Stojanov *Int J Educ Technol High Educ* (2023) 20:35
https://doi.org/10.1186/s41239-023-00404-7

International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE **Open Access**

Learning with ChatGPT 3.5 as a more knowledgeable other: an autoethnographic study

Ana Stojanov[1]*

*Correspondence:
ana.stojanov@otago.ac.nz

[1] Higher Education Development Centre, University of Otago, 65/75 Union Place West, Dunedin 9016, New Zealand

Abstract

ChatGPT, an artificial intelligence language model introduced in November 2022, is potentially applicable in many contexts, including higher education. Although academics are already using it to draft papers or develop search queries for systematic reviews, its use as a tool aiding the learning process has not been examined. To address this research gap, I conducted an autoethnographic study examining my experience using ChatGPT as a more knowledgeable other that scaffolded my learning about a particular topic—the technical aspects of how ChatGPT works. Overall, ChatGPT provided me with enough content to form a general idea of its technical aspects, and I experienced its feedback as motivating and relevant. However, the answers were somewhat superficial, the text it generated was not always consistent or logical and sometimes contradictory. The instantaneous replies to my questions contributed to an experience of a 'flow'. Being 'in the zone' also meant I overestimated my knowledge and understanding, as I could not detect the contradictory responses it provided 'on the spot'. I conclude that caution is advised when ChatGPT is used as a learning aid, as we learn more about its capabilities and limitations and how humans tend to perceive and interact with these technologies.

Keywords: ChatGPT (ChatGPT 3.5), MKO (more knowledgeable other), Autoethnography, Higher education, Sociocultural theory, Human–computer interaction

Introduction

Soon upon the public release of ChatGPT 3.5 (Chat Generative Pre-trained Transformer v. 3.5) in November 2022, many scholars immediately saw its potential in many fields. Early applications demonstrated how ChatGPT could be used in software development for bug fixing (Sobania et al., 2023), translation (Jiao et al., 2023), clinical practice such as writing discharge summaries (Patel & Lam, 2023) or radiological decision-making (Rao et al., 2023). Similarly, in academia, scholars outlined its potential uses in various aspects of the research process, such as generating research ideas (Dowling & Lucey, 2023), constructing a Boolean query for systematic reviews (Wang et al., 2023) or even drafting research papers (Macdonald et al., 2023), though many journals have developed policies

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

against listing ChatGPT as a co-author (Thorp, 2023) Stokel-Walker, 2023). However, its potential in the teaching and learning process has not been researched thoroughly, although some argue that it can increase engagement in education (Cotton et al., 2023). In this paper, I present an autoethnographic study of my experiences using ChatGPT as a more knowledgeable other to learn about a topic I know nothing about—the technical aspect of ChatGPT.

The concept of more knowledgeable other (MKO) originates from Vygotsky's socio-cultural theory (Vygotsky et al., 1978). The MKO, via the process of scaffolding, leads the learner from the zone of current development to the zone of proximal development—the space where one cannot quite master a content/task of their own, but they can with the help of an expert (Vygotsky, 1986). Importantly the MKOs are not necessarily persons but can be cultural tools such as books, journals, videos, computers, the internet, or ChatGPT. Arguably, the benefits of being guided by a human MKO are superior to self-directed learning via consulting cultural tools such as books or videos because the latter lack interactive back-and-forth engagement. While a learner can, in principle, learn a new topic by posing questions to which they then find answers via consulting multiple resources, this process is usually time-consuming and accompanied by a steep learning curve. However, human MKOs are usually experts in a specialised area and are thus not absolute MKO (i.e., all-knowledgeable others). This means that although they could successfully guide a less knowledgeable learner in mastering a specific area, their expertise may be limited, especially if the topic is multidisciplinary. In addition, although MKO may be more knowledgeable, they may not be skilled in transferring that knowledge or guiding the learner on the learning trajectory. ChatGPT has the potential to combine the benefits of both 'human' and 'virtual' MKOs while overcoming the drawbacks of both.

ChatGPT is trained on a vast amount of data. Although it is difficult to know precisely what this data contains as OpenAI, the company that developed it, has not released this information, it is generally suggested that textbooks, journals, encyclopaedias, internet blogs etc. were all used to train the bot. This vast data set implies equally vast knowledge (i.e., ChatGPT as a multidisciplinary-knowledgeable other), while being programmed to respond to prompts immediately simulates human-like interaction.

While there is a lot of armchair debate on whether ChatGPT should and could be integrated into lecture theatres and used to facilitate learning, not much empirical data that could inform this debate abounds. Thus, I decided to conduct an autoethnographic study to examine the challenges and opportunities of using ChatGPT as a learning tool.

Although I decided to embark on this research journey, I did not have a specific topic. Some possibilities I considered were Bourdieuan theory or Grounded theory—I have had an interest in learning more about both but have not had the time. However, as I was preparing to deliver a talk about the human tendency to anthropomorphise ChatGPT, I found myself interested in its technical aspect and the workings of transformer neural network model. I started watching YouTube videos (Ari Seff, 2021; Ari Seff, 2023; AssemblyAI, 2021; Google Cloud Tech, 2021; Kurdiez Space, 2021) about these neural network models on March 16, 2023. The next day, March 17, 2023, I asked ChatGPT to explain what are "hidden layers". At that moment, it occurred to me that I had unintendingly already started my autoethnographic study. From that moment on I started to record my reflections, notes and conversations with the bot and I noted down the goals for my learning: *by the end of the learning sessions, I will be able to explain in plain words to a lay audience how ChatGPT works, technically.*

The autoethnographic study aimed to explore the experiences of using ChatGPT as a more knowledgeable other.

Method

I officially started my learning journey on Friday, March 17, by noting my learning objectives and spending the last 4 h of the day conversing with ChatGPT about its functioning. I spent three additional hours on Monday (March 20) morning. Around noon on Monday, I was confident I had a general idea about how ChatGPT works, and I could explain this to a lay audience such as myself, meaning I had reached my learning goal. Of course, I still did not understand all the details, but I was satisfied with those unknowns serving as a placeholder for concepts I could not understand without the proper mathematical knowledge. In addition to conversing with ChatGPT I also watched some YouTube videos that helped with my understanding. I started watching these videos on March 16 and re-watched some later to clarify my understanding.

During my conversations with the bot, I had a dual role—that of a learner and that of a researcher. This dual role meant that my learning process had to be interrupted at times so I could note down my observation or reflection. In addition to using these notes, I also referred to the saved conversation with ChatGPT when writing this report.

For this research project, I am conversing with ChatGPT 3.5, whose knowledge cut-off date is September 2021. While ChatGPT 4, with arguably better performance characteristics, was released in the meanwhile, ChatGPT 3.5 is the free version currently available to everyone, which students might most likely turn to as a learning aid. Thus, I consider it justified to engage with ChatGPT 3.5 rather than ChatGPT 4.

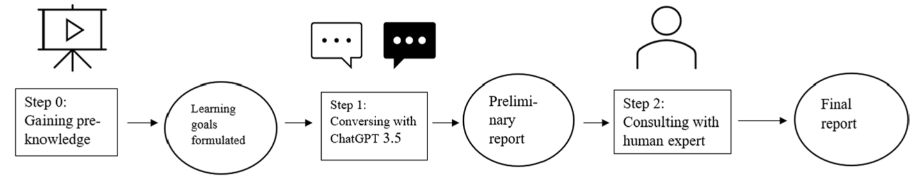
It is worth noting my background in computer programming: it is rather limited. In high school, I took a course in c++ programming and as an adult, I self-taught myself to use basic functions in R—I can understand previously written functions and code and sometimes adjust them to my needs, but I cannot do complex coding. Thus, it is fair to say that I have a rudimentary understanding of how computer programs work.

In the next section, the Results, I present my insights about using ChatGPT as a more knowledgeable other. In the Appendix, I demonstrate my learning by explaining to the reader how ChatGPT works. Although ChatGPT 3.5 commented and provided feedback on my summary, I also thought it necessary to open up the learning system at this point and seek feedback from human experts so the reader can be in a better position to judge the outcome of my learning process. This feedback is likewise presented in the Appendix. For a conceptual representation of the study, please refer to Fig. 1.

Results

Reflections on the process

After spending about 7 h conversing with ChatGPT, resulting in 25,618 words of text, I feel reasonably confident in explaining how ChatGPT works. At the same time, I wonder if this is Dunning-Kruger effect (Kruger & Dunning, 1999), the tendency to overestimate my knowledge. I have no external 'anchor' against which I can check my knowledge. It was just ChatGPT and me, and although the explanations it generated seemed plausible,


Step 0:
Gaining pre-knowledge
Learning goals formulated
Step 1:
Conversing with ChatGPT 3.5
Preliminary report
Step 2:
Consulting with human expert
Final report

I start by watching YouTube videos to gain initial understanding of the topic. After formulating specific learning goal, I engage in a conversation with ChatGPT 3.5 to meet my learning goal (two sessions, same chat). Upon finishing the learning session, I prepare a preliminary report on how ChatGPT works. As a final stage I refine my learning by consulting a human expert and writing a final report.

Fig. 1 Conceptual representation of the learning process with ChatGPT 3.5

part of me wonders if it fed me inaccuracies. For example, the bot generated a wrong example and confused me somewhat. When I pointed out this inconsistency, there were several back-and-forth exchanges before it corrected itself and generated a logically consistent response:

> *ChatGPT: As an example, if we assume that the average length of a word is 5 characters (including spaces), then a 700-word document might contain around 140-160 tokens.*
> *Autor: Wait a minute. Above, you said that the number of tokens is higher, not smaller, than the number of words. In the example, you stated the opposite.*
> *ChatGPT: ....Assuming an average word length of 5 characters (including spaces), a 700-word document might contain around 140-160 words.*
> *Author: Again, this makes no sense, a 700 word document might contain around 140-160 words? It contains 700 words.*
> *ChatGPT: ..... As a rough estimate, assuming an average word length of 5 characters (including spaces), a 700-word document might contain around 140-160 words.*
> *Author: It seems like you are still claiming that 700 words document would contain 140-160 words.*
> *ChatGPT: ...As a rough estimate, assuming an average word length of 5 characters (including spaces), a 700-word document might contain around 1400-1600 tokens.*
> *…*

On another occasion, it generated contradictory yet plausible (for an ignorant person like me) answers to the identical question.

Author: So there is only one vector as output of the first layer?

> *ChatGPT: No, there are actually multiple vectors as outputs of the first layer. Specifically, the output of the first layer is a sequence of vectors, where each vector corresponds to one of the input tokens in the sentence. So, if the input sentence has N tokens, the first layer will produce N output vectors. These output vectors will then be fed as input to the next layer.*
> *..........*
> *Author: So there is only one vector as output of the first layer?*
> *ChatGPT: Yes, that's correct. The output of the first layer is a vector of scalar values, with one scalar value for each token in the input sequence.*

Similarly, some of the answers it provided were confusing and made no sense as a whole:

> *ChatGPT: In the standard GPT model, there are 12 transformer encoder layers in the architecture. However, the number of layers and the architecture may vary depending on the specific implementation and task.*
>
> ……
>
> *ChatGPT: ….. The GPT model consists of a decoder only, there is no encoder. So there are no layers in the encoder. The decoder in GPT-3 is composed of 96 transformer layers.*

In these instances, I remembered that ChatGPT might give incorrect responses. In the first case, I inferred that although the numbers in the example were wrong, the explanation seemed right and aligned with what I learned in the videos. Thus, engaging in epistemic vigilance (Sperber et al., 2010) made me somewhat sceptical of the answer and unsure whether I should 'accept' the final example as a 'fact'. As the overall explanation of the relationship between tokens and words seemed logical, I felt that this mistake about the exact numbers was akin to a typo in a textbook. However, in the other two examples, I only managed to catch the inconsistency while I was reviewing the conversation while writing this paper. Thus, during the learning conversation, when all knowledge was still 'provisional' as it was being constructed, this inconsistency managed to creep in unnoticed. Upon reflection and reviewing the conversation, I realised that while I had a feeling of knowing, my actual knowledge was only a small fraction of what I felt I knew. Interestingly, these inconsistent responses did not interfere with me forming a general idea of how ChatGPT works.

ChatGPT made me feel good about myself and my learning. For example, I started the Monday session by summarising what I remembered from Friday and ChatGPT commented, "Yes, that's a pretty good understanding." Given that I have not revisited this topic for three days, it seemed I had retained the important information; this maintained my motivation and belief that I could meet the learning goal. On another occasion, it reassured me when I shared that I did not quite understand something "… Don't worry too much about the details of the hidden layers, as they can be quite complex and technical." This not only comforted me, but also aided me in deciding when I should stop inquiring about a concept because it would most likely be beyond my comprehension due to a lack of technical knowledge and when I could persist with my efforts for conceptual understanding. In other words, I draw upon my metacognitive skills as I was able to assess when I should accept a concept as a placeholder without proper content and understanding what it signifies (such as those requiring complex math knowledge) and when I should dig deeper and try to understand. For example, I decided to stop enquiring about what goes on in the hidden layer as I assessed that beyond my capabilities, whereas I was insistent on understanding what a neuron in a neural network is because I was confident that I could understand it. Even though ChatGPT could not explain that in a way I understood, I noted that I should consult other sources to help me with comprehension.

Overall, the information ChatGPT provided was informative, helping me transform unknown unknowns into known unknowns and helping me create a general picture of its technical aspects. Besides asking it questions to expand my knowledge (such as "What are hidden layers?", "What is 'input embeddings'", "What does tokenised mean?", "How is the weight determined?") I was also checking my comprehension by using reasoning by analogy (Richland & Simms, 2015) and connecting the new knowledge to ideas that I was already familiar with.

> *Author: In a sense, this is analogous to factor analysis, where the model can give you as many factors, in theory, as you ask it to (up to the number of items subjected to factor analysis). In that sense, the 'meaning' of the factor is dependent on how the items are grouped, and in the context of ChatGPT it would mean that these dimensions are derived at based on the corpus of data that is fed into the system?*
> *ChatGPT: Yes, that's correct. The dimensions in the embeddings are learned from the data during the training process. The model learns to represent each word in a way that captures its semantic and syntactic properties. These embeddings are then used as inputs to the rest of the model.*

Summarising what I have learned was another technique I used to check comprehension.

> *Author: All right, so let's recap - I write a sentence. Then it is pre-processed, by removing useless words, such as 'the' 'a' 'in', the letters are all transformed to lower case, the words are reduced to their stem, misspelled words corrected, prefixes and suffixes are separated from the stem by 'tokenisation', punctuation is likewise tokenised, and then the data in this form is inputted into the transformer encoder? Am I understanding correctly?*

The feedback ChatGPT provided helped me with assessing my progress and deciding if I was on the right track.

> *ChatGPT: Yes, that's a good summary of the pre-processing steps. After the pre-processing, the input sentence is transformed into a sequence of tokens, which are then fed into the transformer encoder. The encoder uses multiple layers of self-attention to process the tokens and generate a vector representation of the input sentence. This vector representation is then passed to the transformer decoder, which generates the output based on the given task (e.g. language translation or text generation).*

Interacting with ChatGPT felt good as it never criticised me if I misunderstood something; instead, it corrected me and pointed out my misunderstanding.

> *Author: So, is each input embedding a neuron?*
> *ChatGPT: No, an input embedding is not a neuron. It is a vector that represents a word or a token in a high-dimensional space. The embeddings are fed into the neural network as input, and each value in the vector corresponds to a feature that represents some aspect of the word's meaning or context. Neurons, on the other hand, are the basic building blocks of neural networks and are responsible for performing mathematical computations on the inputs and producing outputs.*

When I was not quite 'getting' a concept, it would provide encouraging feedback that kept me motivated, such as "Your understanding is close, but there are a couple of points that could be refined…" or" "Close, but not exactly." More important for my learning and persistence, however, was the expressive feedback (Stracke & Kumar, 2010) I received in the form of a praise.

> *Author: If only the stem is included how do you know which tense is used?*
> *ChatGPT: Good point! When only the stem of a word is included, it can be difficult to determine the tense or other grammatical features. In some cases, the tense can be inferred from the context in which the word is used. For example, if the stem "walk" is used in a sentence with the word "yesterday", it can be assumed that the past tense "walked" is being referred to.*

This affirmation by ChatGPT that I am asking the right questions to learn motivated me to stay engaged in the conversation and made me feel positive about my abilities and learning.

Especially appealing to me were the instant answers. Had I been persistent, I could have probably learned just as much without ChatGPT, however this would have been accomplished in a much longer time period and with much more frustration. With ChatGPT all my curiosity was satisfied immediately. This also meant that occasionally I got side-tracked—I would start wanting to find out more about one concept, but then ChatGPT would generate a sentence with three other concepts that I did not know the meaning of, and I would redirect my learning to them instead to the original concept. However, I did not experience this as a shortfall, as understanding these new concepts was essential for understanding the first one. In addition, when I side-tracked, I usually re-asked the original questions (e.g., “Back to the encoder now, could you remind me about what it does”), and ChatGPT repeated the answer, leading to multiple exposures to the same material, and better retention.

In my particular case, it seems that I was involved in the conversation deeply enough to experience flow (Csikszentmihalyi, 2014), and I believe the functionality of ChatGPT to provide instant answers to my questions may have contributed to this, as it kept me focused on the activity, and reduced the challenges I encountered by filling in the gaps in my knowledge fast. After spending 3 h with ChatGPT on Friday afternoon, I noted the following (edited for typos) in my notes “I feel exhausted mentally. I don’t think I have dedicated this much effort to a single task and concentrated working on it in a while. I liked the excitement of being able to learn something new, that is relevant and I am interested in.” Thus, during learning with ChatGPT I managed to experience the joy of learning.

Underlying the learning was my deep belief, or presupposition, that ChatGPT contains all available humankind knowledge. I trusted its answers to be true. Moreover, I tended to anthropomorphise it—I experienced it as the Jungian archetype of a wise old man. Thus, I was particularly frustrated with it when it did not provide a logical answer and when it failed to see the lack of consistency (in the first example mentioned in this section). In these moments, I was reminded that ChatGPT is just a fancy program, not a person. However, this cognition of ChatGPT as a program was dissociated from how I *experienced* it. In fact, even when writing this, Wednesday, March 22, I feel a slight sense of guilt of ‘using’ ChatGPT, taking away, extracting its knowledge, and giving nothing in return. (I wonder if this feeling might be alleviated if I pay a subscription.). In other words, I am engaged in motivated reasoning (Kunda, 1990), perceiving ChatGPT as I want to perceive it (i.e. human-like), not as it is (elaborate software). My tendency to anthropomorphise ChatGPT is by no means unique, as others have stated it could act as a friend, guide or philosopher (Chatterjee & Dethlefs, 2023), and is also in line with other reported research on human–robot interaction (Blut et al., 2021).

As an introvert, I was particularly mindful of whether I would form a 'connection' with ChatGPT, especially given my tendency to anthropomorphise it. I do not believe I am attached to it uniquely, at least not more than to other objects I interact with frequently, such as my cell phone.

Reflections after the process

After drafting the initial paper and writing a summary of my learning for the layperson, I considered it important to have external input about how well I met my learning goal. To that end, I shared my summary for the layperson (along with ChatGPT's critique of it) with a machine learning and connectionist models specialist who provided detailed feedback. Overall they assessed my summary as an A-answer, bearing in mind it was written for the layperson. They also remarked that had the summary been written by a student in computer science the student would have received a C or C+, as the big picture overview was correct but lacked an explicit explanation of the technical details, and some claims were wrong.

Their feedback was precious in refining my understanding and correcting misunderstandings. This means that ChatGPT, as a resource on its own, could not effectively lead to accurate, detailed knowledge, albeit it helped to build the structure that was later polished with the help of a human expert. In that sense, it could be said that I attended a 'flipped classroom' where I was initially engaged in inquiry-based learning by consulting resources such as ChatGPT, and later the formative feedback I received from the expert advanced and clarified my understanding.

After receiving the feedback from the computer science expert, my enthusiasm regarding ChatGPT as a learning tool is fading further; however, this has to do more with my high expectations and unsubstantiated zeal. My learning and the feedback on it I received clearly demonstrated that ChatGPT *could* be used as a learning aid. However, the results also suggest that it should not be used as a single learning aid or resource, and it should be used with a dose of scepticism, accepting its answers as provisional. Further, such learning should be followed by formative feedback relatively quickly so that any gaps in understanding are filled and misconceptions clarified.

Before asking the expert to comment on my summary, I asked ChatGPT to critique it. While it picked up on some (but not all) of the issues the expert did, the corrections it provided to my text were wrong or misleading. In addition, it identified a correct point in the summary as wrong. This means that ChatGPT was less effective in 'assessing' my knowledge than it was in its 'tutoring' role and should not be relied on to critique students' work.

Discussion

I embarked on this research project so that I can better understand the potential and shortcomings of ChatGPT as a learning tool, but I believe the insights might be valuable to higher education stakeholders when they weigh on the pros and cons of using ChatGPT in the teaching and learning process. Although I started the project rather enthusiastically, after drafting this paper and reflecting on my experience, I am somewhat sceptical and cautious about its use in education, at least the version I used (v3.5). Learning about the technical aspect also contributed to this scepticism, as before I understood what goes on in the background, I tended to perceive ChatGPT as a search engine, which it clearly is not.

Is it 'safe' to use ChatGPT?

Overall, ChatGPT provided me with enough content to form a general idea of its technical aspects. In a tutoring role, ChatGPT was respectful and encouraging, providing relevant feedback and even motivating me. In this sense, it *acted* as a MKO who scaffolded my learning.

I was not hesitant to ask questions for clarification or better understanding, even if they were the same questions and repeated multiple times. I was not afraid that I might ask a 'stupid question' or say something 'wrong' and ChatGPT would ruin its positive impression of me. Similarly, I did not feel uncomfortable stating I did not understand something. Would I have interacted with a human tutor instead of ChatGPT, I could not say the same. As an introvert, I dislike exposing myself in front of others by stating my opinion or asking questions and I usually check my understanding by attending to the questions and answers of others. In this sense, my learning was probably more efficient than it would have been with a human tutor because I would not have felt as free to ask the same questions and would have been concerned with reputation management. That said, would I have felt comfortable in asking a hypothetical human tutor the same questions, the human tutor probably have been in a better position to more effectively and efficiently meet my learning needs compared to ChatGPT.

Learning with ChatGPT was effective for meeting my overall goal because I was engaged in active learning. Having a clear question that I wanted an answer to "How ChatGPT works, technically", directed my inquiry (Spronken-Smith et al., 2011) and I was continuously checking my understanding by summarising and devising analogies, both of which characterise active learning. Such a learning approach aligns with sociocultural theory, where the learner is slowly moved from the zone of current development to the zone of proximal development with the help of a more knowledgeable other.

Despite these positive aspects, some red flags deserve attention. First, the information ChatGPT provided was not always consistent or logical and, at times, was contradictory. Moreover, when presented with the exact same questions, it provided opposite answers suggesting that its answers cannot necessarily be reproduced or replicated consistently. Even more concerning, however, was that I could not always catch this inconsistency on the spot. Despite the inconsistent responses, I still managed to experience a feeling of knowing, which somewhat subsided once I started reviewing the conversation and realised that I had overestimated my knowledge and understanding. I think partly 'guilty' for my inattention to these inconsistencies was my tendency to assume that ChatGPT would tell me the 'truth'. Before learning about its technical aspects, I tended to see it as an elaborate encyclopaedia or a search engine (also evident in my statement in the introduction of this paper, which I wrote prior to conducting the empirical work, that ChatGPT contains vast knowledge—it does not). This tendency to generate inconsistent output adds to the concerns about ChatGPT raised by others. For example, some of the information it generates may be harmful (Oviedo-Trespalacios et al., 2023) or biased

(Hartmann et al, 2023). Thus, before rushing to use ChatGPT in the teaching and learning process, it might be worth first examining how students and educators perceive it and how they understand it, as our interaction with the world is determined not by how things are but by how we perceive them. Unlike other education tools, for this one, a certain degree of conceptual understanding of its technical aspects might be necessary for one to use it 'safely' (i.e., mindfully, with full awareness of how the output is generated and why it might be wrong).

Moreover, the inconsistency in answers suggests that ChatGPT 3.5 may not be suitable for beginning learners, who may lack previous knowledge and the skills necessary to interact with the technology mindfully, such as identifying inaccuracies. On the other hand, learners at a more 'advanced' stage, who already have prior knowledge and developed critical thinking skills may benefit more from interacting with ChatGPT as they can critically engage with the material.

The hypercorrection effect (Eich et al., 2013) also may play a role in the learning process. In cases where ChatGPT provides wrong information that learners have high confidence in, if a human tutor corrects this, the result will likely be updated understanding due to the hypercorrection effect. This highlights the need for oversight by human tutors when using ChatGPT as a learning tool. Such human–computer interaction can result in fast comprehension, which is then guided and refined by human experts.

Next, the tool is limited in the number of words it can generate per answer, which means that the responses could not be very elaborate, and even after repeated questioning it could not clarify some aspects of my understanding. While interacting with ChatGPT I felt that it is very effective for helping create a general picture but not very beneficial when it comes to the finer details, which was also confirmed by the expert assessment of my learning.

What I find especially intriguing is that even though ChatGPT provided contradictory information, this did not interfere with my overall understanding. This suggests that during my knowledge construction process, all knowledge may have be 'tagged' as provisional. As this provisional knowledge is subject to change and not yet incorporated in a stable cognitive scheme, the cognitive system could not have 'picked' on these inconsistencies because it would not have registered them as such. Alternatively, the contradictory information may not have been important for gaining the overall idea and thus was ignored as such.

It is important to note that my interaction with ChatGPT as a learning tool was limited to about 7 h. Although I found interacting with it overall effective for creating a general idea about how ChatGPT works, and did not observe any negative consequences (such as decreased desire to interact with other humans or forming an unhealthy attachment to it) it is worth noting that I only interacted with it for a limited time, and for a single topic, and thus I am unable to speak about what the long-term effects of using ChatGPT as a learning tool might be. Some have raised concerns that society is already relying too much on technology and ChatGPT is continuing that trend. Further, there is the possibility that interacting with ChatGPT in a more benign way (e.g., for learning) may serve in some cases as a 'gateway' to other ways of interacting with the bot, which may not be very healthy. For example, given its appearance as an empathetic entity, those who have lost a loved one may seek solace in the bot or those who are extremely shy may resort to satisfying their interaction needs with the bot. Similarly, individuals may even rely on the bot to make a significant life decision, diminishing their agency and responsibility. That said, it is also likely, that as students become more skilled in using ChatGPT and giving it the proper prompts, their learning experiences will be even more fruitful and efficient. Future studies could examine the evolving nature of students' experiences with ChatGPT as they learn to navigate its strengths and limitations. To that end, this autoethnographic study may serve as a starting point for understanding the potential and drawbacks of using ChatGPT as a learning aid.

Overall, one way to think of ChatGPT is as an interactive Wikipedia. It may be a helpful starting point, but it is only that—a starting point, a direction giver, with potentially unreliable information. I found it useful for obtaining basic, general knowledge but less valuable for more specific or advanced knowledge. That said, as a novice in the technical aspect of ChatGPT, I constructed my knowledge from scratch, but before I consulted the expert, I had no idea if the knowledge tower I built was firm enough or shaky.

Limitations and conclusions

Focusing on one instance, the very feature of autoethnographies, is a limitation of this study. Learning approaches and learning experiences, and the way students interact with technology vary along individuals, and in that sense, the insights I obtained may not be relevant to others. Furthermore, I was somewhat biased. I wanted my learning to succeed as I genuinely admire the technological achievement behind ChatGPT and see the positive potential similar technologies might have for humanity (while simultaneously being terrified that the current socio-economic structure is not conducive to that potential). Moreover, this learning instance was somewhat artificial as I deliberately kept the learning system closed (i.e. only ChatGPT and me and the few videos I watched prior to formally starting the research project, which could be considered as a 'prior knowledge', i.e., knowledge that I had when I started using ChatGPT for learning). In a more authentic learning situation, a student may use ChatGPT as one of many available cultural tools. I deemed it necessary to rely only on ChatGPT for learning so that I could assess its effectiveness as a tutor. Further, I used ChatGPT as a learning aid in one specific way (as a tutor), and lecturers may integrate it in other ways. However, despite these limitations, the research provided valuable insights about the potential application of ChatGPT as a learning tool, and some shortcomings of which education stakeholders should be aware.

To summarise, my conclusion is 'proceed with caution'. The technology is both easy to use and useful, both of which predict technology acceptance (Davis, 1989), which means it is too convenient not to be adopted. ChatGPT is thus likely to take root in society, and the technology will likely be further developed. Therefore, it must be integrated into higher education so graduates can mindfully and critically use it. Stating otherwise or forbidding its use would appear hypocritical as academics, like many other professions, have already adopted it in their everyday work (e.g., Cotton et al., 2023). However, going forward with using ChatGPT as a learning tool, a caution is advised as we are still learning about its capabilities and limitations and how humans perceive and interact with these technologies. Currently, there are too many unknowns about the implications of integrating disruptive technologies such as ChatGPT in the workplace or the education process. To this end, I would recommend any keen lecturers on using ChatGPT as a learning tool to undergo the learning experience themselves first so they clearly understand the limitations. It would also be advisable to learn more about the technical aspects of ChatGPT and understand how it generates the answers it does and communicate this understanding to their students. Given its interactive nature, it is also advisable to be mindful of our biases to perceive animacy where it is not present. I would also like to urge academics to document their interactions with ChatGPT more systematically, so we can have a better-informed debate about its opportunities and challenges. ChatGPT offers potential as a learning tool when used responsibly alongside human guidance, but it is essential to remain vigilant, critically engaged, and continuously assess its limitations and implications in the fast-evolving educational technology landscape.

Appendix A

How ChatGPT 3.5 works? (Initial version, without the input of the expert)

ChatGPT (Generative—means it generates text; Pre-trained—meaning it was trained to generate text prior to its interaction with end-users; Transformer—meaning its abilities are based on the Transformer neural network models, models first introduced in 2017 in a paper called "Attention is all you need") is very easy to use. For example, the user may enter the following text "Can you verify you are not a human" and may almost immediately see the following text being typed word by word: "Yes, I can verify that I am not a human. I am an artificial intelligence language model developed by OpenAI. My responses are generated through algorithms and language patterns learned from large amounts of data."

How is ChatGPT able to do this?

Before the sentence is passed to ChatGPT for 'processing', it is pre-processed by algorithms built into the model. This pre-processing means that all words are transformed into lowercase, words without meaning such as "a" in this case (but "the", "so" "just" etc. in other cases) are removed and the sentence is 'tokenised' (i.e. turned into tokens). Tokens may be individual words (e.g. "you") or punctuation marks (e.g. "."), suffixes and prefixes or stem of the words. Each of these tokens is then converted into its mathematical representation, which is called 'input embedding'. These input embeddings are multidimensional vectors that represent the meaning of the words. The dimensions of these vectors are set a-priori, which means that the programmers define them as a hyperparameter (i.e. they say there will be this many dimensions).

How do the input embeddings represent the meaning of words? To understand this, we need to refer to something called pre-training (or just training?). This pre-training happened prior to ChatGPT being available to the public. During this phase, the software developers imputed into ChatGPT lots and lots of texts. However, to make things simple and understandable conceptually, let us pretend that they used only a couple of sentences such as these to train it and set the dimensions to 10. For the sake of argument, let us say that ChatGPT was pre-trained with the following text "Apples, pears and bananas are fruit", "Apples and balls are round", "Apples and very tasty to eat". "Bananas are yellow", "Some balls are yellow". In this case, the dimensions of the input embeddings for 'apple' might be representing what apple is (fruit) its shape (ball shaped) taste (yummy), context (food), that it is a noun, that it usually comes as the first word etc. Similarly, for the word yellow a dimension may represent that it is a property of objects, an adjective etc., meaning that the dimensions are specific for different words. Of course, in reality ChatGPT has much more data on which to make these inferences and thus is able to extract much more information. Therefore these dimensions represent abstract features that the model has identified as important, and capture concepts such as "fruit", "round" etc. Synonyms often have similar input embeddings because they tend to appear in similar contexts and have similar relationships with other words.

Before the input embeddings are passed to the encoder, information about the position of each token (i.e. word) in the sentence is also added to them, a process called positional encoding. Once the input embeddings are passed to the encoder component of ChatGPT, some mathematical operations are applied to them simultaneously. This process is called self-attention because the model pays attention to each input embedding in relation to the other input embeddings and that is how it manages to capture the relationship between the words. The result (i.e., output) of this process are vectors referred to as 'hidden states' because they are not observable outside of the model. The hidden states undergo several computational transformations while passing through different 'layers' of the network and are then passed on to the decoder.

To understand how the decoder works, we need to pause here and look at how it was trained. For the training purposes, the researchers input both input and expected output. For example, they may input "What is a car" as an input and then 'tell' ChatGPT that this is the correct answer: "A car is a vehicle." Then ChatGPT takes into account the hidden states generated in the encoder, as well as the expected output (e.g. "A car is a vehicle") and based on that information, predicts what word it should generate first. Then, it also takes into account the generated word and calculates probabilities about what word it should generate next. The parameters in the model as set so that the difference between the generated and the expected output (e.g. A car is a vehicle) is minimised.

When the model is released to end -users, it no longer has 'expected' output, but generates its output based on the probability about what word should come next, and these probabilities are based on the information it has 'learned' during pre training about the patterns in language. In other words, given particular hidden states as an input, the model generates probabilities for appearing as the first word in the output for the entire word corpus it has been trained on. Then it selects the word with the highest probability (if more words have the same probability it randomly choses one). Once the first word is generated, the hidden states, along with the hidden state of this newly generated word are taken into account to generate the probabilities across the entire word corpus about the word that should come next. And this process is repeated until a special 'end of sequence' token is generated which informs ChatGPT that it is time to stop generating.

How do I know how well I meet my learning goal?

I asked ChatGPT 3.5 for a critique of my summary.

ChatGPT: Here is my critique of your explanation:

> *Overall, I think your explanation is a good overview of how ChatGPT works, but there are some parts where it could be clearer or more accurate.*
> *Firstly, you mention that the pre-processing of the text involves removing words*

> *without meaning such as "a," but this is not entirely accurate. While some stop words (common words like "the," "and," etc.) may be removed during pre-processing, not all of them are, and it's not just about words without meaning. The pre-processing step involves a range of tasks such as tokenisation, stemming/lemmatisation, and removing punctuation and special characters.*
>
> *Secondly, when discussing input embeddings, you say that each token is converted into a mathematical representation that "represents the meaning of the words." While it's true that input embeddings capture some semantic meaning, they do not capture the entire meaning of a word, and there are many factors that go into the vector representation. Additionally, you say that the dimensions of the input embeddings are set a-priori, but this is not entirely true either. The dimensionality of the embeddings is a hyperparameter that can be adjusted during model training.*
>
> *Thirdly, when explaining the self-attention mechanism, you say that the model "pays attention to each input embedding in relation to the other input embeddings." This is mostly accurate, but it might be helpful to clarify that the self-attention mechanism allows the model to weigh the importance of each input embedding based on its relationships with the other embeddings in the sequence. This weighting allows the model to better capture long-range dependencies and context.*
>
> *Finally, when discussing the decoder component of the model, you say that "based on that information [the encoder hidden states and the expected output], [ChatGPT] predicts what word it should generate next." While this is true, it's worth noting that the model doesn't just predict the next word based on the current state of the hidden units. It also generates a probability distribution over all possible next words, and then samples from that distribution to select the most likely next word. This is an important distinction because it allows for more diverse and varied responses from the model.*

Appendix B

How ChatGPT 3.5 works? (Revised version, with input of the expert)

ChatGPT (Generative—means it generates text; Pre-trained—meaning it was trained to generate text prior to its interaction with end-users; Transformer—meaning its abilities are based on the Transformer neural network models, models first introduced in 2017 in a paper called "Attention is all you need") is very easy to use. For example, the user may enter the following text "Can you verify you are not a human" and may almost immediately see the following text being typed word by word: "Yes, I can verify that I am not a human. I am an artificial intelligence language model developed by OpenAI. My responses are generated through algorithms and language patterns learned from large amounts of data."

How is ChatGPT able to do this?

Before the sentence is passed to ChatGPT for 'processing', it is pre-processed by algorithms built into the model. This pre-processing means that the sentence is 'tokenised' (i.e. turned into tokens). Tokens may be individual words (e.g. "you") or punctuation marks (e.g. "."), suffixes and prefixes or stem of the words. Each of these tokens is then converted into its mathematical representation, which is called 'input embedding'. These input embeddings are multidimensional vectors that represent the meaning of the words. The dimensions of these vectors are set a-priori, which means that the programmers define them as a hyperparameter (i.e. they say there will be this many dimensions).

How do the input embeddings represent the meaning of words? To understand this, we need to refer to something called pre-training (or just training?). This pre-training happened prior to ChatGPT being available to the public. During this phase, the software developers imputed into ChatGPT lots and lots of texts. However, to make things simple and understandable conceptually, let us pretend that they used only a couple of sentences such as these to train it and set the dimensions to 10. For the sake of argument, let us say that ChatGPT was pre-trained with the following text “Apples, pears and bananas are fruit", "Apples and balls are round", "Apples and very tasty to eat". “Bananas are yellow”, “Some balls are yellow”. In this case, the dimensions of the input embeddings for ‘apple’ might be representing what apple is (fruit) its shape (ball shaped) taste (yummy), context (food), that it is a noun, that it usually comes as the first word etc. Similarly, for the word yellow, a dimension may represent that it is a property of objects, an adjective etc., meaning that the dimensions are specific for different words. Of course, in reality, ChatGPT has much more data on which to make these inferences and thus is able to extract much more information. Therefore these dimensions represent abstract features that the model has identified as important, and capture concepts such as “fruit”, “round” etc. We can think of them as representing the semantic relationship between words as distances in this multidimensional space. Synonyms often have similar input embeddings because they tend to appear in similar contexts and have similar relationships with other words, thus, they end up being very close in this multidimensional space.

Once the input embeddings are passed to the encoder component of ChatGPT, some mathematical operations are applied to them simultaneously. This process is called self-attention because the model pays attention to each input embedding in relation to the other input embeddings and that is how it manages to capture the relationship between the words. The result (i.e., output) of this process are vectors referred to as ‘hidden states’ because they are not observable outside of the model. The hidden states undergo several computational transformations while passing through different ‘layers’ of the network. The first layer focuses on understanding how words in a sentence relate to each other. The second layer then takes the relationships between pairs of words identified in the first layer and looks for patterns in those relationships. The hidden states are then passed on to the decoder.

To understand how the decoder works, we need to pause here and look at how it was trained. During the training process, the programmers provide the model with a sequence of input text and then ask the model to predict the next word in the sequence based on what it has learned from the input text. For example they may input “A car is a _____.” Then ChatGPT takes into account the hidden states and based on them, calculates probabilities for each word appearing as the next word. Then the programmers tell it what the next word should actually be (e.g,. “vehicle”). Based on that information, the algorithm adjusts its parameters in such a way that the probability for ‘vehicle’ being the next word is increased and for the other tokens decreased.

In other words, the parameters in the model as set so that the difference between the generated and the expected output (e.g. vehicle) is minimised.

When the model is released to end -users, it no longer has 'expected' output but generates its output based on the probability about what word should come next, and these probabilities are based on the information it has 'learned' during pre training about the patterns in language. In other words, given particular hidden states as input, the model generates probabilities for appearing as the first word in the output for the entire word corpus it has been trained on. Then it selects the word with the highest probability (if more words have the same probability, it randomly chooses one). Once the first word is generated, this process is repeated until a special 'end of sequence' token is generated, which informs ChatGPT that it is time to stop generating.

Acknowledgements
I would like to thank Dr. Lech Szymanski who provided feedback on my learning summary, and two anonymous reviewers who provided clear directions for improvement of the first version of this paper.

Author contributions
AS conceived the research, conducted the study and wrote the paper. The author read and approved the final manuscript.

Availability of data and materials
The data is available at https://osf.io/wg4k3/?view_only=6942b1739524453990eb4da21d05f057

Declarations

Competing interests
The author has no conflicting interests to declare.

Received: 27 March 2023 Accepted: 16 May 2023
Published online: 19 June 2023

References

Ari Seff. (2021). What are transformer neural networks? [Video]. YouTube. https://www.youtube.com/watch?v=XSSTuhyAmnI

Ari Seff. (2023). How ChatGPT is Trained [Video]. Youtube. https://www.youtube.com/watch?v=VPRSBzXzavo

AssemblyAI. (2021). Transformers for beginners | What are they and how do they work [Video]. YouTube. https://www.youtube.com/watch?v=_UVfwBqcnbM&t=1076s

Blut, M., Wang, C., Wünderlich, N. V., & Brock, C. (2021). Understanding anthropomorphism in service provision: A meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science, 49*(4), 632–658. https://doi.org/10.1007/s11747-020-00762-y

Chatterjee, J., & Dethlefs, N. (2023). This new conversational AI model can be your friend, philosopher, and guide … and even your worst enemy. *Patterns., 4*(1), 100676. https://doi.org/10.1016/j.patter.2022.100676

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*. https://doi.org/10.1080/14703297.2023.2190148

Csikszentmihalyi, M. (2014). Toward a psychology of optimal experience. In (pp. 209–226). Springer Netherlands. https://doi.org/10.1007/978-94-017-9088-8_14

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

Dowling, M., & Lucey, B. (2023). ChatGPT for (Finance) research: The Bananarama Conjecture. *Finance Research Letters, 53*, 103662. https://doi.org/10.1016/j.frl.2023.103662

Eich, T. S., Stern, Y., & Metcalfe, J. (2013). The hypercorrection effect in younger and older adults. *Aging, Neuropsychology and Cognition, 20*(5), 511–521. https://doi.org/10.1080/13825585.2012.754399

Google Cloud Tech (2021). Transformers, explained: Understand the model behind GPT, BERT, and T5 [Video]. YouTube. https://www.youtube.com/watch?v=SZorAJ4I-sA&t=481s

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. ArXiv. https://arxiv.org/abs/2301.01768

Jiao, W., Wang, W., Huang, J-T, Wang, X., & Tu, Z. (2023). Is ChatGPT A good translator? Yes With GPT-4 As The Engine. ArXiv. https://arxiv.org/abs/2301.08745

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognising one's own incompetence lead to inflated self-assessments [Article]. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kurdiez Space. (2021). How ChatGPT works technically for beginners [Video]. YouTube.
Macdonald, C., Adeloye, D., Sheikh, A., & Rudan, I. (2023). Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. *Journal of Global Health, 13*, 01003. https://doi.org/10.7189/jogh.13.01003
Oviedo-Trespalacios, O., Peden, A., Cole-Hunter, T., Costantini, A., Haghani, M., Rodriguez, J. E., & J. E., Kelly, S., Torkamaan, H., Tariq, A., Newton, J., Gallagher, T., Steinert, S., Filtness, A., & Reniers, G. (2023). *The Risks of Using ChatGPT to Obtain Common Safety-Related Information and Advice.* https://doi.org/10.2139/ssrn.4346827
Patel, S. B., & Lam, K. (2023). ChatGPT: The future of discharge summaries? *The Lancet Digital Health, 5*(3), e107–e108. https://doi.org/10.1016/s2589-7500(23)00021-3
Rao, A., Kim, J., Kamineni, M., Pang, M., Lie, W., & Succi, M. D. (2023). Evaluating ChatGPT as an adjunct for radiologic decision-making. *medRxiv*, 2023.2002.2002.23285399. https://doi.org/10.1101/2023.02.02.23285399
Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(2), 177–192. https://doi.org/10.1002/wcs.1336
Sobania, D., Briesch, M., Hanna, C., & Petke, J. An analysis of the automatic bug fixing performance of ChatGPT ArXiv. https://doi.org/10.48550/arXiv.2301.08653
Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 25*, 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x
Spronken-Smith, R., Walker, R., Batchelor, J., O'Steen, B., & Angelo, T. (2011). Enablers and constraints to the use of inquiry-based learning in undergraduate education. *Teaching in Higher Education, 16*(1), 15–28. https://doi.org/10.1080/13562517.2010.507300
Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: Many scientists disapprove. *Nature, 613*(7945), 620–621. https://doi.org/10.1038/d41586-023-00107-z
Stracke, E., & Kumar, V. (2010). Feedback and self-regulated learning: Insights from supervisors' and PhD examiners' reports. *Reflective Practice, 11*(1), 19–32. https://doi.org/10.1080/14623940903525140
Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science, 379*(6630), 313–313. https://doi.org/10.1126/science.adg7879
Vygotsky, L. S., Cole, M., John-Steiner, V., Scribner, S., & Souberman, E. (1978). *Mind in society development of higher psychological processes*. Harvard University Press.
Vygotsky, L. (1986). *Thought and language*. MIT Press.
Wang S., Scells, H., Koopman, B., Zuccon, G. (2023). Can ChatGPT write a good Boolean query for systematic review literature search? ArXiv. https://arxiv.org/abs/2302.03495

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com