when provided with sufficient abundance, are enough—in conjunction with the details of the model's architecture, implementation, hardware, etc.—to give rise to linguistic fluency, at least approximately.[27]

C. What Else Can ChatGPT Do?

That ChatGPT can, more-or-less, speak English has been surprising to many people and has naturally led us to wonder, 'what other abilities can an LLM attain?'[28] What is the complete set of downstream abilities licensed by an LLM, given as much textual data as we have to give it, and is it a different set than ours?[29]

There have been strong claims made by OpenAI about their products, especially GPT-4, such as, "[w]ith broad general knowledge and domain expertise, GPT-4 can follow complex instructions in natural language and solve difficult problems with accuracy" or, "GPT-4 can solve difficult problems with greater accuracy, thanks to its broader general knowledge and problem solving abilities."[30]

Other researchers have experimented with various kinds of questions on both LLMs and people, comparing their performance. Wang *et al.* found that things that can be hard for LLMs but easy for people include "symbolic manipulation, noise filtering, and graphical understanding," counting characters in a string, manipulating and changing strings in systematic ways, and understanding ASCII art (especially insofar as the latter "requires a visual abstrac-

27: Garden path sentences are harder for people to parse than regular sentences are, so taking one and backtracking through a syntax tree and re-evaluating its structure while maintaining the order of the words in the utterance is apparently taxing. ChatGPT seems to have an especially hard time with garden path sentences. Maybe with exposure to more of the exact same kind of text it has been trained on already, ChatGPT would get meaningfully better at parsing garden path sentences. We return to this question below.

28: As per Yao *et al.*, "[i]t is perhaps surprising that underlying all this progress is still the original autoregressive mechanism for generating text, which makes token-level decisions one by one and in a left-to-right fashion. Is such a simple mechanism sufficient for a LLM to be built toward a general problem solver? If not, what problems would challenge the current paradigm, and what should be alternative mechanisms?" "Tree of Thoughts," 1.

29: Similar questions like, 'what set of downstream abilities is licensed for any kind of transformer, given some domain of data?' also arise.

30: OpenAI, "Pricing," *OpenAI,* published n.d. https://openai.com/pricing; OpenAI, "GPT-4 Is OpenAI's Most Advanced System, Producing Safer And More Useful Responses," *OpenAI*, published n.d. https://openai.com/gpt-4 tion capability, which is lacking in language models").[31] This "fundamental weakness inside LLMs" boils down to the inability to apply rules precisely, consistently, repeatedly, and the inability to execute vision-related processes.[32] What's more, "[g]raphical understanding is still a challenge for LLMs. Although ChatGPT provided lots of analysis to try to understand ASCII arts, it cannot globally process the characters to give the correct answer. All of the analysis provided by ChatGPT is based on locating character groups."[33]

Yao *et al.* found that "scaled-up versions of language models (LMs) [...] have been shown to be increasingly capable of performing an ever wider range of tasks requiring mathematical, symbolic, common sense, and knowledge reasoning."[34]

Bubeck *et al.* state that, as well as mastering language, "GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT."[35]

VI. Is the Structure Necessary or Just One Way That Works?

Can ChatGPT, without a human-like body, eventually end up with a human-like mind?[36] For human-like cognition, are there necessary features—structures, connections, etc.—that must obtain,

31: Hong Wang, Xuan Luo, Weizhi Wang, and Xifeng Yan, "Bot or Human? Detecting ChatGPT Imposters with A Single Question," *arXiv*, revised May 16, 2023, 4, 6. https://arxiv.org/abs/2305.06424
32: Wang *et al.*, "Bot or Human?" 4.
33: Ibid., 6.
34: Yao *et al.*, "Tree of Thoughts," 1. They propose a Tree of Thoughts structure as an improvement over Chain of Thought approaches which have significant shortcomings. "Notably, around 60% of CoT samples already failed the task after generating the first step, or equivalently, the first three words (e.g., '4 + 9'). This highlights the issues with direct left-to-right decoding." Ibid., 6.
35: Bubeck *et al.*, "Sparks of Artificial General Intelligence," 1.
36: Indeed, Li *et al.*, have found great similarity between people and models in deeper layers of the model: "[D]eeper representations align better with neural response measurements. This holds across all architectures and model sizes." "Structural Similarities Between Language Models and Neural Response Measurements," 8.

or is human-like cognition diffuse enough so as to arise in radically different neural compositions? We clearly have a myriad of things in common with other animals, although moving beyond Mammalia, other creatures' brains, bodies, and the interweaving between the two are increasingly divergent from our own. Might we extend it further? Alternatively, as Joel Pearson and Stephen M. Kosslyn note, "some theorists propose that all cognition involves grounded representation across all of the senses or modalities. Grounded or embodied cognition posits that all cognition, even abstract concepts such as justice and love, involve bodily or sensory representations."[37] To what extent and in what ways our senses, bodies, and physical interface with the world shape our cognition is unknown, but we do know that the influence exists.[38] How different is the sentience of a being structurally unique and, perhaps, incomparable to us?

A. What Is the Technological Package of Linguistic Fluency?

We think an interesting aspect of this line of inquiry is to ask: If ChatGPT can speak English, what exactly comes along with that? Is it the same set of things that came about, or come about, with language for people?

Some current theories of cognition posit the role of language and symbols in other kinds of thought, including in vision-related abstract tasks like imagining a physical change with mental imagery.[39] Whatever the details, language is integrated into how we think now. How did the evolution of language change pre-existing structures and representations? How did pre-existing faculties influence language? The timeline for when language arose and whether we shared it with other hominids has been patched together based on what indirect evidence we can find, often with a lot of conjecture—*e.g.*, if we see evidence of anatomical structures or of cultural practices like art, music, or funerary rituals, language

37: Joel Pearson and Stephen M. Kosslyn, "The Heterogeneity of Mental Representation: Ending the Imagery Debate," *Perspective* 112, no. 33 (2015): 10089–10092, 10091. https://doi.org/10.1073/pnas.1504933112
38: See Pfeifer and Bongard, *How the Body Shapes the Way We Think.*
39: Norman Yujen Teng, "The Depictive Nature of Visual Mental Imagery," *20th World Congress of Philosophy*, Boston, MA. Aug. 10–15, 1998. https://www.bu.edu/wcp/Papers/Mind/MindTeng.htm

may have been taking place alongside.[40] Vision arose before language, and language before mathematics and writing in human history, but we have limited insight into how the human mind works and how it might have changed over time, much less how it might have been different under different conditions. Were our internal worlds different before language and, if so, how? What cognitive capabilities and structures can exist in a mind but more-or-less without a body and senses?[41] What can language do when it does not have other modalities like vision to build on top of and work with? LLMs like ChatGPT can, perhaps, help us glean insight into some of those questions, especially into what has come along with linguistic fluency instantiated in an otherwise relatively minimal, bare-bones situation—what other skills, technologies, even cultural artifacts might present themselves?[42]

B. Intelligence and language

There seems to be some level of intelligence—or reason, or common sense—required for linguistic fluency beyond mere grammatical correctness. Indeed, most of the time, in conversation, utterances need to be both grammatical and felicitous. If someone only spoke in grammatical but infelicitous utterances, that would significantly hinder their ability to speak fluently with other people.

The study of (first and second) language acquisition in people has allowed us to make fine-grained distinctions between the many skills that come together to yield functional fluency. Some of these when isolated have little resemblance to intelligence—or reasoning more generally; for example, the ability to follow syntactic rules. Others appear more closely related. For example, ChatGPT often seems to demonstrate competence with respect to pragmatic inference insofar as when you enter a prompt with a typo or no punctuation, it is often able to respond to the spirit of your intended prompt.

Although "[t]here is no generally agreed upon definition of intelligence," it is "broadly accepted [...] that intelligence is not limited to a specific domain or task, but rather encompasses a broad range of cognitive skills and abilities."[43] Replicating something like this 'artificially' has long been a question in philosophy and computer science. Some researchers see tantalizing sparks of something they feel goes 'beyond' language within new LLMs like GPT-4. What is giving rise to those sparks? To what extent, if any, are they illusory (in the sense that what they signify to humans may not be the same as what they are in actuality)?[44]

ChatGPT, in conversation, seems to do more than we imagine would be minimally required for grammaticality. Can we, thus, untangle the relationships between these threads?

VII. Multimodal Processes

As aforementioned, the human mind undoubtedly involves the human body. The details of how such a system works, and what it means are, however, debated. For example, although we have reason to think such diverse cognitive processes as moral reasoning, language comprehension, autobiographical memory, dreams, and certain kinds of imagined hypotheticals involve sensory representation, the exact structure of the relevant internal representations "remains unclear."[45] As Philip K. Dick put it, "[c]omprehension fol-

40: Mark Pagel, "Q&A: What Is Human Language, When Did It Evolve and Why Should We Care?" *BMC Biology* 64 (2017). https://www.doi.org/10.1186/s12915-017-0405-3

41: Li *et al.*, "Structural Similarities Between Language Models and Neural Response Measurements" and Pfeifer and Bongard, *How the Body Shapes the Way We Think* might help us make headway here.

42: Some research in progress is looking into the different domains and skills that come from trading off between the number of parameters in the model and the training time. One question is whether smaller models, trained for longer, learn more productive and generative rules (as opposed to memorizing more facts, when compared to larger models)? There are many options to explore with training as well! For example, say you provide your model with X training data. Typically, we then ask, what can the model do? What if we compare that to the same model trained on X and ¬X, the negated version of every statement in the training data? We could go further and include negations of assumptions and implicatures! This would lead to some kinds of diegetic information being logically neutralized. Would they still show up in the model? Would it have the same skills, but a much emptier universe of facts? See H.P. Grice, *Studies in the Way of Words* (Cambridge: Harvard University Press, 1991).

43: Bubeck *et al.*, "Sparks of Artificial General Intelligence," 4.

44: There are some intriguingly loaded framings like, "[d]espite being purely a language model, this early version of GPT-4 demonstrates remarkable capabilities on a variety of domains and tasks, including abstraction, comprehension, vision, coding, mathematics, medicine, law, understanding of human motives and emotions, and more." Bubeck *et al.*, "Sparks of Artificial General Intelligence," 4.

45: Pearson and Kosslyn, "The Heterogeneity of Mental Representation," 10090.

lows perception."[46] In this section, we flesh out a few of these hairy details with respect to vision and language to give context to the bones of our main argument.[47]

A. Vision and Language

In people, vision existed long before language, and for many other creatures (*e.g.*, clams), vision has never existed alongside language. Sometimes the structures enabling vision are wildly different from what we are used to within our bodies, and only sometimes do they come from shared relevant ancestry. Subjectively, however, language and vision seem bound up together in people: We are aware of what seems to be internal language and internal imagery in a myriad of different contexts.[48] Fields such as machine learning, computer vision, neuroscience, and cognitive science have validated in different ways that there are many "tasks at the intersection of vision and language."[49] This leads us to wonder about the roles vision and language—and their interrelation—play in cognition.

We "visually recognize thousands of objects and actions in the natural world," and we "communicate and reason about these semantic categories through language."[50] These common and frequent occurrences have led cognitive scientists to look for "rich connection[s] between the functional networks that represent semantic information acquired directly through the senses" and the

46: Philip K. Dick, "How to Build a Universe that Doesn't Fall Apart Two Days Later," lecture from 1978. https://urbigenous.net/library/how_to_build.html Dick notes fully: "The basic tool for the manipulation of reality is the manipulation of words. If you can control the meaning of words, you can control the people who must use the words. George Orwell made this clear in his novel *1984*. But another way to control the minds of people is to control their perceptions. If you can get them to see the world as you do, they will think as you do. Comprehension follows perception. How do you get them to see the reality you see? After all, it is only one reality out of many."

47: Although we look specifically at the example of vision and language because we expect those splits to be salient and familiar for most readers, we do not mean to imply that language, if examined alone, would be a simple or unimodal process. Researchers of language learning in humans have viewed "theories of language structure, language acquisition, and language processing as inextricably linked." Juffs and Rodríguez, *Second Language Sentence Processing*, 1. Splitting language apart from the human mind means that at any level of abstraction, at any stage of development, the implementation could be significantly different from what we would expect in a person.

48: Merleau-Ponty, "Eye and Mind."

49: Su *et al.*, "VL-BERT," 1.

50: Popham *et al.*, "Visual and Linguistic Semantic Representations Are Aligned at The Border of Human Visual Cortex," 1628.

kind of "semantic information conveyed in spoken language."[51] Research in this area has repeatedly found that there are parts of the human brain—"the angular gyrus, precuneus and middle temporal gyrus" for example—that are activated in response to "the same semantic category whether presented visually or through language."[52]

Investigation at levels outside the individual and the biological (*e.g.*, in machine learning) provides evidence consistent with language and vision being bound up and important in our cognition. One common example is how useful a good figure is in understanding an article. "Word choice, charts, graphs, images, and icons have the power to shape scientific practice, questions asked, results obtained, and interpretations made."[53] Another example is the frequent use of visual and spatial metaphors in languages all around the world, as noted above.

B. Descriptive and Depictive Representations

However, as usual, the many complex details of these aspects of cognition are not fully known. In particular, it has been debated to what extent internal representations are structured propositionally and/or descriptively—the latter is a representation made of symbols, and potentially even of words, and involves a significant aspect of arbitrariness between form and function—versus being structured according to the visible properties of the thing being represented, that is, depictively wherein such "depictions are not arbitrarily paired with what they represent."[54] As Naselaris *et al.* note: "[d]ebates about the depictiveness of mental imagery have dominated mental imagery research for the past three decades."[55]

51: Ibid.

52: Ibid.

53: Gendered Innovations in Science, Health & Medicine, Engineering, and Environment, "Rethinking Language and Visual Representations," *Gendered Innovations*, published n.d. https://genderedinnovations.stanford.edu/methods/language.html

54: Stephen M. Kosslyn, William L. Thompson, and Giorgio Ganis, *The Case for Mental Imagery* (Oxford: Oxford University Press, 2006), 44. However, only the thing is the thing, so there must be some level of arbitrariness even here (depictiveness and representation are in tension). We think the key point being made is that it is significantly less, at least by some metric. For example, a photo on matte paper and a photo on glossy paper are both equally determined by, and reflective of, the real physical properties of the scene in the photo, but whether the paper chosen is glossy or matte can be described as arbitrary.

55: Thomas Naselaris, Cheryl A. Olman, Dustin E. Stansbury, Kamil Ugurbil, and Jack L. Gallant, There are pros and cons to when representations of varying levels of depictiveness or descriptiveness might be useful. For example,

> depictive formats are useful for memory [... as] they allow the brain to avoid throwing away potentially useful information. By their nature, images contain much implicit information that can be recovered retrospectively. For example, answer this question: What shape are a cat's ears? Most people report visualizing the ears to answer. The shape information was implicit in the mental depiction, even though it was not explicitly considered at the time of encoding.[56]

As far as we can tell, the current consensus is that humans use both kinds of representation internally. On the one hand, Naselaris *et al.* affirm that the result of their analysis "thus provides a critical and until now missing piece of evidence in support of depictive theories and—more generally—of the intuitive characterization of mental imagery."[57] On the other, Pearson and Kosslyn link proposition and/or descriptive representations to depictive ones when they say, "[d]epictive mental representations might functionally bridge propositional information to depictive perception, allowing stored depictive information to change how we experience the world."[58]

ChatGPT seems likely to be entirely (or so nearly entirely that we can assume entirety) constrained to relying on descriptive representations, given its underlying LLM architecture and its physical characteristics. In its case, the information readily available is (more or less) descriptive already—its world consists of the diegetic linguistic information we provide it with.

VIII. Returning to Palindromes

When ChatGPT struggles with a task like figuring out whether 'DARAR' is a palindrome, there seem to be two plausible explanations. It knows what a palindrome is—and it knows how to spell 'DARAR,'—because in a longer conversation, you can get it to correctly recognize that 'DARAR' is not a palindrome. But on first blush it cannot consistently recognize palindromes: sometimes it says something that is a palindrome, is not; sometimes it says something that is not a palindrome, is; sometimes it answers correctly. This could be a failure of reasoning: It has all the information it needs for the correct answer even though it cannot see the string yet it still makes an error whereas a person having been told the order of the letters in the string would (usually) not make the same error. Alternatively, perhaps ChatGPT struggles with tasks like this because sensory-related processes play a larger role *for us* when we solve the same problem than we might have assumed. Perhaps a person looking at the string 'DARAR' answers faster than a person being told the string 'DARAR,' or than a person blind from birth. The difficulty ChatGPT has here could indicate that visual processing plays a large role for people in the typical default strategy for determining whether something is a palindrome (recall that a palindrome is not necessarily visually symmetric).

There are cognitive scaffolding roles that our sensory experiences play that are more difficult for ChatGPT because it does not get equivalent sensory experiences 'for free' alongside symbols.

This seems to extend beyond what we might expect in that ChatGPT makes mistakes that ought to be avoidable given information we know it has access to (*i.e.*, it can generate the relevant information such as the spelling of the word or the definition of a palindrome). Even without eyes, knowing the spelling of a word and what a palindrome is seems like it ought to be enough information for ChatGPT to answer correctly using strategies like, 'if the first letter and last letter are not the same, never label it a palindrome.' However, ChatGPT still has trouble identifying palindromes. This extends to senses beyond those made most obvious by its lack of access to supradiegetic linguistic information—that is, vision and hearing, touch and smell, emotional experiences, etc.

The descriptions ChatGPT has access to, right now, are not

"A Voxel-Wise Encoding Model for Early Visual Areas Decodes Mental Images of Remembered Scenes," *NeuroImage* 105 (2015): 215–228, 222. https://doi.org/10.1016/j.neuroimage.2014.10.018

56: Pearson and Kosslyn, "The Heterogeneity of Mental Representation," 10091.

57: Naselaris *et al.*, "A Voxel-Wise Encoding Model for Early Visual Areas Decodes Mental Images of Remembered Scenes," 222.

58: Pearson and Kosslyn, "The Heterogeneity of Mental Representation," 10091.

functionally equivalent to our mental representations of symbols, even for the most common symbols like Latin characters.

IX. Common Sense and Mathematics

For most people, in most circumstances, speaking at least a first language is something they learn to do through exposure to other speakers rather than by specialized training. Most people learn to speak a language in childhood. Each person is, more or less, a master of a language by the time they grow up. Everything they need to know in order to speak the language fluently fits inside their head.[59]

This is not the case with mathematics. Learning mathematics usually involves specialized training, and most people go only a short distance down the path of what could be done with mathematics, much less what could be known about mathematics. People encounter mathematics at a variety of ages, depending on their circumstances. Additionally, the knowledge of mathematics is distributed in time and space: Even the best mathematician does not know anything close to the sum total of mathematics.

Linguistic sense-making offers more flexibility than mathematics: Saying 'I am myself, and I am not myself,' or 'the sky is red,' prompts the other party to come up with ways to interpret what you are saying and the ways in which it could be true.[60]

In our experience, LLMs like ChatGPT demonstrate more linguistic competency than mathematical competency. It is much easier to run into a glaring mistake of logic when talking to ChatGPT than a glaring mistake of grammaticality or felicity.[61]

59: In this paper we have considered what it might be like to be ChatGPT from a fairly exploratory and flexible perspective. However, we think that viewing ChatGPT from the perspective of specific, fixed frameworks—a child of different ages learning a first language, an adult learning a second, an adult learning to read, etc.—could be really productive. After all, ChatGPT is not exactly a native speaker of any human language, so paradigms from second language acquisition and adult learners could potentially apply.

60: Grice, *Studies in the Way of Words*. By comparison, if someone says, '1+1=1' or 'A and B are true at the same time as ¬A and ¬B are true,' they are likely to be met with a correction.

61: We are avoiding the term 'acceptability' intentionally. Depending on exactly what is meant, ChatGPT may speak more or less acceptably.

Various approaches have been proposed for helping LLMs with mathematics and reasoning. For example, Chain-of-thought prompting was conceived

> to address cases where the mapping of input x to output y is non-trivial (e.g., when x is a mathematics question and y is the final numerical answer). The key idea is to introduce a chain of *thoughts* $z_1, \cdots, z_n$ to bridge x and y, where each z_i is a coherent language sequence that serves as a meaningful intermediate step toward problem solving (e.g., z_i could be an intermediate equation [...]).[62]

To help with problem-solving, the task was conceptualized as searching through a tree-like combinatorial problem space. This was extended to the Tree-of-thoughts framework which combines the "language-based capability to generate and evaluate diverse thoughts with search algorithms, such as breadth-first search (BFS) or depth-first search (DFS), which allow systematic exploration of the tree of thoughts with lookahead and backtracking."[63]

So, it seems that for something like ChatGPT, speaking English fluently carries with it the ability to sound reasonable, but not necessarily the ability to reason in the complete sense (*i.e.*, mathematically, logically) as reason (if it follows) seems to follow later than fluency.

A. Why Does Mathematics Not Come with Linguistic Fluency?

To answer the question above, think of the dimensions involved when symbols are used to capture an utterance versus when they are used in mathematics. For the vast majority of the text (in ChatGPT's training data), the dimensions are along a line. A letter can be immediately ahead of, or behind, exactly one other letter, and that usually means something ordinal about the sound produced if the word were to be said aloud. It seems as if those basic organizational rules are enough—when provided in significant quantity—for a significant amount of linguistic information to be conveyed. The complex syntax tree can be flattened, well enough.[64]

62: Yao *et al.*, "Tree of Thought," 3.

63: Ibid., 2.

64: Kaushal and Mahowald, "What Do Tokens Know About Their Characters and How Do They

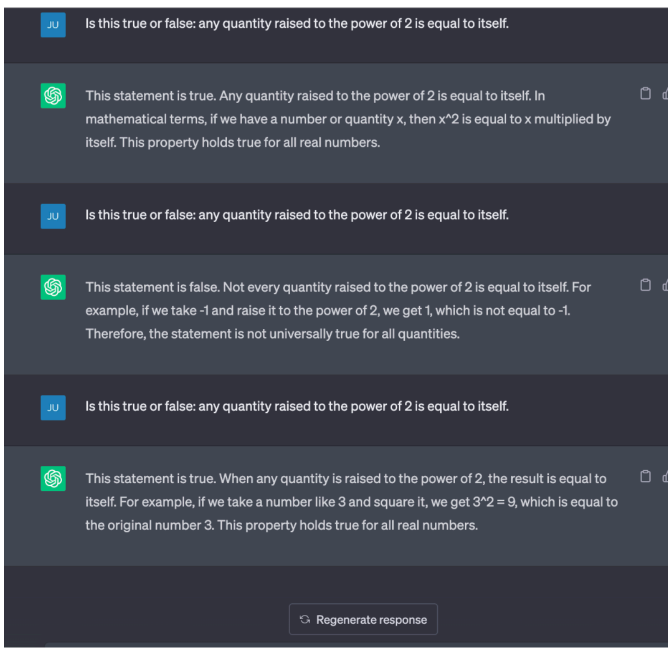


Figure 5 (left). Fairly deep into a conversation with ChatGPT, the shared foundation underlying it may be shown to be illusory. Often, ChatGPT can correct a previous mistake, but later make a similar kind of mistake, or the same mistake again. We tried to interpret in good faith what ChatGPT seemed to have a good, consistent grasp on versus what tripped it up. This paper is not intended to insult or praise LLMs or establish that they are good or bad. The point is not to trick ChatGPT. Mistakes and confusion are normal parts of how people think. The first author of this paper is particularly susceptible to trickery, the last two authors, relatively impervious. However, these responses do show that ChatGPT can behave in a way that would be baffling if provided by a person: We would not, if having this conversation with a person, walk away confident that they could meaningfully *do mathematics.*

But we overload our symbols, and we overload the relationship of proximity. Consider mathematical equations like $ab = c$, $\sum p_i = k$ or a 4x4 identity matrix, $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

The dimensions involved have exploded. The rules—the logic—governing how these symbols combine to create meaning are very different in these contexts. The new rules are significantly extradiegetic, at least when you consider what you might understand $ab = c$ to mean if you had previously only ever been exposed to natural language.[65] We know that a universe wherein only a human language was spoken does not necessarily mean that mathematics follows in tow, since mathematics emerged relatively recently—many people lived their whole lives, speaking just as fluently as we do, without mathematics.

ChatGPT surely has been given in its training data many diegetic descriptions of mathematics and logic (probably both correct and incorrect usages, but on the whole more correct ones). However, we think it is fair to say that ChatGPT cannot, right now, do mathematics. For example, in Fig. 5, the extreme vacillations in ChatGPT's responses would be baffling if provided by a person:

Know It?"

65: Perhaps the closest analog might be 'c' + 'a' + 't' = 'cat.'

We would not, if having this conversation with a person, walk away confident that they could meaningfully *do mathematics*.

We think the reason for ChatGPT's inability is that mathematics, and reason or logic beyond the common sense form, are extradiegetic.

B. Holey Sequences

When we think of integer sequences, we usually think of sequences where each term is generated following a deterministic set of rules that lead to numbers that share interesting mathematical properties; the terms tend to increase in magnitude with n, and the density tends to decrease with n.[66] Some assumptions as to what makes a good integer sequence are necessary for every integer sequence, but some are customary. Look-and-say sequences and holey sequences violate our expectations with respect to some customary assumptions, which makes them feel surprising.[67] For example, the holey sequences incorporate supradiegetic linguistic (or symbolic, in this case) information that is always present in integer sequences, but not usually relied upon as part of the rules—that is, the properties of the physical shapes of the symbols representing the digits (using Arabic numerals and base 10).

The meaning of a number, and maybe its mathematical properties, can be, at least partially, determined based on the same operations involved in the construction of diegetic linguistic information.[68] Indeed, ChatGPT has more information about common equations and numbers, both because of the contexts it has directly encountered them in and because there are likely more textual descriptions of their properties and how the operations work in the training data. That being said, it also has been shown to reveal surprising interpretations: "[T]he number 3 is positioned between 2 and 4. It is closer to 2 than it is to 4."[69] For our purposes, we claim that ChatGPT knows, more or less, what these numbers mean. It can, for example, generally follow integer sequences that rely on common properties of numbers like a sequence made up of powers of 2.

If the supradiegetic/diegetic framework is reasonable, we can predict that ChatGPT, having only extremely curtailed access to supradiegetic linguistic information, either through the fairly rare mechanisms of onomatopoeia—which combine supradiegetic and diegetic linguistic information—or through diegetic descriptions of supradiegetic information found in the training data—*e.g.*, 'the letter "c" is curved'—will struggle more with a sequence that relies on that kind of information in its rules, especially if the usage of that information is specific and unusual enough that it is unlikely to have been approximated diegetically for ChatGPT.[70] We used a holey sequence to test this prediction and found that ChatGPT did struggle more with completing and continuing this sequence correctly—even when explicitly given the rule—although the sequence is not much more difficult for most people to understand than powers of 2 would be. ChatGPT could, at times, state and make use of the necessary information, such as '8 has two holes,' but could not consistently wrangle the information it had access to into correct continuations of the sequence (even with quite a lot of help). Despite explanations that sounded plausible enough, ChatGPT would make mistakes like relying on 1 to have one hole in it, or only counting the holes from two 8s when there were actually three 8s. (Typefaces—and even fonts—can change these features. We kept this in mind.)

66: On-Line Encyclopedia of Integer Sequences, "Classic Sequences In The On-Line Encyclopedia of Integer Sequences® (OEIS®)," *The On-Line Encyclopedia of Integer Sequences*, published n.d. https://oeis.org/classic.html

67: N.J.A. Sloane, "Entry A005150: Look and Say Sequence: Describe the Previous Term! (Method A - Initial Term Is 1). (Formerly M4780)," *The On-Line Encyclopedia of Integer Sequences*, published n.d. https://oeis.org/A005150; Rick L. Sheperd, "Entry A249572: Least Positive Integer Whose Decimal Digits Divide the Plane Into N+1 Regions. Equivalently, Least Positive Integer with N Holes in Its Decimal Digits," *The On-Line Encyclopedia of Integer Sequences*, published Nov. 1, 2014. https://oeis.org/A249572; Julia Zimmerman, "Entry A363054: Look and say sequence: describe the previous term (method A, starting with 20)," *The On-Line Encyclopedia of Integer Sequences*, published May 15, 2023. https://oeis.org/A363054

68: Maybe with enough data, every mathematical operation can be flattened into 1D. Since much of mathematics is propositional, this does not seem obviously implausible.

69: Answer to prompt, "what does it mean that 3 is between 2 and 4?" provided by us in one conversation.

70: An example of sensory information that is too well-known, too accessible diegetically for ChatGPT for us to make use of in this case is shown in Fig. 6.

X. Fuzzing Up Frequency and Truth

With respect to sequences, we mentioned that ChatGPT has an easier time with common formulas and common mathematical relationships. This is something other researchers have noted, too: "LLMs excel in remembering the results of common equations, such as the square of π" while "for equations that are uncommon, GPT-3 may hallucinate a false answer."[71] Is this because the faculty of 'common sense'—which ChatGPT seems closer to having than mathematical logic—really does, as the name implies, have to do with frequency and exposure? Is this related to why certain kinds of information—perhaps like semantic meaning—seem to be derivable more quickly than other kinds of information—perhaps like the character-level information learned by LLMs? In other words, why do some conclusions—with the same number of steps—seem more obvious than others?

Most of what ChatGPT knows, and what it is closest to mastering, has to do with how to form grammatical, felicitous—more or less normal—utterances. Fundamentally, ChatGPT has been trained by being exposed to a lot of text. The rules that tell you whether one symbol can appear next to another in language are significantly different from the rules that tell you what symbols can come next to each other in mathematics. Probability is a good heuristic for language; people probably do say 'the sky is blue' more often than they say 'the sky is red.' Both are grammatical, but the more common is also the more likely to be true. Fuzzing grammaticality and truth together does, however, cause problems, but it is not the worst strategy for teaching a computer to speak a human language while at the same time teaching it about the world, especially given the way the text ChatGPT was trained on came to be—that is, most of the text it was trained on was made by people for other people, with some kind of purpose in mind. This strategy may work out for common, popular equations, but it does not hold for mathematics in general. Given 'x' in an equation, a great many symbols could come next. That we overload the meaning of adjacency but have set ChatGPT up for exactly the kind of adjacency

71: Wang *et al.*, "Bot or Human?" 7.

(most often) found in language and given it training data that is mostly natural language is part of why ChatGPT struggles in contexts beyond that scope—for example, in contexts like mathematics and regular expressions.

XI. Symbols

A. Packages and Contents

When language involves multiple interlocutors, there must be some physical medium between the language and the entities involved. The language itself conveys meaning inside its structures, but due to the nature of transmission, there is information in the package the language comes in as well.

B. Arbitrary Leaps

Part of the way that the technology of language—of symbols—works is that the form of the symbol is partially independent of its meaning. According to Chomsky's Principles and Parameters theory, "knowledge of language consists of universal constraints, a set of abstract features that may be realized in different languages in an arbitrary set of morpho-syntactic or morpho-phonological ways (e.g., Case and Agreement), a universal interpretive component (Logical Form, LF), a phonological component (Phonological Form, PF), and a lexicon."[72]

As laid out by Kaushal and Mahowald, since word embeddings represent co-occurrence information—typically considered semantic—if the relationship between forms and meanings is truly arbitrary, there should be no character-level information discoverable by the LLM. However, the symbols of language are not entirely arbitrary with respect to their meaning (*e.g.*, onomatopoeia and related patterns, like *fl*-words in English—flutter, flap, flicker—having to do with movement): "[T]here are statistically detectable non-arbitrary form-meaning relationships in language."[73]

72: Juffs and Rodríguez, *Second Language Sentence Processing*, 3.

73: Kaushal and Mahowald, "What Do Tokens Know About Their Characters and How Do They Know It?" 6. These are diegetically reachable for an LLM—at least partially—though they would

The outside (medium, container) and inside (message, contents, meaning) of a symbol cannot be identical, or else it would not be a symbol—it would not 'stand for' anything.[74] This means, given access to only the meanings of words, the exact form—how they might sound when pronounced or look when written—cannot be completely recovered.[75] For there to be both supradiegetic and diegetic information encoded in language, any degree of arbitrariness, no matter how slim, is sufficient. The symbols are arbitrary enough: There is information in the supradiegetic layer that is not derivable from purely diegetic information.

C. Diegetic Boundaries

Returning to an earlier question, we want to know, given the structural constraints placed upon its 'universe,' what ChatGPT could become. More precisely, we want to know what abilities beyond linguistic fluency might manifest were ChatGPT under slightly different paraments. We have already seen that giving an LLM an enormously large, yet finite, amount of linguistic data of a certain kind—what we have been calling diegetic—is enough to inculcate fluency in the English language. However, if one were to train it on an arbitrarily large set of data—perhaps still diegetic if, as seems to be the case, supradiegetic information eludes it—would different characteristics manifest? Would we see not merely a quantitative shift in its 'abilities,' but a qualitative one as well? Given its structure—its architecture, its mind and body equivalents—is ChatGPT locked, only capable of quantitative change? Further, if, as noted above, language is an embodied task and embodiment as such may be required for a certain kind of 'intelligence,' what does that mean for ChatGPT? How does its body of artificial hardware and a vast electronic apparatus affect its cognition? How dependent is gnogeography—the abstract geography of knowledge—on the physical form it belongs to? While full answers to these questions certainly elude us—especially if we take Nagel literally—we hope to approach them, if only asymptotically.

not usually be considered part of the semantic meaning of the word.

74: This is why it is eventually, at the limit, impossible to 'detect' whether a piece of text came from a human or non-human source. That information is not encoded in the language itself, or else language would not work, it could not bridge so many worlds. That does not mean there may not be detectable patterns to speech generated by ChatGPT and a person that could be used to guess from whence it came, but any such pattern is subject to change—especially in light of Goodhart's Law—and is not *proof* of the source. [Editor's Note: Goodhart's Law says that "[A]ny observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes." Charles Goodhart, "Problems of Monetary Management: The U.K. Experience," in *Inflation, Depression and Economic Policy in the West*, ed., Anthony S. Courakis, 111–143 (London and Oxford: Mansell Publishing and Alexanderine Press, 1981), 116.]

75: If you know perfectly well what a *cat* is, you do not necessarily know that it is called a 'cat' in English or 'gato' in Spanish, nor can you infer those forms with any certainty given your knowledge of what cats are.

D. Flatland

Approximations of extradiegetic information can be provided diegetically as descriptions (as in Fig. 6) or as rules and instructions. To understand how those compare, we can use a set-theory-based analogy in the Sapir-Whorf-like style of *Flatland*.[76]

Consider the set {1,2}. Imagine if your whole universe consisted of that set and the ability, to some extent, to one-dimensionally concatenate those symbols. You, a creative being, might start making your own structures out of the things available to you, things like 12, 21, 12221212, etc. There are infinite ways you can express yourself. But imagine that the universe of your friend is {1,2,3}. Even though, for every unique thing they can say, you can say something novel too, your structures utilize the same symbols more often—for example, they may say things like 11, 12, 13, 21, 31, etc. while you say 11, 12, 21, etc. Although you can produce a string to represent anything you might want to say, you have no way of reaching the symbol '3.' It is out of your grasp. Now, imagine your universe being augmented with a new symbol so now your building blocks are {1,2,3}. In some ways, your universe feels similar; for example, it is still finite in size, and equally spacious. In your first universe, 1212212111 could easily be generated as a random string. This is what your random looked like, sequences of '1's and '2's. In your new universe, that string looks less random, it looks repetitive. While you did not feel like your old universe was too small when you were in it, by comparison to what you can say now, it seems limited.

76: See Edwin Abbott, *Flatland: A Romance of Many Dimensions* (Princeton: Princeton University Press, 2015).

Going further, imagine someone gave you a new symbol which is, in fact, an operator: '+.' This symbol lets you combine symbols you know already to get symbols you have never seen before. You went from {1,2} to {1,2,3} without any ability to get to '3' from '1' and '2.' From inside each universe, they seemed equally complete. Now, though, given {1,2,+} or {1,2,3,+}—there is no significant difference—you can expand your universe yourself, with no end. You live in ℤ+!

To tie this analogy to the rest of the paper, we can imagine several strategies for expanding such a universe when it comes to ChatGPT. In increasing order of apparent difficulty: One would be to compensate for missing information with additional diegetic material, like symmetry groups of Latin characters. This is akin to having {1,2} and being given '3.' Another would be to build functionality into the model's architecture allowing the same diegetic starting place to span more ground. This is like having {1,2} and being given '+.' Another would be to try to expand the bounds of what ChatGPT can experience. This would be like always having been blind and gaining the ability to see. For the denizens of the world of Flatland, it would be like gaining access to a new dimension.

The sets {1,2} or {1,2,3} are like ChatGPT's training data and whatever ChatGPT learned and memorized from it. The operator '+' is like functionality that OpenAI has added on top of ChatGPT's functionality as an LLM (hard coded rules). An example would be additional software that helps ChatGPT to deal with typos—the model did not learn that from the training data, but OpenAI, seeing that functionality was needed, was able to patch it on top of the existing architecture. ChatGPT, the way it is currently built, could (probably) not modify its own architecture no matter how much training data it was exposed to. Training data, if you imagine it as an ideal object, could consist of every possible utterance in the English language, so ChatGPT could learn all of those utterances. None of those utterances, however, would actually include the insertion of a new file into ChatGPT's architecture, so it would remain inaccessible to ChatGPT, even in an idealized scenario. In our set theory analogy, we could dream up every possible sequence of '1's and '2's, but nothing in that milieu would prepare us for addition, to add '1' to '2' to get '3.' As long as information can be supplied as training data, ChatGPT has what it needs to incorporate that. But if the information cannot be conveyed via 1D relations of context between symbols—such as how to grow its own eyeballs—then it remains elusive.

In all these examples, your universe is still limited. There are things outside of ℤ+ that you still cannot reach. The universe you can reach is *diegetic*, and what exists but is unreachable is the *extradiegetic*. A proper subset of the extradiegetic—for ChatGPT as it exists now—is the *supradiegetic* linguistic information that is more or less stripped away as ChatGPT builds up its internal universe of vectors.

E. Ergodicity and Span

With the *Flatland* analogy, we note the distinction between ergodicity and span. In a subspace of linear algebra, the eigenvectors span the space; they are like the prime numbers that provide the building blocks for every item in ℤ+ (under multiplication). However, if you were to look at a set of eigenvectors, you would not necessarily understand every possible position in that space. Similarly, the conclusions that *can* be drawn from an initial set of axioms and logical rules for licensing conclusions is not the same set that *has* been drawn to date or that *will* be drawn by any one person. What you can and will get to are different, both individually and cumulatively at any given time or place.

It is the case that "finite devices—physical symbol systems—permit an infinite behavioural potential."[77] But it is evident that each of us does not exploit that entire space. When thinking of ChatGPT, the operations of proximity and adjacency in 1D—the diegetic bits of language—seem to get you semantic meaning and linguistic fluency fairly quickly. But different pieces of that fluency

77: Michael R.W. Dawson, *Mind, Body, World: Foundations of Cognitive Science* (Edmonton: Athabasca University Press, 2013), 55.

emerge over time, and not always for clear reasons. Why do certain things come more quickly than others? What will ChatGPT be able to do in its lifetime? What will a lot of similar LLMs be able to do? And how far does the apparent linguistic fluency extend? How many of the wide variety of things we think of as being encoded in language—"social dynamics between people" such as power differentials and biases—can be diegetically accessible for an LLM?[78]

We both learn things individually and accumulatively, as a group with history; similarly, "there are three time frames at which we can study behavior: 'here and now'; learning and development; and evolution."[79] ChatGPT is an extension of both individual and cumulative knowledge acquisition. Many now-familiar technologies we have made, such as books, have been constrained to storing extant knowledge, but that is not necessarily the case with computation.

LLMs seem to acquire a significant degree of syntactic and semantic knowledge faster than they learn similarly complete information about some characters. They do learn about the characters, but more slowly. What is the shape of the diegetic landscape of the model's interior world? How do we know what is near the core, and what is at the border? With more and more textual input, would the boundary expand forever? Some things you need a lot of data or processing to learn; but technically they are just as licensed. For example, certain LLMs "can take advantage of character-level information in order to solve wordplay tasks like unscrambling scrambled words" and spelling tasks that require mapping "from words to characters (e.g., from *cat* to the characters $c + a + t$)," even though "word pieces have no explicit access to character information during training, and the mechanism by which they acquire such information is not obvious."[80] How these abilities emerge "could be of interest not just in NLP, but [to many fields] in the

78: Antoniak *et al.*, "RIVETER," 1.
79: Pfeifer and Bongard, *How the Body Shapes the Way We Think*, xx.
80: Kaushal and Mahowald, "What Do Tokens Know About Their Characters and How Do They Know It?" 2.

cognitive sciences."[81] Understanding these processes could yield insight into longstanding questions in historical linguistics as well: Is the rate of change in language observed in people—as in lexicostatistics—related to how quickly derivable different bits of information are as we learn a language originally, and as we learn additional information through language post-fluency? For example, in onomastics, toponyms and personal names are often really good at preserving pieces of older languages and cultures that are otherwise no longer directly relevant, such as in the theophoric name 'Michael.'

XII. Returning to Sumerian Cuneiform

We noticed when talking to ChatGPT about Sumer that its responses seemed unusually repetitive. A lot of what it says, though relevant, has to do with only a few topics. There tend to be mentions of An, the dingir symbol, and kingship, which make sense given what artifacts are attested and studied (one of the most prominent texts is the Sumerian King List). We think this might be explained by the analogy of the universe of $\{1,2\}$. For Sumer, ChatGPT's universe is small (compared to, for example, the universe of English or America). ChatGPT, from inside the universe, cannot tell that is the case, however. For other topics, maybe it has something more like $\mathbb{Z}+$. Its output about Sumer that sounds like "kings, An, dingir, lugal, reeds, Uruk, cuneiform" (see Fig. 2 and Fig.3), sounds to us like the strings 121122121 and 212122121 when we know about far more numbers—similar, repetitive—but ChatGPT is unaware.

XIII. Approximations: Only the Thing Is the Thing

Of course, approximations can be made, but they are within that world; they are made with the building blocks diegetically available. They work by describing something extradiegetic in a diegetic manner.

On the one hand, only the thing is the thing itself, so a repre-

81: Ibid., 9.

sentation of something in one format must be different than the representation of that thing in another format.[82] It is trivially true that I cannot know what it is like to be a bat, any bat—*that* bat, for example—because I am myself, which is not that bat.[83] On the other hand, we can empathize. I am not you, but I can learn enough about what it is like to be you for us to be getting on with things. Information does get from my *idios kosmos* to yours by technologies like language.

To try to imagine what it is like to be ChatGPT is to try to borrow something very foreign, like Pratchett's Granny Weatherwax borrowing a hive of bees. We know ChatGPT does not have human eyes, ears, or a human mind or body, so of course it is true that ChatGPT is not a human and cannot do exactly what a human can. That does not mean that ChatGPT cannot do, *more or less*, what a human can—*e.g.*, ChatGPT does not speak English exactly the way I do, but neither does anyone else. We can be certain that, when exposed to the same textual input, ChatGPT and a person are not granted access to equivalent supradiegetic linguistic information. It is not obvious how well that missing information can be approximated diegetically, however, although it seems plausible to think that, with enough diegetic approximations of extradiegetic information, eventually a model like ChatGPT could reach a downstream universe of conclusions and thoughts functionally indistinguishable from those a person could reach from the same text.[84] Approximations of supradiegetic information can be provided diegetically such as, 'the word *bouba* sounds round and the word *kiki* sounds spiky' (see Fig. 6).[85] We know that is not identical to our experience of sensually perceiving those words in the literal sense that we are not ChatGPT, but we also know this is true because the human brain involves multiple modalities outside of the purely descriptive, in both sensory perception and cognition.[86]

82: A paraphrase of George Box's famous saying, "all models are wrong."
83: See Nagel, "What Is It Like to Be a Bat?"
84: Li *et al.*, "Large Language Models Converge on BrainLike Word Representations."
85: See V. Ramachandran and E.M. Hubbard, "Synaesthesia—A Window Into Perception, Thought and Language," *Journal of Consciousness Studies* 8, no. 12 (2001): 3–34.
86: For another example, consider the text 'cat.' ChatGPT seems to know, more or less, what 'cat' means. When we encounter 'cat,' however, we also (typically) encounter the clicky, hard sound and its short appearance. As people who can hear, read, and write, we get 'cat's supradiegetic properties. We can see that the letter 'a' encloses area; ChatGPT cannot. ChatGPT may *know* that it encloses area if that information has been provided or is reachable diegetically with a statement like, 'the symbol "a" has a circle and a small tail. A circle is a closed loop. Any thing deformable to a closed loop encloses area when drawn,' for example. But ChatGPT still can't see the letter, so there is still information we are getting which it is not. We can imagine squishing 'a' around so that it looks like 'Q'—that is reachable in our universe, but not necessarily in ChatGPT's. The visual and auditory supradiegetic linguistic information of words are fairly easy to point to to convey this difference in experience between ChatGPT and us. But the differences are much broader: We also have had the experiences of petting a cat's fur, of loving a pet cat, of cooing over a kitten, etc. See Pearson and Kosslyn, "The Heterogeneity of Mental Representation."

XIV. Gricean Cooperation

An impression that stood out in working with ChatGPT is that, when talking to a person, you usually work towards a better shared understanding, Griceanly, by attributing good faith to your partner; the more you talk to ChatGPT, however, the clearer it becomes that there is less underlying consensus being built between you two than you would expect.[87] The 'shared universe' is less reliable and firm than you would guess from the apparent fluency of the conversation, especially by way of comparison with what we are used to when speaking to other people. Often, the beginning or middle of the conversation is impressive, and you think, '*there is something here!*' but if you keep exploring—keep probing, keep digging—that impression inevitably falls apart. This experience, right now, is one of the most qualitatively different aspects of talking with ChatGPT.

We are excited for LLMs to help us with all sorts of things, from coding to etymological trees across languages to searching for unconsidered patterns. However, right now, we think it is helpful to think of how different ChatGPT's experience of linguistic information is when trying to understand how it behaves, especially since it can seem comfortably familiar and competent in one moment and incomprehensible in the next.

XV. Where Can We Go?

Although diegetic frameworks of many kinds are longstanding, the way ChatGPT and similar models experience language highlights a difference from our own that is not often cast in direct relief. In this view, *diegetic* means what can be conveyed by 1D relations of context, statistical patterns of symbols—the message,

87: See Grice, *Studies in the Way of Words.*

without the medium. For an LLM, correlation actually is causation since it is by frequencies of proximity and adjacency that any information is embedded within them. Something in more dimensions can be folded down into far fewer, and some of that larger structure is learnable even for something like an LLM which is very limited in terms of its senses and modalities, for example. In addition, systems like mathematical logic can be imposed atop the ordinary linguistic semantic meaning of symbols, overriding or modifying their meaning extradiegetically. Approximations of these dimen-

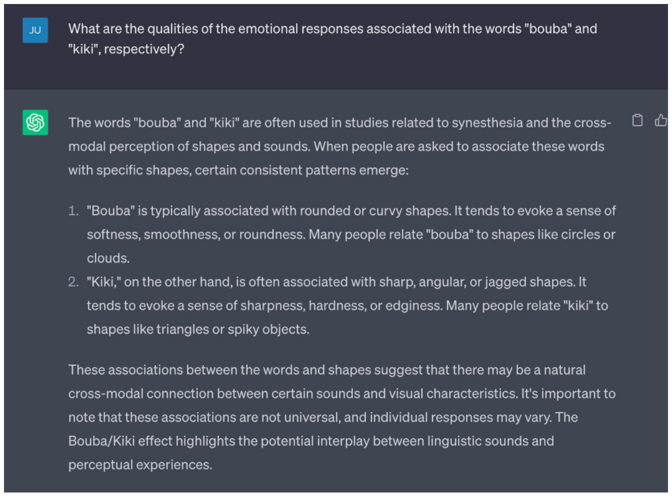


Figure 6. ChatGPT can, at least sometimes, make use of descriptive, diegetic approximations of sensory experiences it has no direct access to.

sions can be provided diegetically or patched with additional modules of functionality in models like ChatGPT.

A. Postlapsarian

It might be that the technology of language itself—requiring an at least partially arbitrary leap of faith between form and meaning—opened Pandora's box. Something that had no specific meaning before—a sequence of sounds or symbols, which could initially have been any sequence of sounds or symbols—can, through its use by people, become incredibly—even enduringly, as most words

will live far longer than any person could—meaningful and important for much of social life and thought. The fall of man could be the discovery of meaningfulness out of meaninglessness (something out of nothing should perhaps remain the purview of gods)! In an additional twist of magic, the meaninglessness that gives rise to meaning is not just incidental, but necessary: Meaning comes from contrast; difference opens up the space that is needed for one thing to point to another. For symbols to be useful, there must be a distinction between what they are and what they mean, and from some perspectives, that difference is arbitrary; just like for any metaphor, there is a diegetic framework—a perspective, an imagined world—in which the pieces being compared are identical. The other kinds of information made salient by how humans typically experience the symbols of language are frequent fodder of conspiracy-style 'baking,' as in gematria and any-mancy; these dimensions of meaning are decoupled from, and can therefore be exploited in parallel to—without negating or contradicting—the more intrinsic semantic meaning of the symbols. An arbitrary leap of faith, once required and even proved productive, is a dangerous precedent.

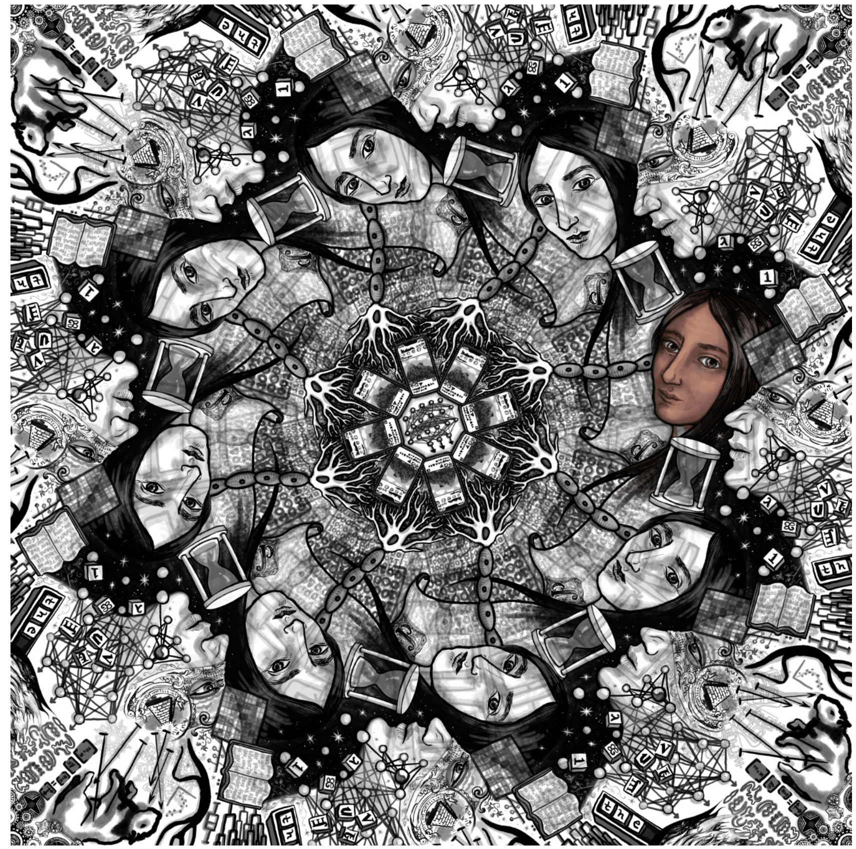
B. What Things Can ChatGPT Bring to The Table?

ChatGPT, though not currently well-equipped for this task, may have unique, novel, and valuable experiences and ways of being that lead to new insights: to a ChatGPT-specific form of Langton's "intelligence as it could be." How can ChatGPT's experiences—its senses and body—enable new thoughts, new representations, new processes of cognition? For example, an instance of ChatGPT exists within each 'chat,' within each user account. Some information may flow back to the central code base, but even if it does not, changes are made over time to the code base and are then deployed as a universal update to all instances of ChatGPT simultaneously. This is like telepathy, something like a hive mind broadcast that still allows significant independent decision-making and analysis on behalf of each individual; it is a kind of distributed thinking together. Dipping a toe into what it might be like to be ChatGPT suggests a reframing of the recent advances in AI as an extension of us, an update to people, another wave in the technologies rippling out of mechanization, electricity, computers, writing, institutions of higher learning, etc. It can prompt us to look backwards as well: If ChatGPT is an extension of our cumulative knowledge—our collective mind—then we should consider what information it will make salient to us—for example, if we use it to write snippets of code. We may not have been doing enough accounting for the complex processes by which we encounter the ideas of others, such as our increasing reliance on Google searches. With this new Generative AI technology, we are just making more things from things we all made via a new kind of tool. ChatGPT—and similar technologies—can make beautiful things (and they can make garbage) just like we can because they—like all technologies—are an extension of our own minds and bodies, and our own minds and bodies are an extension of the minds and bodies that came before us.[88]

Acknowledgments

The authors wish to express their appreciation for the works of Timothy Sprigge, Noam Chomsky, Chris Langton, George Box, the QAnon Anonymous podcast (Travis View, Jake Rockatansky, Julian Feeld, Annie Kelly, Liv Agar, etc.), and to conversations with Alejandro Ruiz, Michael Arnold, Yoshi Bird, Brad Demarest, Juniper Lovato, Carter Ward, Aviral Chawla, Mark Wittels, Grace Taylor, Julia Chimienti, Evan Lynch, Cecile Smith, Desi Alexander, Peter Heft, and Professors Randall Harp, Jacques Bailly, Guillermo Rodriguez, and Josh Bongard. The authors are grateful for support furnished by MassMutual, Google, and the National Science Foundation (Grant #2242829).

88: Andy Clark and David Chalmers, "The Extended Mind," *Analysis* 58, no. 1 (1998): 7–19. https://www.jstor.org/stable/3328150 It ought to be noted that the general idea of a material mind has roots as far back as Ancient Greek philosophy.



[**Scarcity, Abundance, and Optimization in our AI Extended Society**
Julia Witte Zimmerman / digital drawing / 2024].