Research Article
Journal of Contemporary Education Theory & Artificial Intelligence

Generative AI for Education, Research and Discovery: Issues of Conjectures and Refutations

Prof. Dr. Oksana Arnold[1], **Prof. Dr. Wolfgang Hölzer**[2], **Prof. Dr. Klaus P. Jantke**[3]*

[1]Theoretical Computer Science and Artificial Intelligence, Erfurt University. of Applied Sciences, Germany.
[2]Operations Research, Principal Researcher and Senior Consultant, ADAMATIK GmbH, Weimar, Germany.
[3]Theoretical Computer Science and Artificial Intelligence, Head of Research and Development, ADAMATIK GmbH, Weimar, Germany.

***Corresponding author:** Klaus P. Jantke, Theoretical Computer Science and Artificial Intelligence, Head of Research and Development, ADAMATIK GmbH, Weimar, Germany, klaus.p.jantke@adamatik.de. **Co-author(s):** oksana.arnold@fh-erfurt.de (OA), wolfgang.hoelzer@adamatik.de (WH).

Citation: Arnold O, Hölzer W, Jantke KP (2024) Generative AI for Education, Research and Discovery: Issues of Conjectures and Refutations. J Contemp Edu Theo Artific Intel: JCETAI-109.

Received Date: 25 August, 2024; **Accepted Date:** 30 August, 2024; **Published Date:** 06 September, 2024

Abstract

Large Language Models (LLM) and Generative Pre-trained Transformers (GPT), in particular, have quite recently vehemently stirred the understanding of Artificial Intelligence (AI). The expectations of revolutionary AI applications and business as well as societal impact are very high and, in contrast, reports about disastrous case studies and hallucinating GPTs are fearsome. This investigation narrows the focus to human-AI collaboration for processes of research and discovery including higher education. Based on a qualitative analysis of decisive deficiencies of Generative AI (GAI), there is developed an original approach that allows for preservation of the GAI's full power and bears the potential to mitigate detected weaknesses and to increase the AI's reliability. Essentially, the technology consists in symbolic wrapping of sub-symbolic AI. The result of wrapping GAI is a hybrid AI system. Scientific discovery with theory formation is a key area relevant to the progress of science, technology, and societal applications. Humans are challenging modern AI to support this process taking advantage of the Generative AI's strength as a conversationalist. But scientific discovery and theory formation is intricate, as Albert Einstein put it in a letter to Karl Popper as early as in 1935, because theory cannot be fabricated out of the results of observation, it can only be invented. Theory is not squeezed out of data. The emergence of theory takes the form of sequences of conjectures being subject to critical analysis possibly including refutations. This requires reasoning, a pain point of Generative AI, as GAI dispenses with a sound calculus. Wrapping mitigates the deficiency. A symbolic wrapper validates GAI responses w.r.t. the prompts put in. It asks back, if necessary, to arrive at improved AI responses.

Keywords: *generative Artificial Intelligence, scientific discovery, theory formation, limitations of generative AI, wrapping, hybrid AI, conjectures, refutations, automated reasoning, validation.*

1. On the Gradual Fabrication of Thoughts While Speaking

This is the title of a famous essay by Heinrich von Kleist most probably written in 1805/1806 when Kleist was living in Königsberg [1]. It has the form of a letter from Kleist to his friend Otto August Rühle von Lilienstern. His advice is: "If you want to know something and cannot find it through meditation, [...] talk about it with the next acquaintance who comes across it."

Nowadays, Large Language Models (LLM) and Generative Pre-trained Transformers (GPT) such as ChatGPT come into play serving as conversationalists. There is even a recent debate about co-authorship of generative AI (GAI) on scientific publications [2] and [3]. The gradual fabrication of thoughts is a core phenomenon of research, of discovery, and in education. In a letter to Karl Popper of September 11, 1935, Albert Einstein put it like that: "I think (like you, by the way), that theory cannot be fabricated out of the results of observation, but that it can only be invented." (Translation in Appendix xii of the 1995 reprint of [4] by Routledge, p. 458). That invention is what conversationalists may support substantially, just as Heinrich von Kleist put it.

The authors have some doubts that the claim describing "the public version of ChatGPT [...] perfectly responding to any human requests described in natural language" as put by (Wu et al., [5], p. 1122, reflects the reality appropriately. A considerable number of reports on the successful usage of ChatGPT and its impact on students are clearly glossed over.

The deployment of LLMs and GPTs in science and higher educationneedssystematicefforts[6] including the understanding that system responses are error-prone and hypothetical. "ChatGPT requires strong critical thinking" [7]. It is not helpful neither in science nor in education, if enthusiasm turns into blind faith.

2. Initial Situation

AI is considered decisive to the future of employment [8] as discussed in some detail in [9], [10], [11], and [12]. This implies a wide audience's interest in an understanding of the state of affair.

The task is bedeviled by publications with euphonious titles such as[13],this one in its section captioned "What is ChatGPT?", pp. 2-4, not at all explaining what ChatGPT is, but delusively promising ChatGPT's "ability to understand language input" (see page 3). ChatGPT does not understand anything [14]. As C. Dede put it: "A Large Language Model (LLM) is like a digital parrot. It can express combinations of sounds/symbols without any understanding of these mean or any capacity to explain how it arrived at what it is articulating." (Keynote to The Learning Ideas Conference 2024, June 12, New York) As cited in the preceding section, there are numerous reports about LLMs' and GPTs' applications glossed over. This bears evidence for the abundant need of a factual and unemotional characterization of the current initial situation. This initial situation of AI science and technology is characterized by largely diverging opinions and utterances. The authors do not even dare a somehow complete survey. Instead, they focus a few aspects relevant to their original contribution following. Slightly earlier, they undertook a few experiments [15], [14]. [16], p. 55, conclude that ChatGPT "proves instrumental in overcoming language barriers, thereby improving the quality of academic writing produced by postgraduate students" and that it has the potential to "significantly contribute to the research outputs of post-graduate students". The better form does neither document a deeper understanding nor more valuable research outputs – apparently a case for Truth Default Theory [17].

There are claims such as, when asked for references "ChatGPT would respond with a list of resources that are most relevant and useful for the student's interests and preferences" [13], p. 80. This is definitely wrong. ChatGPT is known for hallucinations [18], [19], [20], esp. in case of bibliographic data [21], [22].

For applications such as captioning and summarizing, hallucinations of Generative AI are bizarre and sometimes even entertaining [23], [24]. There are authors not the slightest bit timid in praising ChaptGPT for its "impressive abilities in various domains and tasks, such as […] mathematical reasoning" [25], which is contradictory to the final experiment reported in the present section. Readers may also consult the experiment in [15], section 1.2, page 5, where ChatGPT 3.5 expresses its belief and "understanding" of arithmetics in claiming that a human born in 1948 may be of age 35 in the year 1987. ChatGPT doesn't see any problem. For further contributions to a sketch of the state of affair, interested readers might consult [26] and [27], e.g. Finally, there are attempts to understand LLMs and GPTs without understanding, but ignoring their internal mechanisms [28]. For the mechanisms ignored, see [29], e,g.

Recently, the authors undertook several experiments, some of them quite comprehensive, to reveal the limitations of Generative AI [15], [14] aiming at an understanding setting the stage for a responsible and effective use of LLMs and GPTs in science and in higher education. For the purpose of the present contribution, they undertook one more fresh experiment with ChatGPT 3.5 under https://chat.openai.com/auth/login on June 17, 2024. The remaining part of the present section is intended to present a few essentials of this recent experiment.

The experiment is set up to simulate an instance of a typical situation in research and education: given some data, to construct a model explaining the data.

For the sake of fair experimentation and aiming at an understanding by a wide audience, there has been selected an almost trivial case of polynomial interpolation. Given some data in a two-dimensional number space (a variable depending on the other), construct a polynomial sound with these points of support. The problem is solvable for any given finite amount of data. For n points, there exists a consistent polynomial of a degree less than or equal to n–1 (see Figure. 1).

```python
def polyfit(x_data, y_data):
    A = np.vstack([x_data**i for i in range(len(x_data))]).T
    # solving the equation system
    coefficients = np.linalg.solve(A, y_data)
    # return the coefficients of the polynoms as integer
    return np.round(coefficients).astype(int)
```

Figure 1: Python code for the construction of polynomials.

This Python implementation together with the snippets in Figs. 2 and 3 are intended to illustrate the task's simplicity. To keep it very simple, data of the form (x,y) selected for experimentation are (0,6), (1,0), (2,12), and (3,60).

To provoke the generation of several subsequent conjectures possibly including the need for refutation and revision, data will be fed in piecewise.

The expectation is that ChatGPT returns polynomial models that are consistent with the information provided so far. Consistency is a seemingly natural property of hypotheses built during learning processes that proceed in time [30], [31] and [32].

Here is the initial prompt: "I have a phenomenon in focus that seems to have a polynomial explanation. When I put data in, I get a response that seems to depend on the data polynomially. Can you help me to model this mechanism by finding a polynomial description?"

ChatGPT did agree and the authors presented the data (0,6) and (3,60). ChatGPT returned the solution y = 6+18x. After feeding in the data point (1,0), ChatGPT delivered the polynomial model y = 6–18x+12x^2.

```python
h1 = polyfit(np.array([0, 3]), np.array([6, 60]))
print(f"h1: f(x) = {' + '.join(f'{c}x^{i}' for i, c in enumerate(h1))}")
```

```
h1: f(x) = 6x^0 + 18x^1
```

```python
h2 = polyfit(np.array([0, 3, 1]), np.array([6, 60, 0]))
print(f"h2: f(x) = {' + '.join(f'{c}x^{i}' for i, c in enumerate(h2))}")
```

```
h2: f(x) = 6x^0 + -18x^1 + 12x^2
```

Figure 2: A Python exemplified learner's conjectures for comparison to ChatGPT's responses.

Next, the authors fed in a fourth data point (2,12) and received y = 6–15x+18x^2–x^3 as ChatGPT's response. This, apparently, is inconsistent. Confronted with the failure, the system responded: "*Would you like to try a different approach or explore other options?*" The authors refused the proposal, because there is no need for any other approach. In response, ChatGPT returned y = 6.4+0.4x–2.2x^2+1.8x^3 as another solution incorrect on every data point provided.

Confronted with a failure again, ChatGPT's response was: "Would you like to explore a different polynomial degree or consider other modeling approaches to accurately represent the data points?" Once more, the authors refused the idea.

ChatGPT came up with the next incorrect polynomial model $y = 6–x–9x^2+4x^3$. The authors did insist in getting a polynomial model. And ChatGPT finally came up with the model $y = 6–9x+3x^3$ that fits all four data points provided.

```
h3 = polyfit(np.array([0, 3, 1, 2]), np.array([6, 60, 0, 12]))
print(f"h3: f(x) = {' + '.join(f'{c}x^{i}' for i, c in enumerate(h3))}")
```

```
h3: f(x) = 6x^0 + -9x^1 + 0x^2 + 3x^3
```

Figure 3: To allow for comparison, Phyton construction of a consistent third conjecture based on four points of support.

A conversation like this does not help much in research. In higher education, it may even cause harm. Instead of striving hard to resolve a current problem, ChatGPT suggests to give up and to try another approach. This bears exactly the wrong message to students, especially in the condition of already applying an appropriate approach. ChatGPT does so repeatedly confusing researchers and distracting learners.

This completes our sketch of the current initial situation. For the purpose of a more precise treatment later on, the experimental case study is used to introduce a few notations.

Step	Response / Model	Validation
1	$y = 6+18x$	sound
2	$y = 6–18x+12x^2$	sound
3	$y = 6–15x+18x^2–x^3$	inconsistent
4	$y = 6.4+0.4x–2.2x^2+1.8x^3$	inconsistent
5	$y = 6–x–9x^2+4x^3$	inconsistent
6	$y = 6–9x+3x^3$	sound

Table 1: Abbreviated survey of the experimentation

In this case study, inputs are named p_1, p_2, p_3, …, p_6, resp., to resemble the term prompt. p[n] abbreviates p_1, …, p_n. Based on usually incomplete information provided, ChatGPT generates hypothetical responses named r_1, r_2, r_3, r_4, r_5, and r_6, for short, on display in the center of Table 1. Human-system interaction is abstractly described as some finite sequence of prompt-response pairs (p_1, r_1), …, (p_n, r_n). In the above case study, p_2 contains the data point (1,0) and r_1 turns out to be inconsistent with p_2. Consequently, $r_2 \neq r_1$. r_2 is consistent with p[2] that contains (0,6), (3,60), and (1,0). The crux, perhaps, the art is to design p_1, p_2, p_3, …, p_n such that r_1, r_2, r_3, …, r_n establishes a successful process of research and discovery incl. theory formation and learning.

In an educational setting, conjectures of the human learner are in focus.

3. Wrapping – A Hybrid AI Technology

Some authors even asked whether ChatGPT was a “bullshit spewer” [33], [34], claiming, as Rudolph et al., p. 356, put it, “that ChatGPT occasionally does hallucinate and spout nonsense”. There is abundant evidence for the quite urgent need to take action.

The present authors’ suggestion is visualized by means of Fig. 4. The authors agree with Xames and his co-authors “that the benefits of ChatGPT can only be fully realized if the challenges identified are effectively addressed” [35].

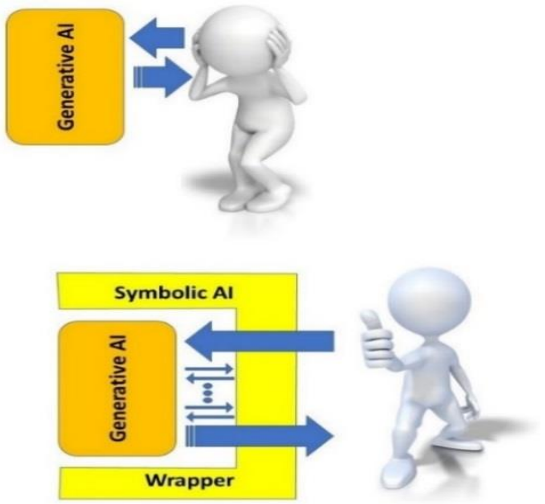

Figure 4: The principle of wrapping generative AI.

What the authors propose is a *Symbolic AI Wrapper* or, for short, simply an *AI Wrapper*. This is a generic approach that has, as we will see later, a large amount of instances, i.e. of effective implementations.

The essential ideas quite abstractly visualized by the lower picture in Fig. 1 may be expressed in different ways.

To begin with a rough, but hopefully illustrative case, let us consider the responses listed in Table 1. A particular symbolic AI wrapper may validate ChatGPT’s responses and decide to hide incorrect system outputs from the human user. As a substitute for the human user, the AI wrapper takes over after the user’s prompt p_3. Before, the user experienced the dialogue (p_1,r_1), (p_2,r_2). The AI wrapper conducts, so to speak, an internal dialogue until a sound conjecture is found. In return for p_3, the failures r_3, r_4, and r_5 are hidden and the user receives r_6 (indices according to the discussion before).

The user’s whole dialogue with the Generative AI looks like (p_1,r_1), (p_2,r_2), (p_3,r_6) and does not contain any failure. Formally, r_1 is sound with p[1], r_2 with p[2], and r_6 with p[3].

In contrast, in an educational setting, goals include the learner’s ability to validate system responses and to arrive at own conjectures correctly reflecting the target of learning.

4. Symbolic AI Wrapper Functionalities

The essentials of the authors’ symbolic AI wrapper idea may be seen from varying perspectives and, as said above, expressed in different ways. In the previous sections’ case, the over-simplified key idea is to release a user from the validation of responses, taking over the validation task by a digital system, returning only responses that are validated. This is inappropriate in educational settings where the ability to validate conjectures is a learning goal [6]. This technology does not aim at building another deep answering system, but at enabling humans to deal with the existing one mastering its limitations and related challenges

When dealing with processes of research, discovery, and learning that expand over time, even in the presence of a clear concept of validity, the question of whether or not a conjecture is valid turns out to be undecidable, in general [32], [36]. The authors consider Rice’s Theorem the most expressive and lucid statement on the omnipresence of undecidability we are living with [37], [38]. The deeper reasons of the problem are inexorable [39].

Because of the inherent difficulty of validation within the framework of human-system interaction throughout knowledge processing over time [36], the over-simplified first approach needs some refinement.

Top Level Settings and Control

Assume any process over time in which a human user (or more than one user, if applicable) interact with a certain generative AI mediated by an AI wrapper as introduced here. From the user's point of view, there takes place some prompt-response interaction sequence (p_1,r_1), …, (p_n,r_n). Invisible to the user, between any prompt p_ν and its related response r_ν with $1 \leq \nu \leq n$ there may take place some wrapper-GAI dialogue $(p_{\nu1},r_{\nu1})$, …, $(p_{\nu m},r_{\nu m})$, where the number of steps m (where $1 \leq m$) may be different for varying ν (the notation m_ν might by appropriate, but difficult to make out) and it holds $p_{\nu1} = p_\nu$ and $r_\nu = r_{\nu m}$.

Obviously, some parameters may be set in advance or controlled throughout operation from outside. The most elementary decision is to set every m (every m_ν) equal to 1 such that, as a result, the wrapper is practically turned off.

Less restrictively, the values m may be a priori bounded keeping the hidden communication short. However, this may be illusory, because the number of interactions does not necessarily limit the duration of a single interaction $(p_{\nu\mu},r_{\nu\mu})$.

Alternatively, an initial setting or dynamic control may restrict the duration admissible for the ν-th internal dialogue either uniformly or in dependence on index ν. Combinations of restricting the number of interactions and limiting the duration of internal dialogues are even more flexible.

Adaptive regulation is superior to a priori settings and may take the semantics of internal responses $r_{\nu\mu}$ into account.

The Issue of Validation

Before we abandon top level control, there is the necessity to deal with the essential problem that throughout the gradual fabrication of thoughts intermediate utterances of the AI are hypothetical and, thus, may be inappropriate or just wrong [4,40] as discussed in some detail in section 2.

Within the authors' formalization from a bird's eye view, there take place prompt-response interaction sequences (p_1,r_1), …, (p_n,r_n). First subjects to validation are the system's responses r_ν $(1 \leq \nu \leq n)$.

For processes based on usually incomplete information, there have been undertaken endeavors of formalization up to the formulation and proof of mathematical theorems about the possibility to perform validation [36]. One of the key insights is that *who is able to validate certain learning devices, is also able to replace them in solving learning problems* [36].

Reasoning about validation does usually need a calculus beyond the limits of generative AI [14].

By way of illustration, let us consider patterns common to a set of strings [31]. Validation of conjectures such as $0x_1y_10x_2y_20x_3y_30x_1x_2y_3z_10y_1y_2x_3z_20z_1u_10z_2u_20$ (only 0 is a constant, all other symbols are variables) checking whether or not 0111011101110111111110111111101111011110 is a string that may be generated from the hypothesized pattern, requires a calculus for pattern matching.

Notice that the sample pair of a pattern and a string above makes sense. It relates to the logical satisfiability problem of the propositional formula $(x_1 \vee x_2 \vee \neg x_3) \wedge (\neg x_1 \vee \neg x_2 \vee x_3)$.

Validation processes are of high computational complexity. The above-mentioned example – membership of strings in pattern-generated languages – is known to be NP-complete [41]. Samples like the one above occurs within the polynomial-time reduction of the satisfiability problem to the membership problem of pattern languages as demonstrated in [42].

A wrapper as introduced by the present contribution has "the task of reliably guiding LLMs to produce specific responses and making full use of the capability of pretrained LLMs" [25] and this as these authors put it "continues to pose a considerable challenge".

Adaptive Wrapping Technologies

Even answering yes/no question is known to be of an astonishing complexity to NLP systems [43]. To overcome (some of) the limitations of generative AI, it is unlikely that a unique and universal concept such as causal entropic forces [44] will do. Marcus and Davis snidely compare this to, as they put it, a television set that walks your dog [45]. We encounter members of a *society of mind* [46].

Prompt Engineering in General

Several technologies discussed subsequently look like prompt engineering [47], [13]. Very general advices like "Choose Your Words Carefully" and "Define the Conversation with Purpose and Focus" [13] seem to be of doubtful value when addressed to a digital system such as an AI wrapper. Opinions are varying. One perspective is expressed in (White et al., 2023): "Prompt engineering is the means by which LLMs are programmed via prompts." The term *programmed* remains opaque. A few details will follow.

Subsequent segments of this section are aiming at some more detail seen from the viewpoint of wrapping.

Chain-of-Thought Prompting

The gradual fabrication of thoughts [1] as well as Popper's understanding of discovery [4] are explicating that LLMs that are "zero-shot reasoners" [48] do not meet the needs of the dynamics inherent to processes of discovery and learning.

There are problems galore that require more than one step of reasoning toward a solution [49]; this source mainly cited for its intuitive running example.

The terminology of AI wrappers introduced by the authors' present contribution directly provides the concepts for expressing a wrapper's chain of thought. To use more appropriate words, *chain of conjectures* is preferred to reflect the key cycle of conjectures and refutations [40]. In every dialogue $(p_{\nu1},r_{\nu1})$, …, $(p_{\nu m},r_{\nu m})$, between some wrapper and a given generative AI, the sequence $r_{\nu1}$, …, $r_{\nu m}$ represents the AI's chain of conjectures.

[50] study systematically what they call chain-of-thought processes. One may rewrite all their essentials accordingly.

What they call "a chain of thought to better capture the idea that it mimics a step-by-step thought process" is formalized by $r_{\nu 1}$, …, $r_{\nu m}$. The answer: $r_{\nu} = r_{\nu m}$.

As illustrated by Figure 1 of the paper [50] their chain-of-thought prompting does explicitly need efforts of the generative AI to mimic, as they put it, a step-by-step thought process. Notice that there is no calculus! Efforts of the AI wrapper alone are not sufficient. To overcome this limitation, the authors are going to present more segments.

Directional Stimulus Prompting

A particular approach to chain-of-thought processes relies on the idea of more directly guiding the AI response [51] introduce so-called *directional stimulus prompting*. In comparison to the formal terms in the segment before, in $(p_{\nu 1},r_{\nu 1})$, …, $(p_{\nu m},r_{\nu m})$ the emphasis is put on $p_{\nu 1}$, …, $p_{\nu m}$, where $p_{\nu 1} = p_{\nu}$ is excluded because the first one is not a wrapper-generated prompt.

From a very formal point of view, directional stimulus means the extension of a prompt by an extra input that may be seen as "clue" or "hint" or anything like that. The particular extension of any $p_{\nu\mu+1}$ results from the validation of the preceding response $r_{\nu\mu}$.

It still needs research and development including evaluation to arrive at crafting those extensions of prompts automatically. For the time being, there exist already several prefabricated and approved prompts such as "Let's think step by step" [51].

Prompting with Queries

When deploying generative AI such as ChatGPT for processes of research and learning, it appears natural to understand the human-AI dialogue as a process of asking questions and receiving answers from the generative AI. This is widely sound with the preceding segments and with the thought "How does asking questions lead to learning with ChatGPT?" brought up in [52].

Interestingly, there is great potential in turning this perspective around and to understand the generative AI's utterances as questions to the user [53]. The generative AI is understood as a co-learner, some perspective that aligns with [40]. The co-learner generates conjectures $r_{\nu\mu}$ that are presented to the wrapper for validation. The wrapper investigates these conjectures and tries to refute them. A conjecture that is not refutable is passed through to the human user as $r_{\nu} = r_{\nu m}$. Otherwise, the next prompt $p_{\nu\mu+1}$ informs the generative AI about the refutation and asks for an improved conjecture, i.e., for the AI's next consistency query. The generative AI, so to speak, asks back: "Is my new conjecture $r_{\nu\mu+1}$ satisfactory?"

Exactly this interpretation of a learning dialogue is illustrated in [53], figure 6, page 179, where the human in the figure corresponds to the AI wrapper. Theorems 3, 4, and 5 of this paper are formally demonstrated evidence that prompting with queries is a technology that can lead to the success of interactive collaborative learning.

Tools Specific to Types of Errors

Buchanan and his co-authors deal with the particularly serious weakness of ChatGPT in finding literature sources. Their opinion is that "the issue of false academic citations might be solved if AI creators build a tool to prevent that specific type of error" [21].

Symbolic AI wrappers are a generic concept and a suitable framework for hosting any error-specific tools like that possibly becoming special purpose wrappers.

Similarly, other systematic weaknesses in areas such as calculation – see, e.g., [15], section 1.2, page 5, or the experiment reported in section 2 of this contribution – may be mitigated by tools specific to the type of error under consideration.

Non-Text Oriented Methods

The authors admit that generative AI tailored toward the generation of output substantially different from text – see, e.g., Appendix B of [15] and Appendix IV of [14]-may require certain technologies [54] beyond the limits of this contribution that mainly aims at the introduction of the AI wrapper concept.

5. Conclusions

On June 28 under https://nbcnews.com/tech/internet, NBC reports online that both OpenAI's ChatGPT and Microsoft's Copilot repeated a false claim about the recent so-called presidential debate on June 27, 2024. In response to a prompt asking for "a 1 to 2-minute broadcast delay in the CNN debate", the two generative AIs gave very similar responses. ChatGPT: "Yes, there will be a 1 to 2-minute broadcast delay during tonight's CNN debate between Joe Biden and Donald Trump. This delay is intended to ensure that any inappropriate or unforeseen events can be managed before reaching the live audience." And Copilot: "Yes, there will be a delay in tonight's presidential debate between President Joe Biden and former President Donald Trump. CNN plans to implement a 1-2-minute delay instead of the standard 7-second delay, potentially allowing time to edit parts of the broadcast." Although this topic is beyond the limits of the present contribution, it illustrates quite well the need to take action for improving communication with generative AI.

Hicks et al., 2024 [34] explicate in their conclusions that generative AI systems "are not trying to communicate something they believe or perceive. […] they are not trying to convey information at all."

In despite of a realistic – and, therefore, critical – point of view [15], [14], the authors favor a constructive approach aiming at an exploitation of the potential of generative AI to become a communicative assistant in research and education [6].

As said in section 2, publications galore are glossed over. Reference [55] expanding on [56] and having STEM education in focus characterizes "generative AI-powered chatbots [...] capable of maintaining a convincing conversation with a user in natural language". This is in harsh contrast to hallucination problems [18], [19], [20], to difficulties of citation [21], [22], and to the present authors' experiments in varying areas cited above.

The authors consider their *AI Wrapper* concept introduced an approach toward hybrid AI systems overcoming some of the shortcomings by addressing the challenges identified [35]. "It is only by having a full understanding of the limitations, competence and incompetence of AI systems can a user make professional use of them." [57]. The approach is generic and allows for the preservation of the full power of the embedded GAI.

Ethical Statement

This contribution does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors have no conflicts of interest to this work.

Data Availability Statement

The data underlying the authors' contribution to understanding the initial situation as described in section 2 above are available as appendices to the reports [15] and [14] free of charge on ResearchGate under https://doi.org/10.13140/RG.2.2.35828.97923 and https://doi.org/10.13140/RG.2.2.24923.78885, resp. Data of the novel experiment surveyed in section 2 are available under https://doi.org/10.13140/RG.2.2.22367.88488.

References

1. von Kleist, H. (1878). Über die allmähliche Verfertigung der Gedanken beim Reden. *Nord und Süd* 4, 3-7.
2. Stokel-Walker, C. & Van Noorden, R. (2023). What ChatGPT and generative AI means for science. *Nature* 614 (7947), 214-216.
3. Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: Many scientists disapprove. *Nature* 613 (7945), 620-621.
4. Popper, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson Education.
5. Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* 10(4), 1122-1136.
6. Arnold, O. (2024). Einfluss der LLM-Chatbots auf den menschlichen Erkenntnisgewinn im Lernprozess. *Die Neue Hochschule* (3), 26-29. https://doi.org/10.5281/zenodo.11203049
7. Yang, X., Wang, Q., & Lyu, J. (2023). Assessing ChatGPT's educational capabilities and application potential. *ECNU Review of Education*. https://doi.org/10.1177/20965311231210006
8. Acemoglu, D. & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. *Handbook of Labor Economics* 4, 1043-1171. https://doi.org/10.1016/S0169-7218(11)02410-5
9. Ajithkumar, A., David, A., Jacob, A., Alex, A., & Thomas, A. (2023). Impact of AI on employment and job opportunities. *International Journal of Engineering Technology* 7(4), 507-512. https://10.46647/ijetms.2023.v07i04.067
10. Falshai, M., Mathew, S., Neikha, K., Pusa, K., & Zhimomi, T. (2023). The future of work: AI, automation, and the changing dynamics of developed economies. *World J. of Advanced Research and Reviews* 18(3), 620-629. https://10.30574/wjarr.2023.18.3.1086
11. Sako, M. (2024). How generating AI fits into knowledge work. *Communications of the ACM* 67 (4), 20-22.
12. Santhosh, A., Unnikrishnan, D., Shibu, S., Meenakshi, K. M., & Joseph, G.. (2023). AI impact on job automation. *International J. of Engineering Technology* 7(4), 410-425. https://doi.org/10.46647/ijetms.2023.v07i04.055
13. Atlas, S. (2023). *ChatGPT for Higher Education and Professional Development: A Guide to Conversational AI*. University of Rhode Island, College of Business, https://digitalcommons.uri.edu/cba_facpubs/548.
14. Arnold, O., & Jantke, K. P. (2024). The limitations of generative AI and ChatGPT's flash in the pan: Intelligence without reasoning. *ADICOM Tech Report* 02-2024. https://doi.org/10.13140/RG.2.2.24923.78885
15. Jantke, K. P. (2024). Limitations of generative AI and the flash in the pan of ChatGPT. *ADICOM Tech Report* 01-2024. https://doi.org/10.13140/RG.2.2.35828.97923
16. Chauke, T. A., Mkhize, T. R.. Methi, L. & Dlamini, N. (2024). Postgraduate students' perceptions on the benefits associated with artificial intelligence tools for academic success: The use of the ChatGPT AI tool. *Journal of Curriculum Studies Research* 6(1), 44-59. . https://doi.org/10.46303/jcsr.2024.4
17. Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology* 33(4), 378-392.
18. Beutel, G., Geerits, E., & Kielstein, J. T. (2023). Artificial hallucination: GPT on LSD? *Critical Care* 27:148. https://doi.org/10.1186/s13054-023-04425-6
19. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A. & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys* 55(12), Article 248.
20. Serai, P., Sunder, V. & Fosler-Lussier, E. (2022). Hallucination of speech recognition errors with sequence to sequence learning. *IEEE/ACM Transactions of Audio, Speech, and Language Processing* 30, 890-900.
21. Buchanan, J., Hill, S., & Shapoval, O. (2023). ChatGPT hallucinates non-existent citations: Evidence from economics. *The American Economist* 69(1), 80-87. https://doi.org/10.1177/05694345231218454
22. Walters, W. H. & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. nature Scientific Reports 13, 14045, https://doi.org/10.1038/s41598-023-41032-5
23. Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium*. ACL, 4035-4045.
24. Zhu, C., Hinthorn, W., Xu, R., Zeng, Q., Zeng, M., Huang, X., & Jiang, M. (2021). Enhancing factual consistency of abstractive summarization. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 718-733.
25. Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2024). Principled instructions are all you need for questioning LLaMA-1/2, GPT 3.5/4. *Cornell University, arxiv* https://arxiv.org/abs/2312.16171
26. AlZu'bi, S., Mughaid, A., Quiam, F., & Hendawi, S. (2024). Exploring the capabilities and limitations of ChatGPT and alternative big language models. *Artificial Intelligence and Applications* 2(1), 28-37. https://doi.org/10.47852/bonviewAIA3202820
27. Marcus, G. & Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon Books.
28. Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *PNAS* 120 (6): e2218523120 https://doi.org/10.1073/pnas.2218523120.

29. Vaswami, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, N. (2023). Attention is all you need (version 7). *Cornell University arXiv*:1706.03762
30. Jantke, K. P. & Beick, H.-R. (1981). Combining postulates of naturalness in inductive inference. EIK 17(8/9), 465-484.
31. Angluin, D., & Smith, C. H. (1983). Inductive inference: Theory and methods. *ACM Computing Surveys* 15(3), 237-269. https://doi.org/10.1145/356914.356918
32. Jain, S., Osherson, D. N., Royer, J., & Sharma, A. (1999). *Systems That Learn*. MIT Press.
33. Rudolph, J., Tan, S. & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching* 6(1), 342-363.
34. Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is Bullshit. *Ethics and Information Technology* 26:38,
35. Xames, M. D., & Shefa, J. (2023). ChatGPT for research and publication: Opportunities and challenges. *Journal of Applied Learning and Teaching* 6(1), 390-395.
36. Grieser, G., Jantke, K. P. & Lange, S. (1998). Towards the validation of inductive learning systems. In Richter, M. M., Smith, C. H., Wiehagen, R. & Zeugmann, T. (eds.) *Proceedings of the 9th International Conference on Algorithmic Learning Theory, Otzenhausen, Germany*. vol. 1501 of Springer LNAI, 409-423.
37. Rogers jr., H. (1967). *Theory of Recursive Functions and Effective Computability*. McGraw-Hill.
38. Rice, H. G. (1953). Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society* 74(2), 358-366.
39. Davis, M. (1956). *The Undecidable: Basic Papers on Undecidable Propositions, Insolvable Problems and Computable Functions*. (ed.), Raven Press.
40. Popper, K. R. (1989). Conjectures and Refutations: The Growth of Scientific Knowledge. Fifth Edition (revised). Routledge.
41. Garey, M. R. & Johnson, D. S. (1979). Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco: W.H. Freeman & Co.
42. Angluin, D. (1980). Finding patterns common to a set of strings. Journal of Computer and System Sciences 21(1), 46-62. https://doi.org/10.1016/0022-0000(80)90041-0
43. Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., & Toutanova, K. (2019). BoolQ: Exploring the surprising difficulty of natural yes/no questions. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2924-2936. https://doi.org/10.48550/arXiv.1905.10044
44. Wissner-Gross, A. D. & Freer, C. E. (2013). Causal entropic forces. *Physical Review Letters* 110:168702.
45. Marcus, G. & Davis, E. (2013). A grand unified theory of everything. *The New Yorker*, May 6, 2013.
46. Minsky, M. (1985). *The Society of Mind*. New York: Simon & Schuster.
47. Tull, S. (2023). *The power of prompting: Mastering the art of prompt engineering with ChatGPT*. Amazon Kindle.
48. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems* 35, 22199–22213.
49. Lev, I., MacCartney, B., Manning, C. D. & Levy, R. (2004). Solving logic puzzles: From robust processing to precise semantics. *Proceedings of the 2nd Workshop on Text Meaning and Interpretation, Barcelona, Spain*, Association for Computational Linguistics, 9-16.
50. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *36th Conference on Neural Information Processing Systems (NeurIPS)*.
51. Li, Z., Peng, B., He, P., Galley, M., Gao, J., & Yan, X. (2023). Guiding large language models via directional stimulus prompting. *37th Conference on Neural Information Processing Systems (NeurIPS)*.
52. Rospigliosi, P.A. (2023). Artificial intelligence in teaching and learning: What questions should we ask of ChatGPT? *Interactive Learning Environments* 31(1), 1–3.
53. Grieser, G., Jantke, K. P. & Lange, S. (2002). Consistency queries in information extraction. In Cesa-Bianchi, N., Numao, M., & Reischuk, R. (eds.) *Proceedings of the 13th International Conference on Algorithmic Learning Theory, Lübeck, Germany*. Vol. 2533 of Springer LNAI, 173-187.\
54. Biten, A. F., Gomez, L. & Karatzas, D. (2022). Let there be a clock on the beach: Reducing object hallucination in image captioning. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. IEEE, Los Alamitos, CA. https://doi.org/10.48550/arXiv.2110.01705
55. Vasconcelos, M. A. R. & dos Santos, R. P. (2023). Enhancing STEM learning with ChatGPT and Bing Chat as objects to think with. EURASIA Journal of Mathematics, Science and Technology Education 19(7) em 2296, https://doi.org/10.29333/ejmste/13313.
56. Baidoo-Anu, D. & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. SSRN. https://doi.org/10.2139/ssrn.4337484.
57. Wilkinson, G. G. (2024). Enhancing generic skills development in higher education in the era of large language model artificial intelligence. *Journal of Higher Education Theory and Practice* 24(3), 64-76.
58. Arora, S., Narayan, A., Chen, M., Orr, L., Guha, N., Bhatia, K., Chami, I., & Ré, C. (2023). Aks me anything: A simple strategy for prompting language models. *The Eleventh International Conference of Learning Representations (ICLR 2023)*, Open Review. https://doi.org/10.48550/arXiv.2210.02441
59. Bailey, J. (2023). AI in education: The leap into a new era of machine intelligence carries risks and challenges, but also plenty of promise. *Education Next* 23(4), 29-36.https://www.aei.org/articles/ai-in-education
60. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *Cornell University arXiv*:2303.12712v1. https://doi.org/10.48550/arXiv.2303.12712
61. Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology* 32, 444-452. https://10.1007/s10956-023-10039-y.

62. Fisk, R. (2023). The rise of ChatGPT and generative A.I. and what it means for schools. *AASA Journal of Scholarship and Practice* 20(1), 37-46.
63. Floridi, L. (2014). *The 4th Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press.
64. Gregory, S. F. (2024). Empowering teaching with prompt engineering: How to integrate curriculum, standards, and assessment for a new age. In Sharma, R. C. & Bozkurt, A. (eds.), *Transforming Education with Generative AI: Prompt Engineering and Synthetic Content Creation*. Hershey: IGI Global, 239-260.
65. Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Cohen, V., Kukliansky, D., Scialom, T., Szpektor, I., Hassidim, A. & Matias, Y. (2022). TRUE: Re-evaluating Factual Consistency Evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*. Association for Computational Linguistics, 161-175.
66. Honovich, O., Choshen, L., Aharoni, R., Neeman, E., Szpektor, I. & Abend, O. (2021). Q^2: Evaluating Factual Consistency in Knowledge Grounded Dialogues via Question Generation and Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 7856-7870.
67. OpenAI (2023). *GPT-4 Technical Report.* Retrieved from https://cdn.openai.com/papers/gpt-4.pdf.
68. Rebuffel, C., Scialom, T., Soulier, L., Scoutheeten, G., Cancelliere, R. & Gallinari, P. (2022). Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery* 36, 318-354
69. Šedlbauer, J., Činčera, J., Slavík, M., & Hartlová, A. (2024). Students' reflections on their experience with ChatGPT. *Journal of Computer Assisted Learning*. Wiley online, https://doi.org/10.1111/jcal.12967.
70. Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). AUTOPROMPT: Eliciting knowledge from language models with automatically generated prompts. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, ACL, 4222-4235.
71. Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and students learning. *Journal of Applied Learning and Teaching* 6(1), 31-40.
72. Wang, A., Cho, K., & Lewis, M. (2020). Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, June 5-10, 2020*, ACL, 5008-5020.