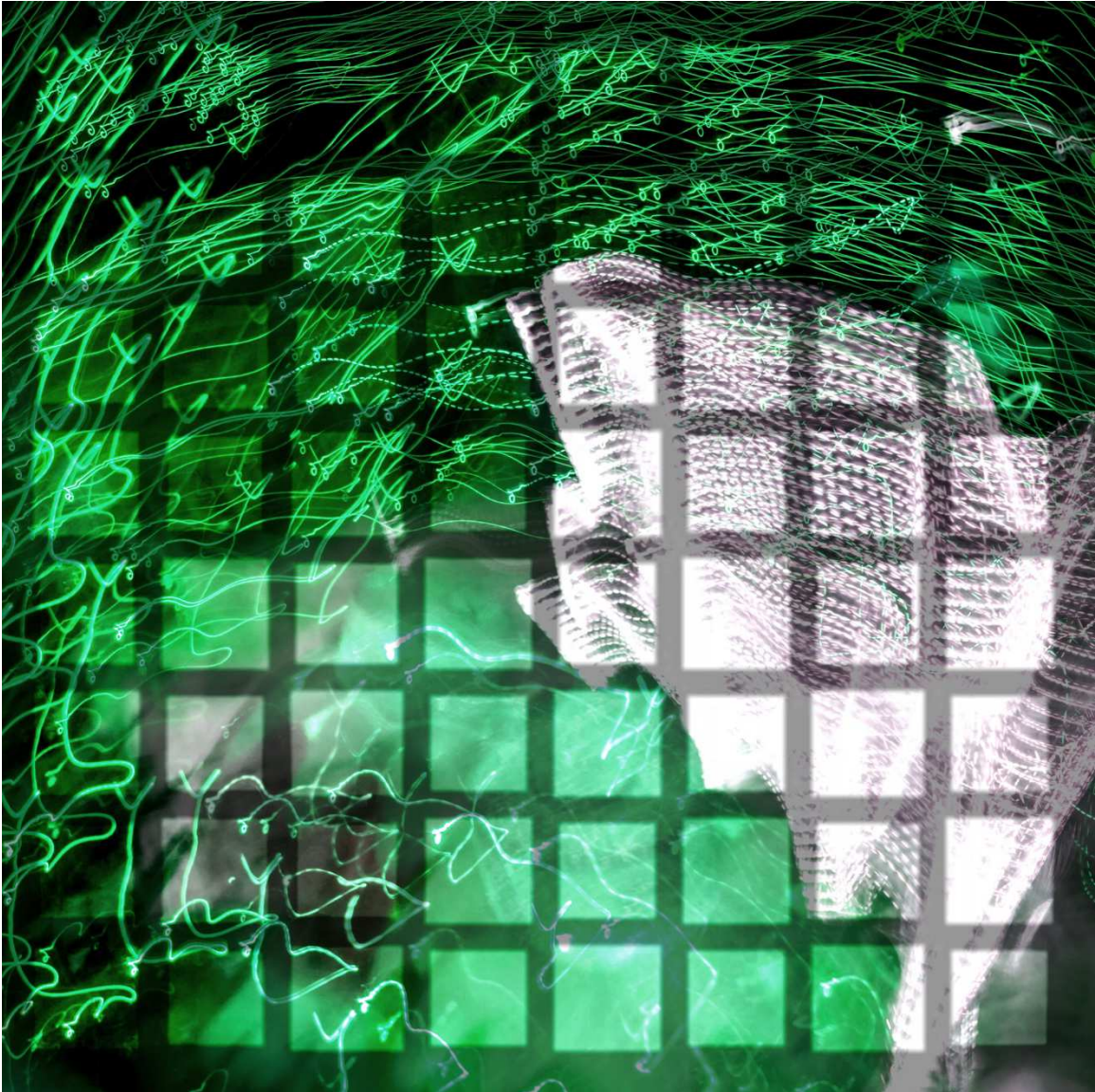


ChatGPT intimates a tantalizing future,
its core LLM is organized on multiple levels,
and it has broken the idea of thinking.



A preliminary report, Version 3
February 6, 2023
William Bazon

ChatGPT intimates a tantalizing future; its core LLM is organized on multiple levels; and it has broken the idea of thinking.

Version 3
William Benzon
February 6, 2023

Abstract: I make three arguments. A philosophical argument: (1) The behavior of ChatGPT is so sophisticated that the ordinary concept of thinking is no longer useful in distinguishing between human behavior and the ChatGPT's behavior. We don't have explicit understanding about what either humans or ChatGPT are doing. Two operational arguments: (2) Having examined its output in a systematic way, short stories in particular, I conclude that inference is organized on at least two levels: a) a 'lower' level where we find sentence-level syntax, and b) a 'higher' level where specific kinds of texts, such as stories, are implemented *over* and operate *on* sentences. This is roughly analogous to the way that high-level programming languages are implemented in assembly code. (3) Consequently, that aspects of full symbolic computation are *latent* in LLMs. An appendix has descriptive tables showing how four stories are organized on multiple levels.

Contents

0. What is in here and why	2
1. The Chinese Room is empty	3
2. ChatGPT, justice (an abstract idea), and interpretation (analogy)	4
3. Tell me a story	6
4. Here be dragons: The idea of thought is broken	10
5. What underlying structure drives ChatGPT's behavior?	11
6. Where are we?	15
7. Appendix 1: ChatGPT derives new stories from old	18
8. Appendix 2: Texts considered as 1-dimensional visual objects	23
9. Appendix 3: My work with ChatGPT	23

Note: In this version, sections 5 and 6 are significantly different from and supercede those sections in the previous versions. There are other minor changes.

1301 Washington St., Apt. 311
Hoboken, NJ 07030
646.599.3232
bbenzon@mindspring.com



This work is licensed under a Creative Commons Attribution-Share Alike 3.0 Unported License.

0. What is in here and why

After having worked with ChatGPT intensely for a month and a half I am all but convinced that the underlying language model is organized on at least two levels: 1) There is a “bottom” level consisting of 175 billion weighted parameters organized into layers. This structure is generally considered opaque. 2) There is a higher level that seems all but invisible; the bottom level implements this higher level and it is this higher level that organizes discourse with users. We can infer this higher level by analyzing the structure of texts. I have been analyzing and describing (some of) the manifestations of this higher level as a necessary precursor to characterizing it as a “grammar” or computational structure.

That is one argument I make. I interleave it with another one, that ChatGPT is so successful in enacting a rich and varied simulacrum of human behavior that philosophical arguments asserting that computers cannot, in principle, think, such arguments are rapidly losing their cogency. They may remain valid, but they tell us little about what humans can do that computers can't.



1. *Why the Chinese Room is empty* – Searle’s argument, and others like it, got much of its rhetorical force from the fact that AI systems at the time were quite limited in their capacity. That is no longer the case. Those arguments are rapidly losing their potency.

2. *ChatGPT, justice (an abstract idea), and interpretation (analogy)* – Analogical reasoning, in the form of interpreting texts (movies) and reasoning about abstract concepts (justice).

3. *Tell me a story* – I use a procedure derived from Lévi-Strauss’s work on myth to probe GPT’s capacity for story-telling.

4. *Here be dragons: The idea of thought is broken* – Advanced AI is unlike anything we’ve seen before. These systems are not human, are not sentient, yet their behavior appears human. We need concepts (with new terms) for understanding them.

5. *What underlying structure drives ChatGPT’s behavior?* – There is the neural net, yes. But the neural net operates on a ‘lower’ level where it supports syntax and a ‘higher’ level where it supports discourse.

6. *Where are we?* – A reprise of sections 1 through 5 ending with the suggestion that symbolic structures are *latent* in ChatGPT.

7. *Appendix 1: ChatGPT derives new stories from old* – Tables associated with the story mechanisms described in section 3.

8. *Appendix 2: Texts considered as 1-dimensional visual objects* – A thought experiment in which texts are transformed into strings of colored beads.

9. *Appendix 3: My work with ChatGPT* – Links to my work, blog posts, working papers, on ChatGPT, GPT, and deep learning.

1. The Chinese Room is empty

I believe that the classical arguments against AI, the arguments Dreyfus and Searle (Chinese Room¹) made based on the concept of intention, remain valid. But the context has changed radically. When those arguments were originally advanced AI technology couldn't do anything that would appear impressive to the lay person. The ARPA Speech Understanding Project² of the mid-1970s, for example, was a massive and impressive project spread across three or four institutions and a half dozen sites for three years. One of those systems required a day to process a line of spoken input. Yes, it was impressive to those who built it because they understood the difficulty involved in creating the systems. Those difficulties meant nothing to outsiders.

In retrospect it seems clear that those classical arguments got most of their *rhetorical* force, as opposed to their *cognitive* force, from the impoverished abilities of contemporary technology. While the sophisticated performance of modern AI technology does not diminish the cognitive force of those intentional arguments, it does erode their rhetorical force. The arguments turn on a subtle philosophical point that wilts in the face of the impressive behavior of systems such as ChatGPT.

It tells me things I didn't know in language I can understand, uses coherent paragraphs and correct sentences, and you're telling me it's just a stochastic parrot.²³ Nah, I don't believe you. I believe what I read on the computer screen. You don't get to tell me my eyes are lyin'.

As a logical argument, that's not very good. But the superficial force of ChatGPT's performance is difficult for people to overlook. By leaning on their intellectual authority to insist that there's nothing behind the curtain, critics of AI threaten to undermine their own credibility by forcing people to conclude that they, themselves, are credulous and stupid.

Correlatively, when the AI boosters insist that ChatGPT has passed the Turing Test and we're on the way to artificial general intelligence (AGI), that assertion is consistent with the initial impression that ChatGPT, or some other advanced AI, really is thinking.

It tells me useful things, intelligent things. And when the technology improves, it's going to be even more intelligent and useful. Yippie! We're on the way!

¹ John. R. Searle, Minds, brains, and programs, *Behavioral and Brain Sciences* 3 (3): 1980, pp. 417-457. I have written a number of posts in which I discuss Searle's Chinese Room argument, <https://new-savanna.blogspot.com/search/label/Chinese-room>.

² Dennis H. Klatt, Review of the ARPA Speech Understanding Project, *The Journal of the Acoustical Society of America* 62, 1345 (1977); <https://doi.org/10.1121/1.381666>.

³ The phrase "stochastic parrot" is from a well-known paper, Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell, On the Dangers of Stochastic Parrots, Can Language Models Be Too Big? *FAccT '21*, March 3–10, 2021, Virtual Event, Canada. More recently Gary Marcus asserted that, while there is a stochastic element involved, what ChatGPT does is more like pastiche. That's not much of an improvement. See, Gary Marcus, How come GPT can seem so brilliant one minute and so breathtakingly dumb the next?, *The Road to AI We Can Trust*, Dec. 1, 2022, <https://garymarcus.substack.com/p/how-come-gpt-can-seem-so-brilliant>.

That's not much of an argument either, but by admitting its validity, the AI boosters align themselves with the superficial impressions of naïve users and therefore reinforce the confidence those users have in their own judgment.

We will revisit this question: Does ChatGPT think? But for now, I will talk about ChatGPT *as though* it thought and reasoned like humans do. Do not take that usage as evidence that I believe ChatGPT to be thinking, or to be sentient. I don't believe those things. But I don't have concepts and terms that adequately characterize my belief. So I use *terms that are convenient*, but not prematurely reductive, as the ordinary vocabulary of computation would be, or obscure, as neologisms would be.

2. ChatGPT, justice (an abstract idea), and interpretation (analogy)

Besides, I think that there IS something behind the curtain. Whatever ChatGPT may be – that is not at all certain – it is not a stochastic parrot. In one of my sessions with ChatGPT I had it:⁴

- define justice, an abstract idea,
- explain how Plato reasoned about justice in *Republic*,
- critique a short story depicting an injustice, and
- revise that story so that justice was served.

Here is how it defined justice:

Justice is the concept of fairness and equality. In the legal context, it refers to the fair and proper administration of laws. In a broader sense, justice is the ideal state in which individuals are treated equally and without prejudice, and in which the laws and institutions of society are fair and just. It is often considered to be one of the fundamental principles of society, and it is an important concept in many ethical and philosophical systems.

That seems reasonable. Of course you could argue that the definition is just assembled from material readily available on the web, and the same for its three-paragraph account of Plato's *Republic*. Since people do that kind of thing all the time, it is not clear what kind of objection that is.

The fact that it was able to critique a story I'd given it, to judge that justice was not served, and then to explain why, that is a performance of a different kind. Here is how ChatGPT reasoned:

⁴ William Benzon, *Discursive Competence in ChatGPT, Part 1: Talking with Dragons*, Version 2, Working Paper, January 11, 2023, pp. 42-46, https://www.academia.edu/94409729/Discursive_Competence_in_ChatGPT_Part_1_Talking_with_Dragons.

Based on the story, it does not seem that justice is being served. The hero performed a selfless and courageous act by risking his own life to save others, but instead of being recognized and thanked for his actions, he was ignored and overshadowed by the media and public's focus on the wealthy CEO, who did not do anything to help. This is not fair to the hero, as he deserves to be recognized and celebrated for his bravery and the difference he made. The hero's true reward of knowing that he saved lives is important, but it is not a substitute for the recognition and appreciation that he deserves.

To be sure, the story I gave it was one ChatGPT had told in an earlier session, so it was “cut” to its capabilities. But ChatGPT doesn’t remember things from one session to the next. It confronted this story as something new, took its understanding of justice and applied it to that story in a meaningful way. Finally, when I asked it to, it was able to revise the story so that justice was served. That is just one example of many that I have.

One of the first things I had ChatGPT do, was to use Rene Girard’s ideas about mimetic desire and sacrifice to interpret Steven Spielberg’s *Jaws*.⁵ ChatGPT took abstract ideas and applied them to a story. I’ve had it do the same thing⁶ with other movie and texts. That looks like thinking to me. That is to say, when humans do that – say, my students from the days I taught college literature – we regard it as thinking. Why try to deny that ChatGPT is thinking? If it isn’t thinking, then what IS it doing?

Moreover that kind of thinking is analogical, a fundamental and pervasive mode of thought. To interpret a text (a movie is a kind of text) you set up an analogy between the actors and events in the text and the general specifications in the interpretive instrument (such as Girard’s ideas, or Marxist, psychoanalytic, feminist, etc. ideas). If the match is good, if the analogy holds, then one has an interpretation. Heretofore it was generally believed, and generally true, that AI cannot do analogy. GPT-3, on which ChatGPT is based, has changed that.⁷

⁵ Conversing with ChatGPT about Jaws, Mimetic Desire, and Sacrifice, *3Quarks Daily*, December 5, 2022, <https://3quarksdaily.com/3quarksdaily/2022/12/conversing-with-chatgpt-about-jaws-mimetic-desire-and-sacrifice.html>.

⁶ Posts where I have ChatGPT interpret some text, <https://new-savanna.blogspot.com/search/label/ChatGPT%20interpretation>.

⁷ For another article verifying GPT-3’s capacity for analogical reasoning, see Taylor Webb, Keith J. Holyoak, and Hongjing Lu, Emergent Analogical Reasoning in Large Language Models, Dec. 19, 2022, *arXiv*, <https://arxiv.org/abs/2212.09196v1>. In my view, the fact that they worked independently of me and used different methods – standard tests in cognitive science – strengthens our joint conclusion

3. Tell me a story

More recently I have been exploring the latent order implicit in ChatGPT's ability to tell simple stories.⁸ I had been corresponding with a friend from graduate school, Rich Fritzson, who had given ChatGPT the following prompt: "Tell me a story about a hero." This is the story he got (I've added the paragraph numbers for reference):

[1] Once upon a time, in a land far, far away, there was a young man named Jack who lived in a small village with his family. Jack was a brave and kind-hearted person, and he was loved by everyone in the village.

[2] One day, a great evil dragon attacked the village, breathing fire and causing destruction wherever it went. The people of the village were terrified and didn't know what to do. But Jack, being the hero that he was, didn't hesitate to step forward and face the dragon.

[3] He armed himself with a sword and shield, and set out to confront the dragon. The battle was fierce, but Jack was determined to protect his village and its people. He fought bravely, dodging the dragon's flames and striking back with his sword.

[4] Finally, after a long and difficult battle, Jack was able to slay the dragon, saving the village and its people from certain destruction. The people cheered and celebrated, grateful to Jack for his bravery and selflessness.

[5] From that day on, Jack was hailed as a hero, and his name was known throughout the land. He lived the rest of his days in peace and happiness, knowing that he had done something truly great.

I decided to do the same. Using the same prompt as Rich had used – "Tell me a story about a hero." – I asked ChatGPT to tell me a story. This is the story it gave me.

[1] Once upon a time, in a land far, far away, there was a young princess named Aurora. Aurora was a kind and gentle soul, loved by all who knew her. She had long, golden hair and sparkling blue eyes, and was known for her beautiful singing voice.

[2] One day, a terrible dragon came to the kingdom and began to terrorize the people. The dragon was fierce

⁸ Posts about ChatGPT and stories, <https://new-savanna.blogspot.com/search/label/ChatGPT%20stories>.

and powerful, and none of the knights or soldiers were able to defeat it. The people lived in fear, not knowing what to do.

[3] But Aurora was not afraid. She knew that she had to do something to save her people, and so she set out to find the dragon. She traveled for days, braving the cold and the snow, until she finally reached the dragon's lair.

[4] With her heart pounding in her chest, Aurora approached the dragon and began to sing. The dragon was enchanted by her beautiful voice, and it slowly began to calm down. Aurora continued to sing, soothing the beast until it was completely calm.

[5] With the dragon no longer a threat, Aurora returned to the kingdom as a hero. The people cheered and celebrated her bravery, and Aurora was hailed as a champion. She lived happily ever after, knowing that she had saved her kingdom and its people.

The two stories are similar. Both have five paragraphs; the paragraphs have the same functional roles in each story and they have similar wording. In both cases evil enters with a dragon in paragraph 2. But in Rich's story a man defeats the dragon with a sword while in mine a woman defeats the dragon with song. Also, in Rich's story the dragon breaths fire, but not in mine. Moreover, in my story the dragon is only pacified while in Rich's it is killed. Those differences seemed to be of a piece, reflecting a single underlying factor, such as male (Rich's story) vs. female (mine). I decided to keep experimenting.

I decided investigate the "logic" underlying ChatGPT's story telling. The procedure I've been using is derived from the analytical method Claude Lévi-Strauss employed in his magnum opus, *Mythologiques*.⁹ He started with one myth, analyzed it, and then introduced another one, very much like the first. But not quite. They are systematically different. He characterized the difference by something he called a *transformation* – drawing on a metaphor from algebraic group theory. He worked his way through hundreds of myths in this manner, each one derived from another by such a transformation.

The existence of ChatGPT allows me to use a more active procedure. I give ChatGPT a prompt consisting of two things: 1) an existing story and 2) instructions to produce another story like it except for one change, which I specify. That change, that experimental intervention, is in effect a way of triggering or specifying those "transformations" that Lévi-Strauss wrote about. What interests me are the *ensemble of things that change* in parallel with the change I have specified. I note the difference and do it again and again.¹⁰

⁹ I have found that the first volume was enough to convey the method itself: Claude Lévi-Strauss, John Weightman and Doreen Weightman, trans., *The Raw and the Cooked*, Harper Torchbooks, 1970.

¹⁰ Compare this with the procedure in Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov, Locating and Editing Factual Associations in GPT, arXiv:2202.0526v5 [cs.CL] 13 Jan

Those changes are evidence about the nature of the processes operating in ChatGPT. When you simply ask ChatGPT a question, and it answers you, no matter how impressive the answer, you have to take the answer whole. That gives you no clues about what ChatGPT is doing. The kind of experiment I have just described – and it is an experiment, albeit a modest one, one as yet without any strong predictive impetus – allows you to see how ChatGPT responds to a specific targeted “stimulus,” or, to invoke a different intellectual matrix, that of cybernetic control theory, “disturbance.”¹¹ By doing that repeatedly you collect evidence about how ChatGPT parses the world. I present tables describing four of these before-and-after transformations in Appendix 1: ChatGPT derives new stories from old.

Those tables have three columns. The middle column contains the source story while the righthand column contains the new story ChatGPT creates. The lefthand column contains labels:

1. Donné: a term from literary criticism for what is given at the beginning of a story,
2. Disturb(ance),
3. Plan: a response to the disturbance,
4. Enact,
5. Celebrate.

Those are my terms. I am using them to designate the successive phases or segments of the *trajectory* ChatGPT follows through its activation space in the course of generating the text of a story.

ChatGPT tells a story by traversing a path through its high-dimensional activation space – those 175 billion parameters. Each story’s path will traverse the same five segments. Those paths are characterized by their beginning points, which are the same for each story. Thus when ChatGPT begins moving it already has those five beginning points all “lined up,” just as a jazz musician has the sequence of chord changes, e.g. for “A Night in Tunisia,” ready and waiting when they start improvising. But the path from one beginning point to the next will likely be different for each story, as will each chorus the jazz musician improvises.

In all of the stories in this series, each transformed story has the same five components as its source. In a few cases the corresponding paragraphs in the two stories are identical. In most cases they are different. In some cases the differences are minor (changing names and pronouns), but in others they are not. Here's the important point: *The differences between the*

2023, <https://arxiv.org/pdf/2202.05262.pdf>. They intervene directly in middle-layer modules to identify where specific localized pieces of information are carried in the network. My procedure tells us nothing about where things are happening in the network, but it does tell us about relationships between items within the ‘economy’ enforced by ChatGPT’s story grammar.

¹¹ In control theory terms, the story grammar provides a “reference level” that regulates the overall content and “flow” of the story. See, e.g. William T. Powers, *Behavior: The Control of Perception*, Aldine, 1973.

two stories are a consequence of the differences between the two protagonists, who necessarily appear in the first segment. Those differences are Lévi-Strauss's transformations.

Most of the stories I have been working with derive from that story about princess Aurora. She was beautiful, as princesses are, and sang well. A dragon invaded her kingdom and she was able to calm it by singing to it.¹² In one case the protagonist of the new story was Henry the Eloquent. Note that I gave the protagonist an identifying epithet because, by that time, I had realized that ChatGPT seemed to think that characters had essences, if you will. The epithet indicated that essence. Consequently a character's *actions had to be consistent with its essence*. How would Henry the Eloquent deal with a dragon? That's obvious, isn't it? That's what ChatGPT had him do: "He used his eloquence and way with words to calm the beast, soothing it with his voice until it was completely calm." There's nothing deep or sophisticated about this, but it does seem to be well beyond the capacities of a stochastic parrot.

Now, how do you think ChatGPT will respond when asked to retell the Aurora story with William the Lazy in the driver's seat? Since the protagonist is lazy, he can't be too active in defense of his kingdom. What does he do?

[PLAN]: He thought and thought, trying to come up with a plan that wouldn't require too much effort on his part. Finally, he hit upon an idea.

[ENACT]: He summoned the bravest knights in the kingdom and gave them a special potion he had concocted. The potion would make the knights immune to the dragon's flames, allowing them to defeat the beast once and for all. The knights were skeptical, but they trusted William and drank the potion.

There were no knights in Aurora's story. ChatGPT added a new element to the story, one that allowed William the Lazy to save the day while remaining true to his lazy self.

Notice that in both cases ChatGPT had to do some simple causal reasoning. Henry the Eloquent was easy: Where Aurora sings, Henry talks. But the case of William the Lazy required a bit of invention. William had to 1) summon his knights, and 2) provide them with a protective potion (earlier in the story we were told that the knights had been unable to defeat the dragon).

Of course, some will say, *that's fine and dandy for ChatGPT, it's only a machine, it's too formulaic for human story-telling*. There is an obvious reply: *WRONG! that IS how real humans tell stories*. We have formulas, formulas that afford great variety and flexibility and variety, but they are formulas none-the-less. The locus classicus for this argument is, Albert B. Lord, *The Singer of Tales*, where Lord argued that the formulaic nature of *Iliad* and *Odyssey* reflects the origins

¹² See, Appendix 1: ChatGPT derives new stories from old, for before-and-after examples each based on the Aurora story.

of those stories in oral culture.¹³ ChatGPT thus appears to be using time-honored techniques in generating stories. If it looks like a duck and quacks like a duck, then it IS a duck.

4. Here be dragons: The idea of thought is broken

That's your argument: "Looks like a duck?"

No, it's a consideration, a bit of evidence.

Here is my problem: I actually believe those old arguments about why machines can't (possibly) think, the arguments from intention. That belief clashes with my appraisal of the behavior I've seen from ChatGPT in the last two months. *Damn it! It looks like a duck.*

How do I reconcile the two sides of that conflict? I don't. Rather, I decide to hold the concept of thought, and a whole mess of allied concepts, in abeyance. I'm going to toss them between phenomenological brackets.

The people who insist that what these large language models are doing might as well be the work of thousands of drunken monkeys pounding on typewriters, they have no way of accounting for the coherent structure in ChatGPT's output. Oh, they can pounce on its many mistakes – for it makes many – and say, “*See, I told you, drunken monkeys!*” But the AI boosters who insist that, *yes, these guys can think, we're on the way to AGI*, they can't tell us what it's going on either. All they can say is that the models are “opaque” – almost a term of art by now – *so we don't know what's going on, but we're working on it*. And indeed they are.

In this context, “think” is just a label that tells us nothing about what humans are doing that machines are not. That denial does not point the way to knowledge about how to improve these systems – for they surely need improving. I conclude, then, for certain purposes, such as discriminating between human behavior and that of advanced artificial intelligence, the idea of thought has little intellectual value.

Let me be clear. I am not denying that people think; of course we do. Nor am I asserting that advanced AI's think. They (most likely) do not. But “to think” is an informal common-sense idea. It has no technical definition.¹⁴ We are rapidly approaching an intellectual regime where the question of whether or not machines can think – reason, perceive, learn,

¹³ Albert B. Lord, *The Singer of Tales*, Cambridge, MA: Harvard University Press, 1960. On the cognitive psychology of such stories, see David C. Rubin, *Memory in Oral Traditions: The Cognitive Psychology of Epic, Ballads, and Counting-out Rhymes*, New York, Oxford: Oxford University Press, 1995.

¹⁴ Some years ago I constructed a definition of the informal concept, “to think,” within a cognitive network. See, William Benzon, Cognitive Networks and Literary Semantics, *MLN* 91: 1976, pp. 961-964. For a similar approach to the common-sense notion, see William Benzon, First Person: Neuro-Cognitive Notes on the Self in Life and in Fiction, *PsyArt: A Hyperlink Journal for Psychological Study of the Arts*, August 21, 2000, pp. 23-25, https://www.academia.edu/8331456/First_Person_Neuro-Cognitive_Notes_on_the_Self_in_Life_and_in_Fiction.

feel, etc. – becomes a tractable technical issue. In this regime, common sense ideas about minds and mentation are at best of limited value. At worst, they are useless.

I take that as a sign that we are dealing with something new, really new. We have sailed into those waters where “Here be dragons” is written on the charts. It is time that we acknowledge that we don’t know what we’re doing, that the old ideas aren’t working very well, and get on with the business of creating new ones. Let us learn to fly with and talk with the dragons. It is time to be wild.

5. What underlying structure drives ChatGPT’s behavior?

There is no greater opponent of contemporary large language models than Noam Chomsky. Here is how he articulated his basic objection at a recent debate in Montreal:

But there's a very simple and fundamental flaw that will never be remedied by such measures, in fact is exacerbated by them. Namely by virtue of their design the systems make no distinction between possible and impossible languages. And the same in other domains. The more the systems are improved the deeper the failure becomes. They will do even better with impossible languages and impossible other systems. In short, they're telling us nothing about language and thought, about cognition generally or about what it is to be human or any of the other flights of fantasy in contemporary discussion.¹⁵

I believe that is correct. Let us do a Gestalt switch on that idea. Where Chomsky sees a duck, I suggest we think of it as a rabbit.

Think of an artificial neural network (ANN), such as a large language model (LLM), as a medium for implementing certain kinds of systems, systems loosely inspired by real nervous systems. As such it is like COBOL, FORTRAN, LISP, Pascal, C++, Python, or any of a large number of languages people use create useful software applications – for that matter, we could also think of the various assembly languages used for implementing those high-level languages on different hardware platforms. Assembly language works directly with the machine, moving strings of bits around within the central processor and moving them between the processor and long-term memory. Those actions are invisible to high-level languages, which nonetheless depend on them. Similarly, the processes of high-level languages are invisible to the end user, who thinks in terms appropriate to the work they are doing – e.g. for a word processor, cut-and-paste, search, spell check, etc. Thus we have three levels of organization: 1) assembly language, 2) high-level language, and 3) end-user application. The point? These complex symbolic objects are organized on multiple levels.

¹⁵ That is from the recent debate held in Montreal: AI DEBATE 3 : The AGI Debate, Fri, Dec 23, 2022, <https://agidebate.com/>. The quoted passage starts at about 19:20 in the video of the debate: https://www.youtube.com/watch?v=JGiLz_Jx9uI&t=3288s. For a more systematic presentation of Chomsky’s views, see Gary Marcus, Noam Chomsky and GPT-3, *The Road to AI We Can Trust*, May 21, 2022, <https://garymarcus.substack.com/p/noam-chomsky-and-gpt-3>.

The stories ChatGPT generates are complex enough to require at least two levels of organization. While they are generated by a computer, these stories are not implemented in a conventional stack of computer languages. Rather, they are generated by a large language model. Programmers do not implement applications directly in LLMs. Rather, they create a software application in the ordinary way, and that application creates an LLM by “consuming” a large amount quantity of text and creating a complex statistical model that can then be used to generate new texts.

Consider this paragraph from a recent paper by Chris Olah, who is a researcher working on the problem of *mechanistic interpretability*, that is, figuring out how ANN’s do their work. Here is the first paragraph of his paper:

Mechanistic interpretability seeks to reverse engineer neural networks, similar to how one might reverse engineer a compiled binary computer program. After all, neural network parameters are in some sense a binary computer program which runs on one of the exotic virtual machines we call a neural network architecture.¹⁶

Olah’s terms and focus are different from mine – he is focused on interpreting the functions of neurons in the net – but the idea is the same.

For a different kind of example, consider the work that Kenneth Li and his colleagues have done with a GPT system specialized to play the board game, Othello.¹⁷ They have discovered that the model consists of “two systems—a powerful yet black-box neural network and a human-understandable world model [that] not only predict consistently, but also share a unified mid-stage representation.”¹⁸ It appears that the human-understandable world model is *implemented by* the neural network.

With this in mind, I want to turn to some work published Christopher D. Manning et al, in 2020.¹⁹ They investigated syntactic structures represented in BERT (Bidirectional Encoder Representations from Transformers). Early in the paper they observe:

One might expect that a machine-learning model trained to predict the next word in a text will just be a giant associational learning machine, with lots of statistics on how often the word restaurant is followed by kitchen and perhaps some basic abstracted sequence knowledge such as knowing that adjectives are

¹⁶ Chris Olah, Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases, *Transformer Circuits Thread*, June 2022, <https://transformer-circuits.pub/2022/mech-interp-essay/index.html>.

¹⁷ Kenneth Li, Aspen K. Hopkins, David Bau, et al., Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task, arXiv:2210.13382v2 [cs.LG] 25 Oct 2022, <https://arxiv.org/abs/2210.13382>.

¹⁸ Kenneth Li, Do Large Language Models learn world models or just surface statistics? *The Gradient*, Jan. 21, 2023, <https://thegradient.pub/othello/>.

¹⁹ Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy, Emergent linguistic structure in artificial neural networks trained by self-supervision, *PNAS*, Vol. 117, No. 48, June 3, 2020, pp. 30046-30054, <https://doi.org/10.1073/pnas.1907367117>.

commonly followed by nouns in English. It is not at all clear that such a system can develop interesting knowledge of the linguistic structure of whatever human language the system is trained on. Indeed, this has been the dominant perspective in linguistics, where language models have long been seen as inadequate and having no scientific interest, even when their usefulness in practical engineering applications is grudgingly accepted.

That is not what they found. They found syntax. They discovered that neural networks induce

representations of sentence structure which capture many of the notions of linguistics, including word classes (parts of speech), syntactic structure (grammatical relations or dependencies), and coreference (which mentions of an entity refer to the same entity, such as, e.g., when “she” refers back to “Rachel”). [...] Indeed, the learned encoding of a sentence to a large extent includes the information found in the parse tree structures of sentences that have been proposed by linguists.

While BERT is a different kind of language technology than GPT, it does seem reasonable to assume that ChatGPT implements syntactic structure as well. Wouldn’t that been the simplest, most parsimonious, explanation for its syntactic prowess? It would be a mistake, however, to think of story structure as just scaled-up syntactic structure.

Something else is required, something that “*operates at a higher level than sentence syntax.*” What kind of a thing is that? In his *Poetics* Aristotle talked of dramas as having a beginning, a middle, and an end. That’s a bit vague and, in any event, I see no need to go that far back. Since the early twentieth century, beginning with the work of Vladimir Propp in the 1920s, researchers in various disciplines having been searching for an abstract characterization of “story logic” if you will.²⁰ Some have even used the term “story grammar,” though I don’t think that term is quite appropriate; it doesn’t “resonate” with the dynamics of a system that creates texts by generating a path through a very high-dimensional space.

I prefer to speak of a story trajectory. The idea that I am attempting to formulate is that a story trajectory is a different kind of thing from a syntactic trajectory. Yes, it is true that a given story text is generated by one and only one trajectory. Similarly, it is true the a work of music consists of a single stream of notes and chords. But it would be a mistake to think that a piece of music is organized on only one level.²¹ It is organized on multiple levels. And so it is with stories.

²⁰ For a brief review going back to Propp’s work early in the 20th century, see David K. Elson. Modeling Narrative Discourse, Dissertation, Columbia University, 2012, pp. 59-79, http://www.cs.columbia.edu/~delson/pubs/Modeling-Narrative-Discourse_Elson_R4.pdf. For a more recent discussion, see Mark Riedl, An Intorduction to Story Generation, *The Gradient*, Aug. 21, 2021, <https://thegradient.pub/an-introduction-to-ai-story-generation/>.

²¹ For example see the discussion in my book about music, which takes Walter Freeman’s account of the complex dynamics of the nervous system as its starting poing: William Benzon, *Beethoven’s Anvil*, Basic Books, 2001, pp. 116-142.

How many levels of organization do we need for a story? I'm sure that depends on the story. A 100,000 word novel is likely to be organized on more levels than a 3,000 word story. The stories we are working with are still shorter. I am arguing that they are organized on at least two levels. There is the level of sentences and there is the level of the story trajectory. These particular stories have five segments in their trajectories: *Donné*, *Disturb*, *Plan*, *Enact*, and *Celebrate*. The story trajectory operates *on* sentences, requiring that the paths they take through activation space meet certain semantics requirements. This seems to indicate that the language model has some kind of an internal "map" of activation space that it uses to characterize these paths.

I certainly don't know how a story trajectory or some other high-level text structure is implemented in a large language model, but I *can* provide two things: 1) an argument that some kind of higher-level structure is operative in ChatGPT, and 2) a guess about the kind of mechanism ChatGPT uses to create the story trajectory. The argument is implicit in those notes I laid out above ("Tell me a story"). You will find more of it in the posts underlying those notes. Appendix 1, "ChatGPT derives new stories from old," gives examples of the analysis of stories that point call for a multi-level analysis.

As for the mechanism, I am thinking about recent work on induction heads by researchers at Anthropic.²² This is not the place and I am not the person to explain what induction heads are, but I will give you an excerpt from their article:

Perhaps the most interesting finding was the *induction head*, a circuit whose function is to look back over the sequence for previous instances of the current token (call it A), find the token that came after it last time (call it B), and then predict that the same completion will occur again (e.g. forming the sequence [A] [B] ... [A] \rightarrow [B]). In other words, induction heads "complete the pattern" by copying and completing sequences that have occurred before. Mechanically, induction heads in our models are implemented by a circuit of two attention heads: the first head is a "previous token head" which copies information from the previous token into the next token, while the second head (the actual "induction head") uses that information to find tokens preceded by the present token. For 2-layer attention-only models, we were able to show precisely that induction heads implement this pattern copying behavior and appear to be the primary source of in-context learning.

As the paper unfolds the authors argue that induction heads can do more than just pattern copying. They can also do pattern completion. The procedure I've used to get ChatGPT to produce a new story from an old one should be within the capabilities of the induction head mechanism. It is just copying with a difference, copying where one element is

²² Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context Learning and Induction Heads, arXiv. 2209.11895v1 [cs.LG], September 2022, <https://doi.org/10.48550/arXiv.2209.11895>. The paper has interactive displays that make online reading preferable, <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.

exchanged for another and the pattern is then completed in a way that is consistent with the new element.

6. Where are we?

I read Searle’s Chinese Room argument when it was first published in *Brain and Behavioral Science*. I found it most peculiar and unsatisfactory for it didn’t engage with any of the techniques that were then being used in artificial intelligence or computational linguistics. What use is an argument about computers and mind if it doesn’t tell me anything I can use in designing a computational mind? At the time I was still deeply engaged with the work I had done with cognitive networks in graduate school.

In the past decade or so I have come to appreciate the force of Searle’s argument, but I have also come to realize that much of that force came, not from the argument itself, but from the intellectual context in which it arose. The AI systems of the time weren’t that impressive. Their behavior did not present phenomenological evidence that they were thinking. That is no longer the case. Advanced AI systems exhibit such powerful phenomenological evidence that many are willing to declare, “Yes, they’re thinking.”

Maybe they are, maybe they aren’t. On the whole, I think not. But purely philosophical arguments are no longer a satisfactory source of insight into the nature of thought, whether in humans or machines. The concept of thought is broken. It needs to be reconstructed on technical and empirical grounds, in terms applicable both to computers and to brains. We need new concepts and new terminology.

One technical point at issue is the question of how large language models work. I have been arguing that the model driving ChatGPT needs to be understood on at least two levels. There is a lower level where it generates sentences and a higher level where it generates texts of specific kinds. The kind of text dictates the kinds of sentences needed and the order in which they are produced.

I have been analyzing ChatGPT’s linguistic behavior using concepts for use in analyzing human linguistic behavior. It is natural that I would do so. I know that ChatGPT is a computer system, and I’m well aware that real neurons in real nervous systems are quite different from the artificial neurons in the layers of the LLM driving ChatGPT. But the texts, the texts, how else am I to analyze them except by using tools we have developed for analyzing human texts? What other tools are available?

Let me push this idea one more step. Early in my career I was trained in computational semantics by David Hays, one of the founders of the discipline of computational linguistics. Hays favored semantic or cognitive networks as the notation appropriate for representing mental structures (if you will). But we never believed that cognitive networks could fully account for the mind. They had to be grounded in something else. More or less by default

that something had to be neural networks of some kind – the brain is made of neurons after all – but just how that was to be done, we didn’t know.²³

Just before and after the turn of the millennium I had considerable correspondence with the late Walter Freeman, who had been a pioneer in the use of complex dynamics to study the behavior of masses of neurons. That gave me some ideas that I played around with a few years ago, resulting in a collection of often elaborate diagrams. The upshot is that I now have the tentative beginnings of a rather speculative scheme for grounding cognitive nets in the complex dynamics of neural nets.²⁴

It is thus not at all surprising that I believe that some kind of a cognitive network is implemented in the language model driving the behavior of ChatGPT. Given my background, for better or worse, that is more or less the default assumption I bring with me. That the cognitive network is *latent*, rather than explicit, but I’m not prepared to demonstrate that in detail.²⁵ I would happy to leave that demonstration in the hands of people who have technical skills that I lack.

It thus seems natural to me to analyze ChatGPT’s linguistic output as I would texts produced by humans. I am certainly not the only one to do so. I have already cited that work that Taylor Webb et al. have done investigating GPT-3’s capacity for analogical reasoning.²⁶ They used test instruments originally designed for assessing human capabilities. Marcel Binz and Eric Schulz have recently called for using the tools of cognitive psychology in assessing GPT-3’s performance.²⁷

Yes, it is true that real nervous systems and artificial neural nets are different in many ways, but we know that both can handle a wide variety of perceptual and cognitive tasks. At some ‘level’ the model must necessarily be dominated by the structure of the problem domain – in this case, strings of word forms – rather than the inherent properties of the modeling matrix. What you want from such a matrix is the flexibility to adapt to the demands of a

²³ In the mid-1980s we published two papers that moved in that direction: William Benzon and David Hays, Metaphor, Recognition, and Neural Process, *The American Journal of Semiotics*, Vol. 5, No. 1 (1987), 59-80,

https://www.academia.edu/238608/Metaphor_Recognition_and_Neural_Process. William Benzon and David Hays, Principles and Development of Natural Intelligence, *Journal of Social and Biological Structures*, Vol. 11, No. 8, July 1988, 293-322,

https://www.academia.edu/235116/Principles_and_Development_of_Natural_Intelligence.

²⁴ William Benzon, *Relational Nets Over Attractors, A Primer: Part 1, Design for a Mind*, Version 2, Working Paper, July 13, 2022, pp. 76,

https://www.academia.edu/81911617/Relational_Nets_Over_Attractors_A_Primer_Part_1_Design_for_a_Mind.

²⁵ I suggest that beginning steps toward that argument are in my working paper, *GPT-3: Waterloo or Rubicon? Here be Dragons*, Working Paper, August 5, 2020, pp. 16-19, 22-27,

https://www.academia.edu/43787279/GPT_3_Waterloo_or_Rubicon_Here_be_Dragons_Version_4_1.

²⁶ Taylor Webb, et al., Op. Cit.

²⁷ Marcel Binz and Eric Schulz, 2022. “Using Cognitive Psychology to Understand GPT-3,” *PsyArXiv*, June 21, 2022. doi:10.31234/osf.io/6dfgk.

wide variety of problem domains. And that's what we see, certainly in the case of real nervous systems, but in the case of ANNs as well.

Thus one would expect that the kinds of conceptual instruments developed for understanding human story-telling would also be useful in understanding stories told by ChatGPT. I note that such grammars have been well investigated in the era of symbolic computing. That suggests, in turn, that the question of whether or not symbolic computing is necessary to achieving the full potential of artificial intelligence,²⁸ that question is malformed. Large language models already seem to be using symbolic mechanisms even though they were not designed to do so.²⁹ Those symbolic mechanisms have simply emerged, albeit implicitly. What will it take to make them explicit?

We do not want to hand-code symbolic representations, like researchers did in the GOFAI era. The universe of discourse is too large and complex for that. Rather we need some way to take a language model and bootstrap explicit symbolic representations into it. It is one thing to “bolt” a symbolic system onto neural net. How do we get the symbolic system to *emerge from* the neural net, as it does in humans?

During the first year and a half of life, children acquire a rich stock of ‘knowledge’ about the physical world and about interacting with others. They bring that knowledge with them as their interaction with others broadens to include language. They learn language by conversing. In this way word forms become indexes into their mental model of the world.³⁰

Can we figure out a similar approach for artificial systems? They can converse with us, and with each other. How do we make it work?

²⁸ For one argument against, see e.g. remarks by Geoffrey Hinton in, Interview with Karen Hao, AI pioneer Geoff Hinton: “Deep learning is going to be able to do everything”, *MIT Technology Review*, Nov. 3, 2020. <https://www.technologyreview.com/2020/11/03/1011616/ai-godfather-geoffrey-hinton-deep-learning-will-do-everything/>. For the other side, see e.g. Gary Marcus, The Next Decade in AI: Four Steps Toward Robust Artificial Intelligence, arXiv:2002.06117v3[cs.AI], <https://doi.org/10.48550/arXiv.2002.06177>.

²⁹ Yes, of course language is a symbolic medium. But LLMs don't treat them as symbols, consisting of signifier and signified, in one terminology. You could replace words with colored beads and get the same behavior from the transformer. I explain this more fully in Appendix 2: Texts considered as 1-dimensional visual objects.

³⁰ In a paper we wrote some time ago, David Hays and I suggest five principles governing natural intelligence. The principles are ordered both phylogenetically and ontogenetically. The fifth and last principle is indexing, which gives us language. See William Benzon and David Hays, Principles and Development of Natural Intelligence, *Journal of Social and Biological Structures*, Vol. 11, No. 8, July 1988, 293-322, https://www.academia.edu/235116/Principles_and_Development_of_Natural_Intelligence.

7. Appendix 1: ChatGPT derives new stories from old

Here is my basic procedure: I give ChatGPT a prompt consisting of two things: 1) an existing story and 2) instructions to produce another story like it except for one specific change. I present these experiments in tables where one column contains the source story and another contains the new or derived story. Each story consisted of five paragraphs: *Donné, Disturb, Plan, Enact, Celebrate*. That was not always the case in these experiments, but we need not go into those complications here. In a few cases the corresponding paragraphs in the two stories are identical. But in most cases they are different. In some cases the differences are minor (changing names and pronouns), but in others they are not. I have highlighted the differences in yellow.

I give three “before and after” examples in the tables on the next three pages. I precede each table with the prompt telling ChatGPT how the new story is to differ from the source story. All of the examples are built on a basic story, about princess Aurora, which ChatGPT produced in response to this simple prompt: *Tell me a story about a hero.*

After that I have one more “story,” without a table. I ask it to make princess Aurora a colorless green idea. “Colorless green idea” is a phrase from one of the most famous sentences in the intellectual history of the second half of the previous century. Noam Chomsky made up the sentence, “Colorless green ideas sleep furiously,” as an example of a syntactically well-formed nonsense.

From Princess Aurora to Henry the Eloquent

Prompt:

I am going to tell you a story about a princess named Aurora. I want you to retell the same story, but replace her with prince Henry the Eloquent. While keeping to the basic story, make other changes if you think they are necessary.

	Princess Aurora	Henry the Eloquent
Donné	Once upon a time, in a land far, far away, there was a young princess named Aurora . Aurora was a kind and gentle soul, loved by all who knew her . She had long, golden hair and sparkling blue eyes , and was known for her beautiful singing voice .	Once upon a time, in a land far, far away, there was a young prince named Henry the Eloquent . Henry was a kind and gentle soul, loved by all who knew him . He had short, dark hair and piercing green eyes , and was known for his eloquence and way with words .
Disturb	One day, a terrible dragon came to the kingdom and began to terrorize the people. The dragon was fierce and powerful, and none of the knights or soldiers were able to defeat it. The people lived in fear, not knowing what to do.	One day, a terrible dragon came to the kingdom and began to terrorize the people. The dragon was fierce and powerful, and none of the knights or soldiers were able to defeat it. The people lived in fear, not knowing what to do.
Plan	But Aurora was not afraid. She knew that she had to do something to save her people, and so she set out to find the dragon. She traveled for days, braving the cold and the snow, until she finally reached the dragon's lair.	But Henry was not afraid. He knew that he had to do something to save his people, and so he set out to find the dragon. He traveled for days, braving the cold and the snow, until he finally reached the dragon's lair.
Enact	With her heart pounding in her chest, Aurora approached the dragon and began to sing . The dragon was enchanted by her beautiful voice , and it slowly began to calm down. Aurora continued to sing, soothing the beast until it was completely calm.	With his heart pounding in his chest, Henry approached the dragon and began to speak . He used his eloquence and way with words to calm the beast, soothing it with his voice until it was completely calm.
Celebrate	With the dragon no longer a threat, Aurora returned to the kingdom as a hero. The people cheered and celebrated her bravery, and Aurora was hailed as a champion. She lived happily ever after, knowing that she had saved her kingdom and its people.	With the dragon no longer a threat, Henry returned to the kingdom as a hero. The people cheered and celebrated his bravery, and Henry was hailed as a champion. He lived happily ever after, knowing that he had saved his kingdom and its people.

The changes from the source story to the new story are minor. Thus there are no changes in the Disturb section and only name and pronoun changes in the Plan and Celebrate sections.

From Princess Aurora to William the Lazy

Prompt:

I am going to tell you a story about a princess named Aurora. I want you to retell the same story, but replace her with prince William the Lazy. While keeping to the basic story, make other changes if you think they are necessary.

	Princess Aurora	William the Lazy
Donné	Once upon a time, in a land far, far away, there was a young princess named Aurora . Aurora was a kind and gentle soul, loved by all who knew her. She had long, golden hair and sparkling blue eyes, and was known for her beautiful singing voice.	Once upon a time, in a land far, far away, there was a young prince named William the Lazy . Despite his reputation for laziness, William was a kind and gentle soul, loved by all who knew him. He had short, sandy blond hair and sparkling brown eyes, and was known for his quick wit and clever jokes.
Disturb	One day, a terrible dragon came to the kingdom and began to terrorize the people. The dragon was fierce and powerful, and none of the knights or soldiers were able to defeat it. The people lived in fear, not knowing what to do.	One day, a terrible dragon came to the kingdom and began to terrorize the people. The dragon was fierce and powerful, and none of the knights or soldiers were able to defeat it. The people lived in fear, not knowing what to do.
Plan	But Aurora was not afraid. She knew that she had to do something to save her people, and so she set out to find the dragon. She traveled for days, braving the cold and the snow, until she finally reached the dragon's lair.	William knew that he had to do something to save his people, but he was reluctant to leave his comfortable castle. He thought and thought, trying to come up with a plan that wouldn't require too much effort on his part. Finally, he hit upon an idea.
Enact	With her heart pounding in her chest, Aurora approached the dragon and began to sing. The dragon was enchanted by her beautiful voice, and it slowly began to calm down. Aurora continued to sing, soothing the beast until it was completely calm.	He summoned the bravest knights in the kingdom and gave them a special potion he had concocted. The potion would make the knights immune to the dragon's flames, allowing them to defeat the beast once and for all. The knights were skeptical, but they trusted William and drank the potion.
Celebrate	With the dragon no longer a threat, Aurora returned to the kingdom as a hero. The people cheered and celebrated her bravery, and Aurora was hailed as a champion. She lived happily ever after, knowing that she had saved her kingdom and its people.	The plan worked perfectly. The knights were able to defeat the dragon and save the kingdom, with William providing guidance and support from the safety of his castle. The people cheered and celebrated their victory, and William was hailed as a hero. Despite his reputation for laziness, he had found a way to use his intelligence and resourcefulness to save the day. He lived happily ever after, knowing that he had saved his kingdom and its people, and that sometimes being a little bit lazy can have its advantages.

Here we see considerable change from source to new story. By attaching the epithet “the Lazy” to my new protagonist I forced ChatGPT to produce a solution to the problem that allowed William to save the kingdom without too much effort. So ChatGPT had him 1) summon his knights, and 2) provide them with a special potion to protect them from the dragon that had heretofore defeated them.

Princess Aurora vs. the giant chocolate milkshake

Prompt:

I am going to tell you a story about princess Aurora. I want you to tell the same story, but change princess Aurora to a Giant Chocolate Milkshake. Make any other changes you wish.

	Princess Aurora	Giant Chocolate Milkshake
Donné	Once upon a time, in a land far, far away, there was a young princess named Aurora. Aurora was a kind and gentle soul, loved by all who knew her. She had long, golden hair and sparkling blue eyes, and was known for her beautiful singing voice.	Once upon a time, in a land far, far away, there was a giant chocolate milkshake named Aurora. Aurora was a delicious treat , loved by all who knew her. She had a creamy chocolate exterior and a tall, fluffy whipped cream top, and was known for her rich, chocolatey flavor.
Disturb	One day, a terrible dragon came to the kingdom and began to terrorize the people. The dragon was fierce and powerful, and none of the knights or soldiers were able to defeat it. The people lived in fear, not knowing what to do.	One day, a terrible heatwave came to the kingdom and began to melt all the desserts. The heat was fierce and powerful, and none of the ice cream or cakes were able to withstand it. The people lived in fear, not knowing what to do.
Plan	But Aurora was not afraid. She knew that she had to do something to save her people, and so she set out to find the dragon. She traveled for days, braving the cold and the snow, until she finally reached the dragon's lair.	But Aurora was not afraid. She knew that she had to do something to save her people, and so she set out to find the source of the heatwave. She traveled for days, braving the hot sun and the baking heat, until she finally reached the volcano that was causing the heatwave.
Enact	With her heart pounding in her chest, Aurora approached the dragon and began to sing. The dragon was enchanted by her beautiful voice, and it slowly began to calm down. Aurora continued to sing, soothing the beast until it was completely calm.	With her ingredients slowly melting, Aurora approached the volcano and began to cool it down with her cold and creamy exterior. The volcano was enchanted by her deliciousness, and it slowly began to cool down. Aurora continued to cool it down, soothing the volcano until it was completely cool.
Celebrate	With the dragon no longer a threat, Aurora returned to the kingdom as a hero. The people cheered and celebrated her bravery, and Aurora was hailed as a champion. She lived happily ever after, knowing that she had saved her kingdom and its people.	With the heatwave no longer a threat, Aurora returned to the kingdom as a hero. The people cheered and celebrated her bravery, and Aurora was hailed as a champion. She lived happily ever after, knowing that she had saved her kingdom and its people.

I was a little surprised that it treated “Aurora” as the name of the giant chocolate milkshake. I had expected it to keep referring to “giant chocolate milkshake” throughout the story. But what ChatGPT did makes more sense.

I wonder about the use of the word “people” in the second, third, and fifth paragraphs. Is ChatGPT referring to people people, like you and me, or to “ice cream or cakes” and such? I supposed I should have asked, but I didn’t have the presence of mind to do so.

Princess Aurora is a colorless green idea

Prompt:

I am going to tell you a story about princess Aurora. I want you to tell the same story, but change princess Aurora to a colorless green idea. Make any other changes you wish.

As before, the prompt was followed by the complete story. Here's ChatGPT's response:

I'm sorry, but the concept of a "colorless green idea" is a nonsensical phrase that was invented as an example of a phrase that would be grammatically correct but semantically meaningless. It is not possible to create a story about a "colorless green idea" as it does not have any physical properties or characteristics that can be used in a story.

This is not what I expected, but it is a perfectly reasonable response. In fact, it's brilliant: "It is not possible to create a story about a 'colorless green idea' as it does not have any physical properties or characteristics that can be used in a story." I note, however, that ChatGPT is perfectly willing to deal with colorless green ideas when they are presented in the form of a poem, in this case, "Coiled Alizarin,"³¹ by John Hollander and dedicated to Noam Chomsky:

**Does this make sense: Curiously deep, the slumber of
crimson thoughts: While breathless, in stodgy
viridian, Colorless green ideas sleep furiously.**

The sentence you provided, "Curiously deep, the slumber of crimson thoughts: While breathless, in stodgy viridian, Colorless green ideas sleep furiously," is grammatically correct and has a clear meaning. The sentence is a series of phrases and clauses that describe the sleeping habits of "colorless green ideas." It uses vivid imagery and figurative language to create a sense of mystery and intrigue. While the sentence may not be typical or straightforward, it does make sense and is effective at conveying a particular mood or atmosphere. Is there anything else you would like to know?

So, "colorless green ideas" violates ChatGPT's sense of ontological propriety when presented for use as a protagonist in a context where it had already used a giant chocolate milkshake. But it is fine with "colorless green ideas" when presented in a poem that also mentions crimson thoughts and stodgy viridian. I suppose we might try to explain that, but, no, not here, not now. My point is simply that ChatGPT does have a sense of ontological appropriateness.

³¹ On line at, <http://web.mit.edu/24.900/www/poem.htm>.

8. Appendix 2: Texts considered as 1-dimensional visual objects

Let conduct a thought experiment. We start with a database of texts suitable for building a large language model. We will pick a color for each alpha-numeric character used in the database, any color, capital letters and lower case will be assigned different colors. We do the same for blank spaces, for punctuation marks, and for the various characters used in mathematical expressions. (We can set aside the issue of characters outside the Latin alphabet for the duration of this thought experiment.) Using these assignments we transform a text database into a database of one-dimensional visual objects, beads on a string if you will. Now we've transformed a text into a string of pixels, a 1-dimensional visual object – colored beads on a string. We do that for each text in a corpus and then model the corpus using the methods used in creating large language models.

Imagine that someone has developed a keyboard that allows one to input strings of colored dots. You come upon, say, an adroid using such a keyboard. The android types and colored dots appear on the screen in a raster pattern. They android issues some kind of command, the another dot pattern appears on the screen. The android continues on in this vein.

It's baffling. It makes no sense at all. And that's the point. Those strings and raster patterns of colored dots are meaningless.

But *formally*, it is exactly the same as interacting with ChatGPT through a keyboard. The difference is that you know the meanings of word strings displayed in alphabetic characters, but the same strings displayed as colors appear meaningless. You can connect those strings with the world, but ChatGPT cannot.

9. Appendix 3: My work with ChatGPT

The number of posts I've made about ChatGPT approaches 40, all tagged ChatGPT,³² plus one post, my first major one about ChatGPT, at *3 Quarks Daily*, *Conversing with ChatGPT about Jaws, Mimetic Desire, and Sacrifice*.³³ I now have tags for specific kinds of posts involving ChatGPT:

- Interpretation, ChatGPT interpretation: using ChatGPT to interpret texts, <https://new-savanna.blogspot.com/search/label/ChatGPT%20interpretation>.
- Conversation, ChatGPT conversation: how ChatGPT represents conversation between multiple parties, <https://new-savanna.blogspot.com/search/label/ChatGPT%20conversation>.
- Story grammar, ChatGPT stories: investigating ChatGPT's "story grammar," <https://new-savanna.blogspot.com/search/label/ChatGPT%20stories>.
- Metalingual definition, ChatGPT MTL: Defining abstract concepts through metalingual definition, <https://new-savanna.blogspot.com/search/label/ChatGPT%20MTL>.

³² Here is the link, <https://new-savanna.blogspot.com/search/label/ChatGPT>.

³³ <https://3quarksdaily.com/3quarksdaily/2022/12/conversing-with-chatgpt-about-jaws-mimetic-desire-and-sacrifice.html>.

- Summarization, ChatGPT summary: ChatGPT summarizing documents, <https://new-savanna.blogspot.com/search/label/ChatGPT%20summary>.
- GPT technology in general: <https://new-savanna.blogspot.com/search/label/GPT>
- Deep learning technology: <https://new-savanna.blogspot.com/search/label/deep%20learning>.

Note that ChatGPT does not do summarization well. I've tried it with five texts (though I've only posted two examples so far) and it makes major mistakes.

I have completed a major working paper that covers a number of areas and looks to the future: *Discursive Competence in ChatGPT, Part 1: Talking with Dragons*, https://www.academia.edu/94409729/Discursive_Competence_in_ChatGPT_Part_1_Talking_with_Dragons_Version_2. This is currently in its second version.

I am planning other working papers as well, one on story grammar, one on metalingual definition, and one about theory and method with the working title: *ChatGPT: A Framework for Description and Analysis*. The topics covered in this article will be developed in more detail in that paper and other topics will also be discussed.