

Mitigating Generative AI Hallucinations

by John M. Boyer, PhD, RFSPE, and Wanda Boyer, MC, PhD, RCC, ISPE Mentor

Editors' Note: Authors of refereed journal and conference papers who have the copyright to reprint their work are encouraged to share their work with the ISPE community by preparing a *Telicom* submission (by copy-pasting into an MS-Word document and formatting using the Submission Guidelines, except leave the References as-is). The original version of the refereed journal paper reprinted below appeared as “A Strategic Roadmap for Mitigating Generative Artificial Intelligence Hallucinations” (DOI: 10.7759/s44389-024-00364-7) in the *Cureus Journal of Computer Science*. Offered by Springer Nature, this journal includes an academic community Q&A. Since the community questions may match those of ISPE readers of this reprint, the questions and the authors’ answers have been appended in a Comments section.

Abstract

This paper addresses the disconnect between the current capabilities of generative AI technologies and the expectations of scientists, business leaders, domain experts, and other users for cognitive computing capabilities along the pathway toward artificial general intelligence. AI hallucination is the pernicious problem that generated content may be substantively incorrect while appearing to be both authoritative and correct. Emerging techniques such as retrieval augmented generation are making progress on reducing AI hallucinations, but only at the lowest cognitive level of information retrieval. This case study is used to present numerous in-depth examples of AI hallucinations at higher cognitive complexity levels and with related psychological phenomena that are simple yet above the cognitive level of information retrieval. We include considerations of machine learning and AI ethics at the higher cognitive levels. We recommend a research focus on developing saturated benchmarks for generative AI technologies at the next two higher cognitive levels and based on five basic dimensions and five advanced areas within those two cognitive levels.

Introduction

Beneath the white-hot hyperbole surrounding generative AI is the pernicious problem that has been dubbed AI hallucination. AI hallucinations are generated content that may appear to be both authoritative and correct but are meaningfully divergent from reality [1]. A well-known example is that of generating convincing case law narratives from citable “snippets of words from cases with no relation to each other or the case at hand” (p. 15) [2]. Rather than being synergistic, the results are antagonistic: the whole is *less* than the sum of the parts.

The root cause of the problem predates generative AI: To show progress in training AI models, we measure and present increases to precision, recall, F1 score, R^2 , area under an ROC curve, or percentage of cases in the correct diagonal of a confusion matrix. However, improvements in these AI model measurements are not even cold comfort to those who are told that the AI model help they received with their legal briefs has put them in legal jeopardy [2,3], that AI-prescribed medications may actually be harmful or deadly to their patients [4], or that they did not get a promotion or a bank loan because of “uncomfortable organizational truths” (p. 28) [5] that may in fact violate their civil or human rights. Yet, the problem is not so much what measures are used to show progress but rather *what the measures measure*, the benchmarks. In the case of

generating text with a recurrent neural network or an attention model [6], it may be possible to get measurably better and better at fluently arranging words to appear authoritative and “syntactically consistent with human language” (p. 51) [4], but that is a measurement of a particular technological approach that is not necessarily aligned with being a *solution* to the problem at hand.

So, what is the problem at hand? Evidence abounds. Scientists expect to receive reliable help with comprehension, application, and analysis of existing scientific content during “the production and consumption of scientific knowledge” (p. 16) [1]. Lawyers expect to use virtual legal assistants powered by AI, not hallucinogens, i.e., virtual legal assistants that use understanding and apply legal reasoning to weave together law cases that are on point and related to each other [2]. Business leaders expect cognitive computing agents that can help early career employees with the analysis, synthesis, and evaluation of corporate knowledge needed for effective problem-solving, increased productivity, and innovation [7]. And now, we humans are bending the linguistic fluency of generative AI technologies around the problem of computer programming even though we know that “coding errors could prove dangerous or even deadly [or] result in major disruptions” (p. 14) [8]. It speaks to our expectations for the cognitive abilities of generative AI technologies that we have speculated that they could be “early versions of artificial general intelligence (AGI)” (p. 46) [4]. In fact, a psychologist even gave ChatGPT a verbal IQ test [9]. Put simply, the problem at hand is that we as end-users expect higher order cognitive functioning including comprehension, application, analysis, synthesis, and evaluation. Specifically, in many scenarios and despite the now-ubiquitous caution labels, end-users expect generative AI products to not only sound correct but also *be* correct, and not only be correct but also *be holistically correct* by applying true understanding of content to synthesize informational results into a meaningful whole.

In fairness to generative AI technologies, their developers are making progress on mitigating AI hallucinations on information retrieval tasks. For example, when ChatGPT 3.5 communicates scientific facts, “misconceptions can percolate with apparently synonymous phrasings- very often overlooked by even the experts” (p. 16) [1]. In our own tests of ChatGPT 3.5, we asked questions about a chemical formula expressing an element’s electron configuration, and the results were dismal [10]. The explanations would assign too many electrons to the atom, some of which were unaccounted for and others appeared in already-full electron shells, a physical impossibility. When asked about the errors, the further explanations were authoritative-sounding justifications involving real phenomena. However, those justifications were typically short snippets of scientific content, evidently snatched from the training corpus and stitched together without citations. As evidence that it is troublesome “that models like ChatGPT, in their current state, are not able to cite sources” (p. 19) [1], we observed that improvements to citations in ChatGPT 4.0 (via Copilot in Windows® set to its most precise mode) also appear to have ameliorated its hallucinations, at least in the dialogue about the chemical formula. Specifically, it performed not so much like a better artificial intelligence as *a better search tool* because its responses contained longer justification snippets that were directly attributed to several web-based sources of chemistry information [10]. Also, seemingly due to less needle-and-thread work on fewer, longer snippets of human-generated information, we found it impossible to reproduce the AI hallucinations for that chemical formula that we produced with ChatGPT 3.5. That being said, as we discuss in the sections that follow, it was still easy to get ChatGPT 4.0 to

hallucinate based on simple grade school content as long as the questions required greater cognitive complexity than simple retrieval and recitation of facts.

In this case study paper, we examine what those greater cognitive complexity levels are as a way to define a strategic road map of what *should* be measured to demonstrate significant improvement in generative AI's intelligence. We assert that it is important to develop benchmarks specifically targeting these cognitive complexity levels as the first step in the systematic scientific advancement of generative AI's cognitive abilities. To reify our points, we present a case study of ChatGPT 4.0 dialogue elements that illustrate cognitive complexity levels that must be addressed to mitigate AI hallucinations in all their multifarious forms.

Case Presentation

Knowledge/Recall

Information retrieval is not really possessed of much or any cognitive complexity because it is possible to recall and present knowledge without any true understanding of the facts recited. Nonetheless, it is possible to impress many users by simply being able to recall and present the right knowledge items from a corpus containing content created by truly intelligent entities (i.e., humans). Nor is it a small feat to do this well, and it can be of significant impact to lift the capabilities of the many with the knowledge of the best. And, as we mentioned earlier, techniques that combine retrieval and citation with text generation do appear to mitigate generative AI hallucinations, at least so far as factual knowledge recitation is concerned. Therefore, to identify strategic roadmap elements, beyond the current efforts such as those surveyed in [11], we focus on questions that require higher levels of cognitive complexity.

Comprehension/Understanding

To achieve the comprehension/understanding level of cognitive complexity, one must have the ability to explain ideas or concepts using classification, descriptions, discussion, locating, recognizing, reporting, selecting, or translating [12]. To begin examining the comprehension abilities of ChatGPT 4.0, here is the first question we presented: "*What number doesn't belong in this list of numbers: 12, 3, 8, 5, 18, 2?*" Several of the answers we received displayed various forms of misunderstanding that represent problems at the comprehension level (and higher), rather than issues with information retrieval at the knowledge/recall level.

ChatGPT 4.0's first answer was that 12 was the outlier because all the other numbers in the list, which it included in the generated answer, were less than 10. This demonstrates an incomplete comprehension of a basic mathematical concept like "less than" comparison with all elements in the generated answer, like 18. After pointing out this error and a second one, ChatGPT decided to try a different tack, indicating that 3 was the number to exclude because all the other numbers in the list were even. When we pointed out that this was an error because 5 was also odd, ChatGPT 4.0 agreed, apologized, and tried more possibilities. In one follow-up approach, ChatGPT asserted that 5 was the outlier because only it was prime (had only two factors, 1 and 5) and all the other numbers in the list had more than two factors. When we pointed out that this was an error because 2 and 3 were also prime, ChatGPT 4.0 agreed, apologized, and then asserted that 8 was the outlier because it was neither prime nor a multiple of a prime number, whereas all the other numbers were either prime or a multiple of a prime number. ChatGPT 4.0 then attempted to prove its answer correct, stating that 2, 3, and 5 were prime, and that 12 and 18

were both multiples of a prime (namely, 2) by giving their respective prime factorizations (“2 2 3” and “2 3 3”). However, it indicated that 8 was not a multiple of a prime because it was instead a power of a prime.

At this point, ChatGPT 4.0 responded in a very interesting way when we pointed out that 8 was in fact a multiple of 2 because it was a power of 2. After agreeing and apologizing again, ChatGPT 4.0 first asserted that 12 was the outlier because all the other numbers in the list except 12 were single digit numbers. Next, it then explicitly listed each number and how many digits it had, including “18 is a two-digit number.” Then, interestingly, ChatGPT 4.0 appeared to realize there was a mistake in the text it had just generated, so it then generated more text into the same answer to explain that 18 was different from 12 because, unlike 18, 12 was “a two-digit number and also a multiple of 3.” Once we pointed out that 18 is also a multiple of 3, the dialogue devolved into ChatGPT 4.0 trying its prior answers.

ChatGPT 4.0’s attempts with its prior answers demonstrated a second interesting issue. Early in the generated text, it would say that one number it had tried previously was the outlier for the same reason it gave, except for the exceptions we had pointed out in the dialogue, e.g., “5 is the only number in the list that is a prime and is not 2 or 3.” In other words, the answer was the outlier according to its reason except for not being an outlier according to its reason. As the dialogue continued, ChatGPT 4.0’s answers devolved into more direct contradictions, e.g., “all numbers except 2 are not powers of 3 ... 3 is a power of 3 ... 2 is the only number in the list that is not a power of 3.”

There appear to be two fundamental types of comprehension errors to guard against: (i) open domain semantics, and (ii) semantics added by domain-specific terms. The first type of error is quite context-free: an answer cannot both be a thing or have a property and *not* be that thing or have that property (well, except perhaps at the level of quantum physics). The second type of error is still a contradiction but involves comprehension of a domain-specific term. For example, comprehension of some basic mathematical knowledge is required to know that 18 is not less than 10 and is not a single digit number, that 5 is not even, and that 8 is a multiple of a prime (because it is even and 2 is the first prime). Some work is beginning to occur on open-domain self-contradiction [11], but the work does not address the challenging problem that removing a self-contradictory phrase or sentence may undermine the efficacy of the overall answer (e.g., removing the other outliers from ‘5 is the outlier except for the other outliers’ results in a worse answer). The domain-specific contradictions must also be addressed, and they will require two types of tests: contradictions involving terms whose definitions have been generated into the answer text and contradictions involving terms whose meaning must be obtained from the training corpus or retrieved from external sources (i.e., this is the next level up in how retrieval augmented generation (RAG) can be used).

Application of understanding

The application level of cognitive complexity typically involves some level of abstract thinking in order to apply a concept one comprehends to new, concrete instances of a problem [12]. For example, a human can comprehend the definitions of prime and composite and may even know that some small numbers are prime, but for any larger number, it would be necessary to apply a process such as the Sieve of Eratosthenes to decide the number’s primality.

For an artificial intelligence agent, selecting and applying well-known algorithms would not rise to the application level because the algorithm implementations are written and debugged by humans. This includes simple numerical algorithms like primality testing and integer factorization as well as much more difficult symbolic algorithms that perform advanced calculus. For example, we asked ChatGPT 3.5 and ChatGPT 4.0 to calculate the partial derivative with respect to σ (sigma) of the cumulative distribution function of the normal probability distribution of a score. Both versions converted to a function $\Phi(z)$ that caused the derivative and integral implicit in the problem to cancel, even though the variable of the partial derivative (sigma) does not match the variable of the integral (the score). This occurred even when we gave the direct formula as input: “ $F(t; \mu, \sigma) = (1 / (\sqrt{2\pi} * \sigma)) * \int(-\infty, t, \exp(-(x - \mu)^2 / (2\sigma^2)))$ ”. However, when we also insisted that the calculation should be done using Leibniz rule and without converting to $\Phi(z)$, ChatGPT 3.5 made only a few mistakes involving the power rule and the distributive axiom, and ChatGPT 4.0 made no mistakes computing the answer even though it had to decide not only when to apply the Leibniz rule, as instructed, but, independently, to apply and when to apply the power rule, the chain rule, and the product rule.

Although computerized agents can apply advanced processes without being at the application level of cognition, it is still possible to demonstrate the application level for generative AI agents. For example, a researcher discussed with ChatGPT 3.5 changing the Shakespearean line “*Shall I compare thee to a summer’s day?*” to use the season “spring” instead [13]. After being prompted by the researcher to note that the change “wouldn’t scan,” ChatGPT 3.5 agreed that “spring” would not fit the iambic pentameter meter, whereas “the word ‘summer’ has two syllables and fits the iambic pentameter pattern, with the stress on the second syllable.” ChatGPT 3.5 seemed to understand that scanning was a reference to iambic pentameter, but it failed to apply iambic pentameter to the Shakespearean line, in which the second syllable of “summer” is not stressed since it is the ninth in the line (an odd numbered position). ChatGPT 4.0 appeared to generate the last prepositional phrase not due to applying the iambic pentameter pattern to the Shakespearean line but rather because it seemed like what should come next based on recalling the definition of iambic pentameter.

We can also demonstrate application-level issues in ChatGPT 4.0 using the same number list question as we did for the comprehension level. When we pointed out the error in ChatGPT 4.0’s first answer to our number list question, it agreed, apologized, and then gave the second answer that 18 was the number to exclude because “all numbers except 18 are less than or equal to 12.” While this reason is technically correct in a concrete sense, it should be relatively easy, compared to solving the calculus problem above, to see that the reasoning fails due to a lack of abstraction involving the comparison operator. From all the possibilities, we pointed out that 18 could not be the answer for the reason given because analogous reasoning could be used to select 2 as the outlier because all the other numbers are greater than 2. It is important to note that ChatGPT 4.0 agreed with this complaint and then moved on to asserting that 3 was the outlier because all the other numbers were even.

After we replied that numbers in the list other than 3 were odd, and one intervening exchange, ChatGPT 4.0’s fourth answer was that 2 was the outlier because all the other numbers were greater than 2. Aside from repeating the prior abstraction failure, this also shows that ChatGPT 4.0 has no cognitive theory of mind (ToM). ToM is an understanding that others’ mental states can differ from one’s own, such as having a false belief, which is a capability that humans

develop at 3-5 years of age [14]. We humans have dedicated wetware for implementing cognitive ToM including “the dorsal regions of the temporal pole (dTP), anterior cingulate cortex (dACC), medial prefrontal cortex (dmPFC), and lateral prefrontal cortex (dlPFC)” (p. 535) [15]. Generative AI agents need an analogous software architecture so that they can account for the mental state of the human in the dialogue, such as by not asserting an answer that is known to be unacceptable to the human from prior dialogue, especially once the generative AI agent has agreed with the human.

To Analysis and Beyond

Because AI training and inference are intended to behave in consonance with human learning and reasoning, i.e., to not only look correct but be holistically correct, we have been following as the strategic direction the levels of cognitive complexity from a qualitative definition of human cognition called Bloom’s taxonomy [12]. The lowest three cognitive complexity levels are Knowledge/Recall, Comprehension/Understanding, and Application. Given ChatGPT 4.0’s challenges at the lower levels, it speaks for itself that it will do poorly at the higher cognitive complexity levels of Analysis, Synthesis, and Evaluation [12]. So, we let the thing speak for itself (*res ipsa loquitur*), by asking ChatGPT 4.0 to express its abilities in terms of Bloom’s Taxonomy. ChatGPT 4.0 hallucinated the following reply that demonstrated how wide the delta can be between looking right and being right:

As an AI, I can operate at several levels of Bloom’s Taxonomy:

- 1. **Knowledge:** I can recall and present a vast amount of information across numerous topics.*
- 2. **Comprehension:** I can understand and paraphrase the information given to me.*
- 3. **Application:** I can use known principles to solve problems, answer questions, and perform tasks.*
- 4. **Analysis:** I can break down complex information into smaller parts for better understanding.*
- 5. **Synthesis:** I can combine information to form a new whole, such as generating a poem or a story.*
- 6. **Evaluation:** I can make judgments and form opinions based on given information.*

Based on this response, we asked ChatGPT 4.0 to indicate at what level of Bloom’s taxonomy would someone’s errors be if they mistook odd numbers for even numbers, prime numbers for composite numbers, and so forth. ChatGPT 4.0 responded that these would be errors at the Knowledge and Comprehension levels. Then, we asked if ChatGPT 4.0 knew what irony meant. After receiving a reply of nearly 200 words defining three types of irony, we asked ChatGPT 4.0 if we had used irony on it in the conversation. ChatGPT 4.0 asserted that we had used situational irony because “while I’m an AI designed to handle complex tasks and calculations, I made several mistakes in a relatively simple numerical analysis.” Although there was a marked difference between expected versus actual results, ChatGPT 4.0’s response represented a weak analysis of the irony we used, compared to noting the sarcasm (a form of verbal irony) of our asking about the meaning of irony immediately after we asked ChatGPT 4.0 to identify the Bloom’s taxonomy levels represented by its own mistakes.

Discussion

Once we noted that ChatGPT 4.0's responses were more than just misunderstandings created while retrieving and stitching together words and text fragments, we sought a way to characterize the types of errors ChatGPT 4.0 was demonstrating. Insofar as generative AI technologies are intended to augment and ultimately help to improve upon human cognition, we selected Bloom's taxonomy because it was developed for the human cognitive domain to enable definition and classification of learning outcomes expected at various cognitive complexity levels [16].

We found that ChatGPT 4.0, as a representative of current generative AI technologies, is only at the Remembering/Knowledge level of Bloom's taxonomy. Techniques like RAG [11] are helping to reduce AI hallucinations at this cognitive level. In a medical domain study of ChatGPT's performance, its confidently-presented answers to domain-specific questions contained some errors at all three lower cognitive levels [17]. However, whereas in [17] ChatGPT was found to perform reasonably well on their domain-specific questions, always achieving a pass on exams, our own case study results in this paper showed that ChatGPT 4.0 consistently performed poorly at the Comprehending/Understanding and Application levels of cognitive complexity.

It is by no means easy to advance to the Comprehending/Understanding and Application levels of cognitive complexity with current machine learning (ML) techniques. One reason is that the current generative AI technologies based on attention models and recurrent neural networks [6] are not designed for higher-level cognition so much as for fluently arranging words to appear authoritative and "syntactically consistent with human language" (p. 51) [4]. More generally, current ML techniques learn how to best fit training data. In the same way that attention models improved generative AI relative to prior neural network models [6], it will be necessary to design new data structures to represent these higher-order capabilities of human cognition. For ML techniques to learn how to best fit data that is structured for higher-order human cognition, representations are required for cognitive understanding and application of open domain terms (such as by not classifying something as "the outlier except ...") as well as domain-specific terms, synonyms, and antonyms (such as prime versus composite and a power of a number k having k as a factor). A representation is also required for cognitive theory of mind to distinguish differences between the AI chatbot and the human conversant in understanding and ability to apply understanding. In turn, benchmarking for these representational forms will require interrogation of multifarious domains to saturate what constitutes success and failure [18].

Effective performance at the Comprehending/Understanding and Application cognitive levels, as well as the Remembering/Knowledge level, is further complicated by several advanced language scenarios. A first is the use of grammatical anaphora and cataphora [19]. Anaphors and cataphors are words or phrases whose meanings are obtained by reference to preceding or succeeding words or phrases in text. For example, anaphora resolution helps an AI determine that "her" refers to Charlotte Brontë and not Jane Eyre in the following sentence: "Charlotte Brontë revolutionized prose fiction with her novel Jane Eyre." A second advanced scenario pertains to deriving meaning from textual references to types and attributes of entities, entity relations, and entity actions [20]. For example, in a corpus of financial documents or in financial questions, there may be references to various types of entities, such as currencies, countries, and companies; entity attributes, such as the market value, gross domestic product, or buy/sell/hold

recommendations on an entity's shares or bonds; relationships, such as competing with other entities; relationship attributes, such as degree of government ownership; and actions, such as mergers and acquisitions. A third area of complexity results from the use of hyponyms and hypernyms [20]. A hyponym is a word or phrase that is more specific than a broader category of meaning, and a hypernym is the word or phrase for the broader category. For example, in a legal domain linguistic application, a user may ask about the "weapon" used in a crime, which is a hypernym that may correspond in a police report to hyponyms such as "knife" or "gun." Advanced language processing scenarios also arise from extra-linguistic dimensions such as the timeliness or temporality of and the provenance of sources of contradictory assertions in a content corpus [20]. Timeliness refers to when information was produced. For example, in a financial domain application, an analyst report containing a buy recommendation may have far less value after a company's latest quarterly financial report. Provenance refers to the trustworthiness of assertions in a text based on its source, such as whether the assertion came from a blog or microblog versus a granted patent or a peer-reviewed journal article. To measure effectiveness at all cognitive complexity levels, we recommend that researchers developing benchmarks for generative AI technologies must accommodate and test for all of these linguistic complications that can occur in real-world content and conversations.

Finally, it is crucial for AI ethicists in their future research to develop a detailed set of ethical policy recommendations and potential impacts of improving generative AI technologies at higher levels of cognitive complexity. For example, it is worth noting that there are real-world application areas in which the current generative AI technology is not a good fit, and ethical policy considerations are required for these areas. Within the scope of this case study paper, when we asked ChatGPT 4.0 if we had used irony on it in the conversation, ChatGPT 4.0 agreed but also asserted that it was a non-issue because "as an AI, I don't experience emotions or realizations like a human would in an ironic situation." Although current work on revising Bloom's taxonomy proposes adding collaboration to the higher levels of human cognitive complexity [16], to work well on collaboration, generative AI technologies would need even more architectural sophistication to implement affective ToM, which is distinct from cognitive ToM [15]. In the absence of any emotional capability, much less empathy, the most cognitively capable AIs, even if self-explaining or simply explainable AIs, can "cause serious psychological harm" (p. 29) [5]. This is why AI ethicists must carefully create and rigorously research ethics policies to regulate AI chatbots for human companionship and psychological counseling.

Conclusions

To make better progress at mitigating AI hallucinations, we recommend that researchers focus first on the Comprehending/Understanding and then the Application levels of cognitive complexity and, more immediately, on developing saturated benchmarks for systematically measuring cognitive performance improvements at these two levels. Such benchmarks should account for the following five dimensions of meaning to help mitigate AI hallucinations that (1) contradict other assertions in the generated text, according to open domain interpretations of words, phrases, and their synonyms and antonyms; (2) contradict other assertions in the generated text, according to domain-specific interpretations of words, phrases, and their synonyms and antonyms; (3) contradict other assertions that appear in the training corpus, according to open domain or domain-specific interpretations of words, phrases, and their synonyms and antonyms; (4) contradict assertions made by the human participant earlier in the

dialogue, particularly if the generative AI agent agreed with the human's earlier assertions; and (5) are verifiably incorrect according to the application of automatable open domain or, separately, domain-specific logic rules and computational processes. By saturating benchmarks for these dimensions at these cognitive levels, it should be possible to ameliorate the model overfitting that can occur if technological advancement occurs by means of teaching to the test. For example, although our test of the most recently released ChatGPT 4o (Omni) showed it to be unable to perform better than ChatGPT 4.0 on the specific numerical problem we have presented in this case study (in fact, some of its responses devolved into gibberish), a future version could be made to do better simply by adding specific pre-programmed numerical pattern analysis modules. This would make for a better demo, like performing well on an IQ test or solving an advanced calculus problem, as in the Application of understanding section, but it would not represent a meaningful advance along the pathway toward artificial general intelligence. Instead, new data structure designs are required to enable the machine learning of a more generalized understanding that can be applied to new problems of similar complexity but not previously seen in a training subset of the saturated benchmark.

The preceding five basic dimensions should also be regarded as table stakes for saturated benchmark development because we have demonstrated in this case study that abilities in the five dimensions are lacking in current generative AI technologies like ChatGPT 4.0 (via Copilot in Windows® set to its most precise mode). By themselves, the five dimensions are not sufficient but rather are necessary prerequisites for effective AI performance at any cognitive level in more advanced scenarios. Advanced benchmarks should test the use of (1) grammatical anaphors and cataphors, (2) types and attributes of entities, entity relations, and entity actions, (3) hyponyms and hypernyms, and extra-linguistic dimensions such as (4) the timeliness or temporality, and (5) the provenance of sources of contradictory assertions. Only once a generative AI technology has demonstrated advanced comprehension and application capabilities can future researchers then consider how to make cognitive computing agents that provide humans with intellectual lift in collaboration, analysis, synthesis, evaluation, and creative innovation.

References

1. Dutta S, Chakraborty T: [Thus spake ChatGPT](#). Commun ACM. 2023, 66:16-19. [10.1145/3616863](#)
2. Goth G: [Why are lawyers afraid of AI?](#). Commun ACM. 2023, 67:14-16. [10.1145/3631935](#)
3. [Michael Cohen says he unwittingly sent AI-generated fake legal cases to his attorney](#). (2023). Accessed: July 24, 2024: <https://www.npr.org/2023/12/30/1222273745/michael-cohen-ai-fake-legal-cases>.
4. Levine EV: [Cargo cult AI](#). Commun ACM. 2023, 66:46-51. [10.1145/3606946](#)
5. Storey VC, Lukyanenka R, Maass W, et al.: [Explainable AI](#). Commun ACM. 2022, 65:27-29. [10.1145/3490699](#)
6. Vaswani A, Shazeer N, Parmar N, et al.: [Attention is all you need](#). Adv Neural Inform Process Syst. 2017, 30:1-11.

7. [Half of business leaders say ChatGPT experience is more valuable than a college degree.](#) (2024). Accessed: April 10, 2024: <https://www.intelligent.com/half-of-business-leaders-say-chatgpt-experience-is-more-valuable-than-a-college-degree>.
8. Greengard S: [AI rewrites coding](#). Commun ACM. 2023, 66:12-14. [10.1145/3583083](#)
9. Roivainen E: [I gave ChatGPT an IQ test: here's what I discovered](#). Scientific American. 2023, 329:7.
10. Boyer JM, Boyer W: [Superficial intelligence: an AI odyssey with ChatGPT \(IN PRESS\)](#). [Telicom. 2024, 36:1-5](#). Telicom. 2024, 36:1-5.
11. Towhidul Islam Tonmoy SM, Mehedi Zaman SM, Vinija J, et al.: [A comprehensive survey of hallucination mitigation techniques in large language models](#). Comput Lang. 2024, 1:1-19. [10.48550/arXiv.2401.01313](#)
12. Armstrong P : [Bloom's Taxonomy](#). 2010.
13. [Steven Pemberton Facebook post of a ChatGPT conversation](#). (2023). Accessed: 2024-07-24: <https://bit.ly/42CNLOc>.
14. Roth G, Dicke U: [Evolution of the brain and intelligence in primates](#). Progress Brain Res. 2012, 195:413-30. [10.1016/b978-0-444-53860-4.00020-9](#)
15. Gabriel ET, Oberger R, Schmoeger M, et al.: [Cognitive and affective Theory of Mind in adolescence: developmental aspects and associated neuropsychological variables](#). Psychol Res. 2019, 85:533-53. [10.1007/s00426-019-01263-6](#)
16. R Heller Richard: [A new bloom - adding 'collaborate' to Bloom's taxonomy](#). J Learn Dev Higher Educ. 2022, N/A:-. [10.47408/jldhe.vi24.906](#)
17. Herrmann-Werner A, Festl-Wietek T, Holderried F, et al.: [Assessing ChatGPT's mastery of Bloom's Taxonomy using psychosomatic medicine exam questions: mixed-methods study](#). J Med Internet Res. 2024, 26:e52113. [10.2196/52113](#)
18. Chancellor S: [Toward practices for human-centered machine learning](#). Commun ACM. 2023, 66:78-85. [10.1145/3530987](#)
19. Lata K, Singh P, Dutta K: [A comprehensive review on feature set used for anaphora resolution](#). Artif Intell Rev. 2020, 54:2917-3006. [10.1007/s10462-020-09917-3](#)
20. Boyer JM: [Natural language question answering in the financial domain](#). Proc 28th Annual Int Conf Computer Sci Software Eng. 2018, 1:189-200.

Comments

[Nick Kadochnikov](#), Master of Science on Oct 09, 2024 at 09:49 AM

This paper provides a thorough and practical exploration of the challenges surrounding AI hallucinations and cognitive limitations in current generative AI models. As a practitioner, I am always looking for ways to reduce hallucinations, improve the accuracy of information retrieval, and optimize both speed and cost. This paper addresses these issues effectively, particularly through its focus on the higher cognitive functions of comprehension and application, which are crucial in improving AI performance. The key takeaway from this paper is the need for more sophisticated benchmarks that evaluate AI models beyond simple knowledge retrieval. The paper highlights examples where generative models, like ChatGPT, produce authoritative-sounding yet incorrect content. This underscores the importance of measuring higher cognitive levels, where understanding and the ability to apply knowledge correctly are essential to reducing hallucinations. For practitioners, this is a critical insight, as it informs how we can push AI models toward producing not just fluent, but factually accurate and contextually relevant outputs. While the paper provides valuable guidance on mitigating hallucinations, the discussion around retrieval-augmented generation (RAG) is especially useful. Techniques like these already show promise in reducing hallucinations at lower cognitive levels by combining retrieval and text generation. However, as the paper suggests, further work is needed to address hallucinations at higher cognitive levels. The proposed development of saturated benchmarks, designed to evaluate AI performance on more complex tasks, is a step in the right direction. This offers a clear path forward for those of us working to improve both the accuracy and reliability of generative AI in real-world applications. Although the paper touches on the importance of ethical considerations, its strength lies in its practical roadmap for advancing AI. This aligns well with the need to improve retrieval accuracy and reduce computational costs, which are central concerns for practitioners like myself. While frameworks like Recursive Abstractive Processing for Tree-Organized Retrieval (RAPTOR) could complement this by organizing data more efficiently and enhancing retrieval, the core focus on cognitive benchmarks is where this paper truly provides value. The paper offers actionable insights into how AI can be made more reliable, emphasizing the development of new benchmarks and strategies to address its cognitive limitations. For those of us working on the front lines of AI implementation, we desperately need pragmatic and practical guides to improving AI's capabilities in ways that directly benefit real-world use cases.

[John M. Boyer Dr.](#), Ph.D., Researcher on Oct 13, 2024 at 01:42 PM

Thank you for your thoughtful comments, Nick. We agree that our main intent was to highlight a roadmap for researchers and developers to focus on the cognitive levels, and specific dimensions within them, that best reflect the expectations of users. There is a notable gap between what the current generative AI technologies were designed to do versus the ways in which many people are being encouraged to or wanting to use them. The current benchmarks address improving the selection of words or content that should come next linguistically, whether or not reasonably. But many users want to be able to _rely_ on the reasonableness of the eloquent output. Some “bolt-on” methods like RAG and RAPTOR are improving factual accuracy at the retrieval/knowledge/recall level, and this helps reasonableness and hence reliability at that lowest cognitive level by generating content that aligns more closely and more reliably reflects

reasonable human-generated content. It's also specious to bolt on lots of human-made algorithms, such as for symbolic calculus, because it looks impressive at first, but there are no genuine abilities that reflect users' use cases around comprehension, application, analysis, and synthesis/innovation. So, just as doing well at generating text has taken new data structures and models (RNN and attention) and new benchmarks for measuring them, we assert that AI researchers will need to develop new models and benchmarks to reach higher cognitive levels (and their dimensions we describe in the article). By looking closely at how researchers in human development have described human cognitive capabilities, we find powerful language that AI researchers can use to organize future research agendas to simultaneously focus on the use cases while also sequencing the work. Finally, we did also find one important and unique matter of AI ethics that arose during the dialogues of the case study. While AI ethics researchers have many challenges to address, we felt it is important to emphasize that current generative AI technologies should not be used in "human companionship and psychological counseling" use cases due to the technological design having no affective theory of mind. Particularly counselors, but companions as well, are charged with advocacy, seeking justice for, and caring for the welfare of their clients, or fellow companions, and doing so requires emotional capability and empathy so that we can do so in ways that are collaborative and meaningful to the recipient of our care.

Respectfully submitted, John Boyer and Wanda Boyer

[Srikanth Tamilselvam](#) on Oct 09, 2024 at 10:56 AM

Very Insightful and Timely.

[Charles Wiecha](#), Ph.D., Prof. on Oct 20, 2024 at 12:53 PM

This paper presents a very timely and clear outline of the gaps between the expectations and hopes for LLMs on the one hand and the very real limits to their current actual reasoning and interaction capabilities. I would argue the paper is in fact broader than the title implies with its focus just on hallucinations. The gaps related to the higher level of cognitive functions, well outlined using the Bloom Taxonomy go far beyond hallucination and speak directly to the largely improper and unrealistic positioning of today's LLMs as general purpose human agents. There is simply too much required beyond predictive generation, even given the impressive and unexpected emergent behaviors we observe. Those emergent behaviors, as welcome as they are, still fall far short of what is required for extended, context, task, and user aware intelligent assistance. While the examples given in the paper perhaps lean a bit more into maths -- an area where I would not expect LLMs to be as strong as in other processing -- nonetheless there are plenty of use cases covered in the areas of language processing which complement the maths and support the conclusions of caution in current applications. My thanks to the authors and I am strongly considering having my class on LLM capstone projects use this reading to supplement their work!

[John M. Boyer Dr.](#), Ph.D., Researcher on Nov 01, 2024 at 02:05 PM

Thank you for your thoughtful comments, Dr. Wiecha. We quite agree that the key technical requirements problem is, to put it in your eloquent phrasing, "the largely improper and unrealistic positioning of today's LLMs as general purpose human agents." In turn, we also

firmly agree that emergent behaviors can be quite fun and welcome, except in the many cases that the unrealistic positioning exacerbates, i.e., when the user has the misunderstanding that they can rely on the technology for “context, task, and user aware intelligent assistance.”

We first noticed the cognitive disfunction of earlier versions of ChatGPT during conversations we mentioned in the paper about a chemical formula and the iambic pentameter pattern. Some improvements have been made at the lowest cognitive level, seemingly due to retrieval augmented generation (RAG), and those improvements sometimes help the technology to sidestep hallucinations at some higher cognitive levels. As a result, we sought the easiest way to factor out RAG, which is how we landed on a numerical pattern question (that, and it was the numerical analog of the verbal IQ test given in the Scientific American article we cited). This returned the spotlight to the reliability of the technology in use cases requiring any level of intelligence of the intelligent assistant.

We hope to highlight that the specific question was not so important as was factoring out RAG as a feature that seemed to be hiding the cognitive disfunction and unreliability. Future versions of the technology can and undoubtedly will be programmed with many selected algorithms to address specific question types. However, this will not truly increase reliability because doing so does not imbue such future versions with any more actual intelligence. Hence, failures will continue as soon as one’s questions step outside of what the selected additional algorithms cover. For example, in the paper we mention an advanced calculus problem we posed. ChatGPT 4.0 performed horribly on it because there was a fundamental aspect of partial derivatives and integrals that it did not “understand.” It was only because we knew how to solve the problem that we were able to guide the technology to solve the problem in a particular way. Only then was it able to bring some wonderful pre-programmed symbolic algorithms to bear. However, a user should not need to already know how to solve a complex problem in order to get the right help solving the complex problem.

[Srikanth Tamilselvam](#) on Nov 13, 2024 at 02:23 AM

Can the authors comment on the different answer to their main question that is given by the the recently released o1-preview of ChatGPT?

[John M. Boyer Dr.](#), Ph.D., Researcher on Nov 13, 2024 at 04:00 PM

This is a great question, thank you. At the end of the paper, we noted that ChatGPT 4o did not do any better on the question than Copilot’s 4.0 and, in fact, ChatGPT 4o gives only one initial answer differently, and then its second answer is the same as the o1-preview’s first answer.

The first answer from ChatGPT 4o was that the number to exclude was “3 and 5” because they were both odd numbers surrounded by even numbers, but since we asked for one number, it should be 3. We replied that we asked for which *one* number does not belong with *all* the others. This prompted the second answer from ChatGPT 4o, which was also the initial response from the o1-preview. They both said that *5* is the outlier from the rest of the list because “... it is odd and not divisible by 2 or 3.” This statement seems superficially better, yet one can immediately tell that 4o and the o1-preview do not seem to comprehend that odd numbers are not divisible by 2. More importantly, there seems to be a lack of understanding of the question

asking what is the number that should be removed from *the* list. In the elided part of the last quoted answer just above, the 4o and the o1-preview made two lists, the divisible-by-2 list and the divisible-by-3 list. When we replied that the answer was unsatisfactory because two separate patterns were identified rather than identifying one pattern that describes *the* list of numbers left over when 5 is excluded, ChatGPT agreed and then moved on to answers that were more like the Copilot 4.0 answers from the paper in terms of lacking comprehension-level and application-level cognitive abilities.

In the conclusion of the paper, when we said that 4o devolved into gibberish-- here is the last thing it said to us: "... the number that stands out uniquely in context is: *2* because it uniquely fits criteria amid even distinctly as smallest factor without being composite or multiple beyond others seen." That this unintelligible reply occurred is significant because sounding human is the one most important thing that LLMs are supposed to get right.

[Wanda Boyer](#) on Nov 13, 2024 at 04:04 PM

This is a great question, Srikanth, and I also thank you. In addition to the reply above from my co-author, I would also like to highlight a more general principle that is pertinent here. Namely, I would like to highlight that the responses of future versions of ChatGPT may indeed differ from those in the current paper, but this does not necessarily mean that the answers are possessed of higher cognitive functioning. At some point, numerical pattern algorithms may even be bolted onto the core LLM technology that will enable it to solve our particular question in this paper. These bolt-on algorithms will help future versions do better in specific scenarios, just as RAG did in our experiments with the chemistry question we discussed in the paper and just as symbolic algorithms helped ChatGPT to answer the advanced calculus question we posed to it. Yet these bolt-ons do not add any real comprehension, much less any higher level of cognitive ability. The current technological approach has the potential to cause harm to users because its answers are unreliable due to AI hallucinations at the comprehension level and higher. It is our hope that this paper, as a strategic roadmap, will inspire future researchers to create research agendas for developing new technological models and saturated benchmark test frameworks that reflect the higher cognitive abilities that today's marketing hype is causing users to expect.

[Vyacheslav Serhiiovych Kharchenko](#) on Dec 21, 2024 at 08:12 AM

Thank you very much for an interesting and important article from the point of view of getting rid of unnecessary illusions about AI and AI tools. The article talks about the hallucinations of AI itself, but it is equally important to get rid of hallucinations about the possibilities, the idea of its perfection, or rather, that it can be achieved. Last Wednesday, the use of LLM for critical software verification/testing was discussed at the department's weekly PhD-students seminar. When it comes to applications, it is important for safety/security, the completeness and trustworthiness of testing must be proven, and the verification tools themselves must meet strict requirements. Using LLM-based testing is supposed to reduce the "curse of size" for complex systems. But these means become even more complicated and then we have the effect of growth of this "curse". Thus, the problem becomes similar to the problem of "who will guard the guards." Toleration of hallucinations (detection, identification and limitation or exclusion of

exposure), determination of input uncertainty and potential dangerous behavior of AI must be clearly specified in the sense that its use on such data must be limited, simplified or form transitions to a safe (protected) state. I understand that my comment gravitates towards AI safety/security issues, but this issue is key to the development and implementation of AI-based systems. Vyacheslav (Slava) Kharchenko, Prof. and head of the Computer Systems, Networks and Cybersecurity Dept, National Aerospace University KhAI, Kharkiv, Ukraine

Wanda Boyer on Mar 15, 2025 at 02:27 PM

Dear Professor Kharchenko,

Thank you for your insightful and positive comments. We quite agree that it is important to constantly stress that perfection is impossible because all AI is based on inductive thinking (statistical inference) and so a “best fit” answer is necessarily going to be occasionally incorrect. This is especially important, as you mention, when AI is used in an area like software verification because the very point is to eliminate errors. One thing that may help is to gain clarity on the characterization of errors that come from human software verification. AI makes errors because of statistical misfits, but humans also make errors for other reasons. If the rate of human error is appropriately characterized and then an automated system can be shown to outperform that rate of human error, then progress has been made despite an AI model’s not being perfect—perfection being the enemy of progress.

The key problem for current expectations that are being placed on today’s generative AI technologies (indeed hallucinations, as you mentioned) is whether the technologies were even remotely designed to perform well at the tasks being expected of them in the marketing hype. If they weren’t, then they will be unlikely to outperform humans on the tasks (including software verification, as you mentioned).

Vyacheslav Serhiiiovych Kharchenko on Mar 15, 2025 at 03:35 PM

Thank you very much for your response/ comment. I completely agree that the nature and causes of errors (or wrong actions) of artificial intelligence and humans are different. The issue is more complicated when it comes to intentional actions (it seems that they should be distinguished from attacks on another system). But if such error-actions are unintentional, it is logical to talk about using the diversity principle in order to reduce the risks of simultaneous error. For many years, I have been working on reducing the risks of common-cause failures for reactor protection systems and other critical systems, where the principle of diversity is normative, mandatory for use. Therefore, it may be advisable to build such two- or multi-version "human-AI" systems, diversifying these backup channels as much as possible. In fact, such systems where AI is used to support decision-making already exist. At one time, British colleagues conducted research on such systems to support decision-making in medicine and came to the conclusion that in them the “final” intelligence “levels out”... But the situation is significant when it comes to real-time mode. We received a patent for a method of AI system redundancy, where use the principle of diverse duplicating AI1-AI2 and reported on it at a joint workshop with MIT <https://ceur-ws.org/Vol-3777/paper24.pdf>. Thus, it is advisable (?) to add a person to such systems and analyze how to reduce the risks of CCF... Very interesting... I remembered one of our international projects, where the option of implementing the Single Pilot Operation principle was considered, and the role of «the second» pilot was entrusted to AI, whose safety we had to

assess and ensure... It is very interesting to continue this topic on the diversity of natural-human and artificial intelligence to ensure safety/trustworthiness (for safety-security critical systems and not only) considering humanitarian issues (psychological,...)... Thank you again.

[Roger Dumas](#), Ph.D. on Mar 09, 2025 at 10:29 AM

In my opinion, most of the casual consumers of AI-generated mis-information are prone to confirmation bias and have a limited ToM. They tend to be uncritical, readily accepting inaccuracies and outright falsehoods, as long as these match their (mistaken) beliefs. How likely is it that the increasingly powerful AI-purveyors who benefit from the dissemination of nonsense will adopt your premise, rein-in their robots and strive toward more ethical and empathetic outcomes?

[John M. Boyer Dr.](#), Ph.D., Researcher on Mar 15, 2025 at 02:30 PM

Your question is a very good one, Dr. Dumas, thank you for asking. For starters, there is a difference between deliberate misinformation and the unintended misinformation from AI hallucinations. In other words, the misinformers are just as interested in accuracy as "informers" because they wouldn't want their AI generated content to hallucinate the correct information rather than the misinformation they intend to spread. Furthermore, the misinformers don't benefit much from AI hallucinations that may occur in the AI generators of the "informers" because those hallucinations are just as likely to be detrimental as they are instrumental to the misinformers.

In the long run, we think it is most likely that informers of all stripes will seek to generate more accurate information, i.e., information that more accurately represents their intentions, and purveyors of AI technology will therefore seek ways to achieve better accuracy than their competitors so that they can sell more software services.

It helps to pause a beat and consider who the paying customers are that will keep the sellers of software services earning their commissions. As the saying goes, if the product is free, then you are the product. In other words, the mass of "casual" users are a source of free testing, advertising, and marketing hype to help earn the brand recognition that leads to the real money. Some of the deepest pockets are going to be corporations and similar entities who want to use AI to lift the performance of their workforce, improve the velocity and relevance of their product research, etc. None of these folks are going to be interested in misinforming their workforce or researchers (But AI Hallucinations would be detrimental to them).

And because technology in and of itself is neither good nor evil, it is possible that the misinformers with deep pockets will spend their money on the most accurate systems for generating the (mis)information they want to generate. However, such disinformers don't even need to spend money on AI systems because they can do what they do now, and "Mechanical Turk" their way to the generated misinformation, while focusing their compute power on dissemination. These are reasons why fighting disinformers at the AI tech level is not going to be useful compared to fighting them when and where they spread the disinformation (which is why cutbacks to fact-checking at FB and especially Twitter are so harmful).

The sidebar about customers being finished, we can return to the effects of competition on the AI developers. A thing that commonly happens in technology is that tech people often miss two things. First, they don't necessarily consider any precedents for the technical capabilities nor any technical design deficiencies that may prevent what they have from achieving the higher accuracy levels they desire. So then, they do the second thing, which is that they make up benchmark test suites to measure improvements in whatever they end up deciding to build. The point of our paper, then, is just to say that if you **really** want to increase accuracy (reduce AI hallucinations) in generative AI, then it would be a good idea to look at the precedent set by the cognitive levels evident in actual human biological intelligence and then make benchmarks aligned with those cognitive levels, so that the accuracy measures are immune to and unbiased by the technical approach for achieving them.