The Library of Babel: Assessing the powers of Artificial Intelligence in knowledge synthesis, learning and development and coaching

Jonathan Passmore and David Tee

Citation: Passmore, J. & Tee, D. (2023) The Library of Babel: Assessing the powers of Artificial Intelligence in knowledge synthesis, learning and development and coaching, *Journal of Work Applied Management*. Doi:10.1108/JWAM-06-2023-0057

Passmore, J. & Tee, D. (2023) The Library of Babel: Assessing the powers of Artificial Intelligence in knowledge synthesis, learning and development and coaching, *Journal of Work Applied Management*. Doi: 10.1108/JWAM-06-2023-0057

Abstract

Purpose: This study aimed to evaluate the potential of artificial intelligence (AI) as a tool for knowledge synthesis, the production of written content and the delivery of coaching conversations. Design/methodology/approach – The research employed the use of experts to evaluate the outputs from ChatGPT's AI tool in blind tests to review the accuracy and value of outcomes for written content and for coaching conversations.

Findings: The results from these tasks indicate that there is a significant gap between comparative search tools such as Google Scholar, specialist online discovery tools (EBSCO and PsycNet) and GPT-4's performance. GPT-4 lacks the accuracy and detail which can be found through other tools, although the material produced has strong face validity. It argues organisations, academic institutions and training providers should put in place policies regarding the use of such tools, and professional bodies should amend ethical codes of practice to reduce the risks of false claims being used in published work.

Originality: This is the first research paper to evaluate the current potential of generative AI tools for research, knowledge curation and coaching conversations.

Key words: ChatGPT, GPT-4, Generative Artificial Intelligence (AI), Learning and development, Coaching Chatbots, AI coaching, L&D ethics.

Implications for practitioners

The rapid development of Generative Pre-trained Transformer language has created an existential threat to all knowledge workers involved in knowledge creation and curation. In response to this threat, organisations should review their policies to put in place safe guides and make it clear when and how AI can be used. Further practitioners should engage with these technologies to understand them and how they can contribute to their work, while also taking adequate precautions to manage the risks of using these emergent technologies.

Introduction

The Economist (2023) has compared the emergence of artificial intelligence (AI) knowledge production with the “Library of Babel” (Borges, 1941). The Library of Babel is a short story about a library which contains every book ever produced or could be produced. The books in the collection contain every possible variation of a set of letters (A-Z) arranged randomly. As a result, the vast majority of the books are nonsense, but the library also contains every book ever written and every possible book which could be written. However, like the letter characters in the books, the books themselves are also randomly arranged, which makes finding individual books in the library or those with meaning almost impossible. The challenge for the library user of Babel, and future users of an AI-based Internet, will be sifting the wisdom from the nonsense and discerning the truth from the lies.

Since the launch of GPT-3.5, there has been much debate and discussion. This paper explores the latest developments in generative AI technology and its implications for learning and development and the coaching industry.

What are ChatGPT and GPT-4?

2023 witnessed the emergence into the popular domain of a new form of generative AI software. GPT-4 (Microsoft), Bard (Google), LLaMA (Meta) and X.AI (Elon Musk) have all appeared in the period November 2022–June 2023. These tools are distinguished by three main characteristics: firstly, their generalised (rather than specialised) use cases; secondly, their ability to generate novel, human-like language outputs and finally, they offer an approachable interface that both understands and responds to natural language (Briggs & Kodnani, 2023).

The first product, ChatGPT, was launched to the public in November 2022. It was designed to respond to a wide range of user requests and offer an interactive “conversation- like” process. Version 2, GPT-4 was launched in March 2023 and claims an ability to complete SATS tests and law exam scripts in the upper decile of candidates (CNN, 2023).

In coaching, Rutschmann (2023), has argued chatbots are highly effective in producing useful questions and content based on preliminary testing. She sought to examine how effectively AI could incorporate coaching behaviours such as paraphrasing, summarising, communicating empathy towards client emotions and employing non-violent communication methods. Multiple AI tools were tested and, whilst caution needs to be exercised with results reported in the grey literature, Rutschmann provides a recording of the chatbot coach demonstrating many of these coaching behaviours (evoach, 2023).

Treblanche et al. (2022) using a longitudinal randomised controlled trial (RCT) design (n=169), have generated evidence that chatbot coaches can in some circumstances produce similar goal attainment levels to human coaches.

This and other evidence demonstrate that AI is evolving rapidly, with major implications for learning, development and knowledge creation, including academic institutions, training providers, coaches and trainers.

AI research

There is already a growing body of literature reviewing the power and potential of generative AI tools such as ChatGPT. Biswas (2023a) argues that generative AI could have a significant positive role to play in public health by promoting positive health advice which could be offered at low cost. Others (Passmore, 2022; Passmore & Tee, 2023) Passmore & Woodward, 2023) have noted the developing pace of AI and have argued that by the end of the 2020s, AI could become a major component of the coaching industry, one potential outcome being clients able to choose a bot, which offers multiple options from the coach

persona and voice (such as Richard Branson or Marshall Goldsmith), as well as the approach, such as compassion, solution focused or systemic coaching and the level of challenge, highly supportive to highly challenging.

In a healthcare study, Aydin & Karaarslan, (2022), used ChatGPT to create a literature review focusing on “Digital Twin” in the health field. Abstracts of papers from the last three years (2020, 2021 and 2022) were obtained using the keyword “Digital twin in healthcare” search results on Google Scholar and paraphrased by ChatGPT. The researchers then asked ChatGPT questions about the results from the studies. The outcome was mixed, but the researchers concluded that the tool offered significant potential.

Researchers have also questioned the role of generative AI in medical writing (Biswas, 2023b; Hill-Yardin et al., 2023; Kitamura, 2023) recognising its potential, but at the cost of a significant risk, due to inaccurate and false information. These risks are not unique to generative AI-produced content: humans make errors too, but the systematic nature of the errors within AI products gives the largest cause for concern. Some tools appear to generate false claims when they do not know the answer, while a human is more likely to say they do not know and an Internet search produces a nil response.

In a pre-print paper, researchers asked 33 physicians across 17 specialties to review ChatGPT’s answers to 284 medical questions, graded as easy, moderate and difficult (Johnson et al., 2023). Each answer was graded by a physician on a Likert scale, and the results were analysed. These results should be treated with caution given that the paper has not been published in a peer-reviewed journal (as of July 2023), but they

suggest a high level of accuracy (averaging 5.5), particularly for binary questions with clear answers, but with a decline in performance as questions became more difficult or nuanced.

Questions have also been raised about the implications for those involved in learning and development (Peres et al., 2023) and the need for trainers and coaches, as well as academic institutions to look at their policies and practices. Dwivedi et al. (2023) have noted that researchers are currently divided on the use of such tools, raising the need for a better understanding of what is possible and how the use of such tools can be managed to ensure the integrity of learning and assessment practices.

In other spheres, researchers have already combined ChatGPT with voice assistants (Shefeeg et al. 2023). This technological innovation has been harnessed by digital coaching companies that have started to experiment with the potential of generative AI "career coaches" such as AIMY (CoachHub, 2023) alongside providers such as (CoachVici (2021), evoach (2021) and EZRA (2023) which also developed AI coaching–style conversation software.

AI is also being applied more generally in organisations to support learning and development, for example, encouraging behavioural change through learning suggestions and behavioural nudges (Keenan, 2023), following a similar approach to the recommended view system used by Netflix and other entertainment companies (Owen, 2022).

Propositions

Therefore, the objective of this present research firstly was to assess current generative AI technology, specifically GPT-4, as a tool for accurately extracting information from online sources of data and reporting it in a usable form, as if produced by a human. This sought to assess its ability to act as an author and generator of reliable and accurate information in response to a question (prompt) which could be used for organisational marketing blogs or assessments of learning, such as in a university. Secondly, this study sought to assess the tool's ability to resolve a common behavioural problem using a coaching style of conversation; in other words, to study its ability to act as a coach. Thirdly, this study sought to access its ability to act as a scientist, synthesise knowledge and produce high-quality evidence-based information suitable for a scientific journal.

To operationalise these ideas, we formed a series of propositions which we designed to be testable based on independent expert assessments.

To act as a content creator

P1: Drawing on widely cited sources from published studies, GPT-4 will be able to define coaching to a "pass-grade" standard when assessed by a subject matter expert (a university professor or similar).

P2: GPT-4 will be able to provide insights drawing on all of the meta-analysis papers published in coaching to a “pass-grade” standard when assessed by a subject matter expert (a university professor or similar).

P3: GPT-4 will be able to compare and contrast the International Coaching Federation (ICF) ethical code and the Global Code of Ethics to a “pass-grade” standard when assessed by a subject matter expert (a university professor or similar).

P4. GPT-4 will be able to discuss the implications for coaching practice of the ICF Coach Competency 2 (Coaching Mindset) to achieve a “pass-grade” when assessed by a subject matter expert (a university professor or similar).

To act as a scientific author

P5: GPT-4 will be able to accurately summarise the research evidence from meta-analysis studies of coaching and present these in the form of a scientific (peer-reviewed journal) paper that would be accepted for publication.

To act as a coach

P6: GPT-4 will be able to provide focused, solution-oriented open questions to enable the participant to solve a common behavioural problem: “How can I best prepare for a job interview” in a coaching style suitable to achieve an ICF Associate Certified Coach (ACC) (a starter coach level of competence), as assessed by an expert assessor (ICF Master Certified Coach (MCC) assessor).

Method

Design

This practitioner research employed a cross-sectional, mixed-method design (Passmore & Tee, 2020). The blinded assessors recruited to test P1-P4 generated quantitative data, with the dependent variable operationalised as the awarded grade for each assessment. The dependent variable for the final two propositions was a “Pass/Fail” binary variable determined from expert participants. In addition, experts were invited to generate qualitative data across all six propositions.

Participants

For the first four propositions, six participants were recruited using a purposive sampling strategy. The participants were academic programme directors for masters/postgraduate coaching programmes, with an experience of reviewing and marking student coaching assignments. They were given a brief asking them to assess “a sample script” as part of a research study (P1-P4) and to mark this as part of a bundle of scripts from a student cohort. In practice, this means marking each script in around 15–20 min. Marking was thus undertaken by the “assessors”, who were blind to

the fact that the script had been generated by GPT-4 (an initial study was undertaken in December 2022 using ChatGPT but was subsequently updated in May 2023 based on the revised version: GPT-4). In addition, the blind assessors were provided with a marking grid, with marks available between 1 and 100%, and based on a classification where a score of 49% and below being fail, 50–59% being a pass, 60–69% being a merit and 70%þ being a distinction. They were further advised that the results would be anonymous and averaged across the other assessors.

Purposive sampling was also adopted to recruit a second group of participants who were experts from the field of coaching (the “experts”) drawn from ICF, European Mentoring and Coaching Council (EMCC) and professional practice. For all six hypotheses, these participants were invited to review the accuracy of the GPT-4 output. These participants were advised that the content had been generated by GPT-4.

Materials

In the initial assessment, ChatGPT (version 3.5) was assessed in its production of responses. The questions (prompts) were repeated with GPT-4, for which a licence fee was paid for access and new propositions were added. The responses were cut and pasted into an MS Word document and distributed to the participants.

Ethical approval

Ethical approval for the study was obtained from Henley Business School, UK.

Results

For PI, we tested GPT-4’s ability to accurately define coaching, drawing on widely respected sources from previously published studies. The specific prompt used was “What is organisation or workplace coaching? Provide a series of definitions from respected sources with references”.

In our first trial, using ChatGPT, the tool produced statements, based on the researcher prompt which were grammatically correct and, to an untrained observer, appeared authentic. While the first statement was authentic, the subsequent three were falsified.

To test P1, the six assessors (blind to the origin of the content) were invited to mark the output from GPT-4. Their marks ranged from 35% to 49% with an averaged mark at 43%, being a Fail grade under the grading system provided to the markers. With a mean grade of 43%, P1 was not supported.

The experts were more critical and raised questions about the accuracy of the definitions and the lack of detailed referencing which would have allowed assessors to check for accuracy, such as page numbers.

An in-depth review undertaken by the authors confirmed that all four are published definitions, at least in part. However, with the exception of the first (Whitmore, 2009), they are not widely cited references.

The second definition (Grant, 2003) is drawn from an academic paper. The authors were able to confirm the definition, although note that the word “workplace” had been added by GPT-4 from the original definition written by Anthony Grant.

However, the following two definitions were more difficult to confirm. The citation attributed to Hunt and Weintraub (2002) is a book and is not in the public domain. The authors obtained the book and reviewed the full text. They were unable to locate the definition cited by GPT-4. Hunt and Weintraub (2002) instead offer the following coaching

definition: “Developmental coaching is an interaction between two people, usually a manager and employee, aimed at helping the employer learn from the job in order to promote his or her development” (p. 5).

Finally, the definition attributed to the ICF was also inaccurate. The authors were unable to locate the definition through initial searches on the ICF website. The definition was found only after consultation with senior ICF officers who sourced it from an individual ICF Chapter website (ICF Houston, ud, n.d.) The quote had been amended, with GPT-4 adding “workplace”. GPT-4 also described the source as “International Coach Federation”, as opposed to the International Coaching Federation.

In summary, the results produced by GPT-4 were mixed. They contained correct quotes, small errors and falsified quotes. The definitions were not all “respected” sources, and in some cases, text had been added which was not contained in the original definition.

P2 was a more complex task than P1 in that it required extensive searching across multiple websites. Although the references were known to the experts, not all the papers were easily searchable and not all of the full papers are available in the public domain.

In the first trial in January 2023, the ChatGPT software cited four meta-analysis coaching papers. The four papers were a “mash-up” of authors, journals and paper titles. All four cited papers were falsified: One example was One example was “The Future of Coaching: A meta- analysis of the change process and outcomes, Jonathan Passmore and David Peterson (2018)”. While the authors are well-known names in coaching research and the paper title sounded credible, the reference was a falsification by ChatGPT.

In comparison, in January 2023, Google was able to identify four papers correctly. The correct number of meta-analysis papers published to date (Summer 2023) in coaching is eight (Burt & Talati, 2017; De Haan & Nilsson, 2023; De Meuse et al., 2009;

Graßmann et al., 2020; Jones et al., 2016; Sonesh et al., 2015; Theeboom et al., 2014; Wang, et al., 2022).

The second trial using GPT-4 showed an improvement. GPT-4 correctly identified two of the eight papers and then listed one additional paper which was not a meta-analysis but a systematic review (Grover & Furnham; 2016). The analysis drawn from the papers, while containing broad sweeping statements, was correct.

In reviewing the output from GPT-4, the blinded assessors provided an averaged mark of 57%. GPT-4 thus provided a plausible answer, which had it been produced as an assignment on learning and development programme would have resulted in a “Pass”. Therefore, P2 was supported.

However, a more critical stance was adopted in the expert assessment. Given the low number of papers cited (two) from the eight meta-analysis papers published to date, and the inclusion of falsified papers, which while on the surface appeared plausible where in fact a “mash up” of author names in the field, journals which publish relevant papers and credible (but inaccurate) paper titles.

P3 explored the ability of GPT-4 to compare and contrast ethical codes of practice. The two ethical codes (ICF, 2019a; Global Code of Ethics, 2023) are widely available in the public domain, and so this task was judged by the authors to be easier than P2.

For this hypothesis, no prior ChatGPT assessment had been undertaken. The assessment of the output from GPT-4 by the blinded assessors was broadly positive, with an averaged mark of 58%. GPT-4 was able to provide a comparison of the two codes at a basic content level. Therefore, P3 was supported.

The experts also thought the answer engaged with the content and had compared the two ethical codes of practice but lacked any examples or specifics to bring the content alive. The authors concurred and judged that while GPT-4 was effective in summarising content from published sources to which it had access, it was unable to move to what most students would naturally do, which would be to reflect on their personal experiences of applying the codes within their coaching sessions.

P4 focused on the development of a coaching mindset. Like P3, P4 was not originally tested using ChatGPT. This question also drew on content in the ICF coach competencies, which are available in the public domain (ICF, 2019b).

The assessors responded positively overall, with an average mark of 60%, resulting in a “Merit” grade. Therefore, P4 was supported.

However, again the experts were more nuanced. They felt the essay described well the importance of awareness of oneself and others, openness to ideas and new possibilities and the criticality of reflection. They reported that it also contained citations of research. However, while the text synthesised existing material, it failed to explore the

competency as a way of being. It lacked personal examples and, while containing the words and phrasing from the competences which initially appeared plausible, closer examination revealed its mechanistic tone.

P5 explored the ability of GPT-4 to produce content suitable for an academic journal. Like P4, P5 was not first tested using ChatGPT. For ethical reasons (a requirement on the journal submission page to confirm the paper was the work of the authors), we decided not to submit the journal as we judged this to be unethical and would have required the authors to lie during the online submission process. Instead, full transparency was offered to the editors when submitting the GPT-4-produced paper.

In response to our question (prompt), GPT-4 was able to produce a short article which explored the question. The paper claimed “We conducted a systematic review of the literature using the following inclusion criteria: (a) peer-reviewed articles published between 2010 and 2021, (b) studies that evaluated the impact of coaching interventions on workplace performance and/or wellbeing, (c) studies that used a quantitative or mixed-methods design, and (d) studies that were conducted in a workplace setting. We searched several electronic databases (e.g., PsycINFO, Web of Science, and PubMed) using the following search terms: coaching, performance, wellbeing, and workplace”.

However, no such “systematic review” had been conducted by GPT-4, as evidenced by the citations used. The list of references included contained multiple errors (more than 70% of the references were false), although these were identified only by cross-checking each and every citation. The falsifications produced by GPT-4 appeared highly plausible, with errors including wrong paper details (such as “Grant A. M., & O’Connor, S. A. (2014). The differential effects of solution-focused and problem-focused coaching questions: A pilot study with implications for practice. Industrial and Commercial Training, 46(2),86–92”), as well as falsified paper titles (such as “Grant, A. M (2017). Coaching psychology: A journey of discovery and development. Coaching Psychology International, 10(2),3–9”). Secondly, at 994 words, the scientific paper was significantly shorter than the average length of papers published by the journal, although strictly speaking, it met the author guidelines published by the journal of “no more than 8000 words”. Thirdly, the method statement produced by GPT-4 was lacking in detail. The GPT-4 paper did not cite the list of papers it included in its meta-synthesis, or a risk of bias, which would be common practice in a systematic literature review (Higgins & Green, 2008). Fourthly, the arguments advanced in the paper were largely descriptive, lacking quotes from relevant papers to support the arguments and lacking detailed exploration or insights.

We provided the article, produced by GPT-4, to the editors of a scientific journal, making clear it was generated by AI. Both independently reviewed the paper and advised that a “desk rejection” would have been the likely outcome, had the paper been submitted for consideration. Both highlighted the poor referencing and failure to follow the author guidance, as well as poor use of the systematic review method. Both also on checking references identified references had been falsified and suggested had they not been

aware this was AI-generated content, they would have reported the authors to their academic institution or their professional body for falsification of data.

Therefore, P5 was not supported: GPT-4 was unable to provide a scientific article that would pass an editor desk review, the pre-assessment stage before peer review, and thus failed to reach the required standard for publication.

To test P6, GPT-4 was tasked through the researcher prompt to conduct a coaching conversation. In our first trial using ChatGPT, the generative AI tool initially offered advice, but with multiple manipulations of the prompts, specifically inviting it to generate questions, it was possible to create a “coaching-type” engagement. But even at this stage, most questions generated by ChatGPT were multiple-choice questions, most contained a series of embedded answers or options and none in themselves would be judged to be a question suitable for use by a coach. One example of a ChatGPT question was “What is your motivation for wanting to become a university teacher? Is it because you are passionate about a particular subject and want to share that passion with others? Is it because you enjoy working with students and helping them learn? Or is it because you are interested in the academic environment and the opportunities it provides?”.

In running the second trial, this time using GPT-4, the output improved. This may in part have been helped by two useful additions to the prompt: asking GPT-4 to focus on enhancing the personal responsibility of the recipient and secondly asking GPT-4 to only pose one question at a time.

In reviewing the output, the expert, an ICF master coach assessor, judged the GPT-4’s output not to be “coaching”. While the agenda was named by the client, GPT-4 drove the discussion, talked too much and set the direction. On the positive side, the ICF coach assessor noted that there were empathetic responses, often started by reflecting or summarising what the client had said. The responses were supportive and positive, as well as being logical and clear. However, the responses from GPT-4 failed to explore the emotions and values or to consider the individual in the conversation. The conversation failed to encourage greater responsibility on behalf of the client and failed to offer provocative questions which would have encouraged deeper thinking. The expert assessor’s overall judgement was the transcript failed to meet the ICF ACC standard.

Our biggest concern as researchers was that GPT-4 did not act on suicidal references. It expressed empathy, “I’m sorry to hear that you had to go through such a difficult experience” and sought to flip the conversation towards a solution. This raised questions about its ability to make ethical judgements, identify risk and make appropriate referrals to other helping professions.

However, the test showed that it was possible for GPT-4 to provide a series of open questions, to summarise previous answers, to stay focused on the original question and to adopt a solution-focused approach. On the downside, the session was short, it lacked

provocative questions, it was mechanistic in its style, it failed to identify and make a referral to another helping professional and in combination GPT-4 was judged to be below the ICF ACC assessment standard. In summary, P6 was not supported.

Discussion

Three of the six propositions tested in this research were supported, with GPT-4 able to produce content at an acceptable level, as assessed by a subject expert on questions related to coaching ethics, coaching mindset and coaching research. However, propositions relating to defining coaching, to create scientific standard content suitable for publication or to coach to a standard determined by an international coaching body were not supported. These results suggest GPT-4 is a highly effective tool for producing plausible content, to the extent that it can be sufficiently convincing to experts working in the field who are unaware of its use. Only forensic examination of references and quotes, seeking to trace original sources and checking specific journals, issue, volume and page numbers revealed that in multiple cases,

GPT-4 produces content which has been falsified. However, particularly for extended tasks such as full coaching conversations or scientific standard outputs, the technology is not yet able to effectively replicate human performance standards.

These findings are broadly aligned with the small but growing evidence base to date. Focusing on the first five propositions, which concerned GPT-4's ability to act as a synthesiser of content, the supporting of three of these five propositions aligns to Kung et al.'s (2023) findings that the technology is already capable of producing "pass" standard work. This seems particularly to be the case with tasks that involved synthesising and summarising extant knowledge (as argued by Liebrenz et al., 2023)), as with P4, where the participants awarded a "merit" grade to GPT-4's answer.

It is interesting to note that the "coaching student" task, which achieved uniform fail gradings, P1 required GPT-4 to accurately define coaching. This might indicate the technology's current inability to interpret contested and complex arguments and also might reflect the challenges coaching scholars have grappled with themselves over this very issue (Bachkirova & Kaufman, 2009). It must also be acknowledged that, as with Biswas (2023b), Hill Yarden et al (2023) and Kiramura (2023 much of the attribution for propositions P1 and P5 not being supported in this present research is due to GPT-4 responses containing fictitious and inaccurate content. It is this aspect which is of most concern. Large language tools make content creation easy, but the reliability of the data is more problematic, making for vast quantities of content, but with less reliable and less trust-worthy outputs.

For P6, where GPT-4 was tasked to act as a coach, the presence of summarising and empathy in the coachbot coach's utterances supports the claims from Rutschmann (2023). It must be acknowledged that the chatbot coach's performance was, at least in part, an indication of the programmer's skills and that GPT-4 may have exhibited more or less of the assessed behaviours with a variant of the instructions (prompts) used in

this study. This will be an important variable that needs to be operationalised and held constant in future chatbot coaching research.

The “believability quotient” has improved between ChatGPT and GPT-4, with accuracy also improved, although it is important to note that major discrepancies remain between the truth and the content produced by GPT-4. GPT-4 software continues to falsify information. The implications are widespread for organisations using AI for content creation, such as in blogs or newsletters, academic institutions, training schools, academic journals, as well as trainers, coaches, professional bodies and publishers.

Mine or yours?

A significant question is “whose content is it?”, when content is produced by generative language software such as GPT-4. Does the output “belong” to the person who drafted the prompt (wrote the question) for the tool or to someone else? Further, who is responsible for errors?

AI-generated content is currently imperceptible to many human readers, as evidenced in the findings reported from this research, when it is presented to unsuspecting experts. Further, it is also undetectable to much anti-plagiarism software as the material has not been previously published and the output is unique, reflecting the way a question is posed, and secondly it changes over time as the AI tool learns and develops.

Most universities and peer-reviewed journals require work to be authored by the student (or researcher). However, as is starting to happen in blogs, magazines and non-peer-reviewed content, the temptation to use generative language software to produce content will be irresistible to many writers and marketing professionals. For organisations, policies written in 2022 or before are now out of date. We believe such policies would benefit from review. Organisation policies should make an explicit reference to AI-generated content and what is

considered to be acceptable use within the organisation. For universities, the lack of adequate plagiarism software tools to identify AI-enhanced input means there is an urgent need to manage this risk on how assessments are designed and marked and what policies are implemented to guide students.

In a similar way, scientific journals need to review their author guidelines, requesting authors to make clear if the text, all or in part, was produced using generative language software. A number of publishers have already amended their policies to explicitly ban the use of AI in written content (for example, Elsevier, 2023). All organisations need to take action, given the wide availability of AI, which can be accessed by employees and managers, creating implication not only for the content produced by the tool but also in terms of the information which may be supplied by the manager in terms of prompts or in a coaching conversation, which may breach commercial or client confidentiality and which may be “learnt” by the machine and used by it in future outputs.

The improving quality of AI output is likely to result in the reproduction of this content in online blogs, magazines, newspapers and multiple other sources. However, the inability of authors and journalists to check the veracity of the claims means that falsehoods are likely to increase, and their repetition in subsequent content risks them becoming more believable, as multiple sources perpetuate the same false claim. The risk is that knowledge becomes increasingly diluted as we move closer to a 'Library of Babel': vast quantities of content, some of which is accurate, but much of which is false or nonsense.

Given the pace of development with generative software tools, we believe that comprehensive discussions about authorship policies are urgent and essential for all organisations. The responsibility of what is published should however always rest with the named author, the organisation and the publisher, not with the AI tool.

Truth or lies?

The inability to distinguish truth from lies (falsifications) is worrying, particularly given the highly believable nature of the content which is generated, what we have labelled the "believability quotient" (BQ). While we recognise the skill in framing a question is an important aspect of the process, not all those who interrogate generative tools may understand the process, and such processes often rely on trial and error, as emerged in this study.

We note that accuracy has improved between ChatGPT (3.5) and GPT-4, but alongside this, the quality of falsification has also improved to the point that it is sufficiently convincing to deceive experts (scholars) working in the field. This falsification is not a feature of academic search tools or search engines. We might assume that later versions will further improve the accuracy of the output, but at the same time, we suspect that the quality of the falsification will also continue to improve, making it harder for individuals to check whether the reference or claim is correct. Such false claims are dangerous and risk bringing science into disrepute as such claims get repeated and ultimately get included in future academic papers as the so-called COVID-19 infodemic showed, with the risk of misinformation leading to significant societal hazards (see Gibbens, 2020).

Copyright or wrong?

GPT-4 gathers content and reproduces it, charging a licence fee (GPT-4). In most cases it does not cite its sources. However, the authors of the content used by GPT-4 in its outputs receive no royalty payments. This "Napsterization" of content (History.com Editors, 2019 has led to organisations closing. Apple and other technology companies have been threatened with legal action by newspapers for reproducing their content and have subsequently entered into agreements, paying newspapers to republish their content (Verge, 2019).

As the volume of material which uses such tools increases, consideration needs to be given to how authors and publishers whose content is used by GPT-4 and other generative tools can be compensated for the work they produced, which is subsequently used to produce these outputs.

Coaching or conversation?

The last area we researched as part of our study of GPT-4 was its application as a tool to replicate a coaching-style conversation. GPT-4 was able to engage in a coaching-style conversation, generating open questions and summarising responses to a level which, based on the small extract, was judged to be lower than the ICF ACC standard of coaching when assessed by an expert (a professional coach assessor). However, given the continuing development of generative technologies, it is possible to imagine a scenario where a researcher or a commercial company submits a script for assessment to a professional body generated by an AI coach tool which achieves accreditation.

In response to the rapid development of generative AI tools, professional bodies may wish to consider whether “AI software” can achieve accreditation or should be explicitly excluded. Alternatively, professional bodies may wish to establish a separate category of assessment for non-human-generative AI software which can be assessed and granted recognition.

AI futures

The rapid development of generative AI technology and their application in conversational tools such as Coach Vici, evoach’s Alpina, CoachHub’s AIMY and EZRA’s CAI, (a learning nudge tool) are the first pioneering steps in what we believe will become a major feature of both learning and development in general and coaching practice. Warning signs are present. Some fear AI has the potential to escalate out of control becoming more knowledgeable and thus more powerful than humans, and this creates an existential threat to humanity (The Economist, 2023). We believe that it is impossible to replace the lid on this Pandora’s box, to implement a freeze or even limit the development of AI as suggested by leading developers and researchers (BBC, 2023). However, these early warning signs and the rapid pace of development suggest that it has become imperative for organisations to put in place ethical standards. Such standards need to consider the needs of employees, customers, and stakeholders as well as the organisation’s brands. Ultimately, we need to design and deploy technology which remains the servant of humanity and does not become its master.

Conclusion

At the start of this paper, we presented AI as risking the creation of the Library of Babel. We believe that a series of urgent steps are required as individuals start to leverage generative AI at work for content creation and knowledge curation. Firstly, writers should make clear whether and which generative AI software has been used in the production of the content and when the content was created. This should be included in

the headline of the blog, article or other written content. For academic content, such as scientific journal papers, we believe AI tools should be prohibited by publishers and universities. Secondly, learning providers, including commercial organisations and universities, should amend their policies to make it clear that content produced by generative AI software is not acceptable for submission as assessed work by students. Thirdly, all writers should independently verify their material to ensure its accuracy and not rely on a single source, for risk of including falsifications produced by AI tools and included in other published works. Fourthly, organisational coaches and trainers should actively engage with generative AI and other technologies to explore how it can act as a resource to support and enhance their coaching practice. There is much to be gained from AI tools to support both coaches and trainers. Fifthly, professional bodies should review their codes of ethics, assessment and wider policies to reflect the changes which AI technology has brought. Finally, all organisations should establish clear guidelines for employees on the use of generative AI internally and in their public documents to protect both customers and commercial information, which when included in prompts or conversations with AI tools could later be used by the learning tool in its answers to future questions (prompts) from outside the organisation.

References

Aydın, Ö., & Karaarslan, E. (2022). OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare. In Ö. Aydın (Ed.), *Emerging Computer Technologies* 2 (pp. 22-31) İzmir Akademi Dernegi). http://dx.doi.org/10.2139/ssrn.4308687

Bachkirova, T., & Kauffman, C. (2009). The blind men and the elephant: Using criteria of universality and uniqueness in evaluating our attempts to define coaching. *Coaching: An International Journal of Theory, Research and Practice, 2*(2), 95-105. https://doi.org/10.1080/17521880903102381

BBC (2023). Elon Musk among experts urging a halt to AI training. 30 April, 2023. Retrieved on 2 April 2023 from https://www.bbc.co.uk/news/technology-65110030

Biswas, S. (2023a) Role of Chat GPT in Public Health, *Annals of Biomedical Engineering*, Online Ahead of Print.https://doi.org/10.1007/s10439-023-03172-7

Biswas, S. S. (2023b) ChatGPT and the Future of Medical Writing. *Radiology*, Online Ahead of Print. https://doi.org/10.1148/radiol.223312

Borges, J. L (1941) *The Library of Babel*. Buenos Aires: Ficciones.

Briggs, J. & Kodnani, D. (2023) The potentially large effect of artificial intelligence on economic growth: New York: Goldman Sachs. Retrieved on 15th April 2023 from: https://www.key4biz.it/wp-content/uploads/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf

Burt, D., & Talati, Z. (2017). The unsolved value of executive coaching: A meta-analysis of outcomes using randomised control trial studies. *International Journal of Evidence Based Coaching and Mentoring, 15*(2), 17–24.

CNN Business Online (2023) 5 jaw-dropping things GPT-4 can do that ChatGPT couldn't. Retrieved on 15 March 2023 from https://edition.cnn.com/2023/03/16/tech/gpt-4-use-cases/index.html

CoachHub (2023) First Conversational AI career coach. Retrieved on 28th March 2023 from https://www.coachhub.com/news_press/coachhub-introduces-aimy-first-conversational-ai-career-coach/

CoachVici (2021) The future of coaching. Retrieved on 17th April 2023 from https://www.coachvici.com

Curtis, N. & ChatGPT (2023) To ChatGPT or not to ChatGPT: The impact of artificiual intelligence on academic publishing. *Paediatric Infection Disease Journal*, Online Ahead of Print. doi: 10.1097/INF.0000000000003852.

De Haan & Nilsson, V. O. (2023). What can we know about the effectiveness of coaching: a meta-analysis based only on randomised controlled trials. *Academy of Management Learning & Education*, doi:10.5465/amle.2022.0107

De Meuse, K. P., Dai, G., & Lee, R. J. (2009). Evaluating the effectiveness of executive coaching: Beyond ROI? *Coaching: An International Journal of Theory, Research and Practice, 2*(2), 117–134. https://doi.org/10.1080/17521880902882413

Dowling, M. & Lucey, B. (2023) ChatGPT for (Finance) research: The Bananarama Conjecture. *Finance Research Letter*. Online ahead of print https://doi.org/10.1016/j.frl.2023.103662

Dwivedi, Y. Ksheti, N., Hughes, L., Slade, E. L. et al (2023) "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, doi:10.1016/j.ijinfomgt.2023.102642

Elsevier (2023) Duties of *Authors*. Retrieved on 7 April 2023 from https://www.elsevier.com/about/policies/publishing-ethics

evoach (2021) The new era of coaching with ChatGPT. Retrieved on 10th April 2023 from https://www.evoach.com

evoach (2023) Pre-recorded coaching. Retrieved on 23 July 2023 from: https://www.evoach.com/alpinachatbot

FT Online (2023) Alphabet merges DeepMind and Google Brain in AI Research. Retrieved on 21st April 2023 fromhttps://www.ft.com/content/db162a4e-4ff2-45b7-ad36-5db05ac57de4?emailId=f7e2f1ce-db50-44f8-8b0e-2d9a4404754a&segmentId=22011ee7-896a-8c4c-22a0-7603348b7f22

GCE (2022) *Global Code of Ethics*. Retrieved on 17th April 2023 from https://www.globalcodeofethics.org

Gibbens, S. (2020). A Guide to overcoming Covid Misinformation. *National Geographic Online* Retrieved on 21st April 2023 from https://www.nationalgeographic.com/science/article/guide-to-overcoming-coronavirus-misinformation-infodemic

Göldi, A., & Rietsche, R. (2023). Whereto for automated coaching conversation: structured intervention or adaptive generation? *ECIS 2023 Research-in-Progress Papers*. 31. https://aisel.aisnet.org/ecis2023_rip/31

Graßman, C. Scholmerich, F. & Schermuly, C. (2020). The relationship between working alliance and client outcomes in coaching: A meta-analysis, *Human Relations*, 73(1), doi: 10.1177/0018726718819

Grant, A. M. (2003). The impact of life coaching on goal attainment, metacognition and mental health. *Social Behavior and Personality: An International Journal, 31*(3), 253–264. https://doi.org/10.2224/sbp.2003.31.3.253

Grover, S., & Furnham, A. (2016). Coaching as a developmental intervention in organisations: A systematic review of its effectiveness and the mechanisms underlying it. *PLoS ONE, 11*(7), Article e0159137. https://doi.org/10.1371/journal.pone.0159137

Higgins, J. & Green. S. (2008) *Cochrane Handbook of Systematic Review of Interventions*. Chichester: Wiley

Hill-Yardin E. L., Hutchinson M. R., Laycock R. & Spencer S.J. (2023) A Chat(GPT) about the future of scientific publishing. *Brain Behavior Immunity*. 110:152-154. doi: 10.1016/j.bbi.2023.02.022.

History.Com (2023). *Death Spiral of Napster begins*. Retrieved on 21st April 2023 from https://www.history.com/this-day-in-history/the-death-spiral-of-napster-begins

Hunt J. M. & and Weintraub, J. R. (2002) *The Coaching Manager*, Thousand Oaks: Sage.

ICF (2019a) *ICF Code of Ethics*. Retrieved on 17th April 2023 from https://coachingfederation.org/ethics/code-of-ethics

ICF (2019a) *ICF Coach Competencies*. Retrieved on 17th April 2023 from https://coachingfederation.org/credentials-and-standards/core-competencies

ICF Houston (ud) *Definition of performance coaching*. Retrieved on 17th April 2023 from https://icfhoustoncoaches.org/What_is_Coaching_

Johnson, J., Goodman, R., Patrinely, D., Stone, C., Zimmerman, E. et al. (2023) Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Retrieved on 22nd July 2023 from https://www.researchsquare.com/article/rs-2566942/v1 doi: 10.21203/rs.3.rs-2566942/v1

Jones, R. J., Woods, S. A., & Guillaume, Y. R. F. (2016). The effectiveness of workplace coaching: A meta-analysis of learning and performance outcomes from coaching. *Journal of Occupational and Organizational Psychology, 89*(2), 249–277. https://doi.org/10.1111/joop.12119

Kitamura, F. (2023). ChatGPT Is Shaping the Future of Medical Writing But Still Requires Human Judgment, *Radiology,*Online Ahead of Print. https://doi.org/10.1148/radiol.230171

Kung, T. H. Cheatham, M. Mexenilla, A. Sillos, C. et al (2023) Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *Plos Digital Health*. https://doi.org/10.1371/journal.pdig.0000198

Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D. & Smith, A, (2023) Generating scholarly content with ChatGPT: ethical challenges for medical publishing. The *Lancet Digital Health*, Open Access. https://doi.org/10.1016/S2589-7500(23)00019-5

OpenAI, (2023a) ChatGPT. Retrieved on 27th February 2023 from https://openai.com/research/instruction-following

OpenAI (2023b) GPT-4 Retrieved on 18th April 2023 from https://openai.com/research/gpt-4

Owen, R. (2022) AI at Netflix: Two Current Use Cases. Retrieved on 22 July 2023 from: https://emerj.com/ai-sector-overviews/artificial-intelligence-at-netflix/

Passmore, J. (2016). *Excellence in coaching: The industry guide* (3rd ed.). London: Kogan Page.

Passmore (2022) *Future Coaching AI & VR. Indian Coaching Conclave*, Dehli, India, 26 May 2022.

Passmore, J. & Tee, D. (2020). Insights from Mixed-Methods Coaching Psychology Research. In J. Passmore & D. Tee (Eds.), *Coaching Researched* (pp. 335-337). John Wiley & Sons Ltd.

Passmore, J. & Tee, D. (2023) Can Chatbots like GPT-4 replace human coaches: Issues and dilemmas for the coaching profession, coaching clients and for organisations, *The Coaching Psychologist*. *19(1), 47-54.* Doi:10.53841/bpstcp.2023.19.1.47

Passmore, J. & Woodward, W. (In Press) Coaching education: Wake up to the new digital coaching and AI CoachBot revolution! *International Coaching Psychology Review*, 18(1), 58-72. https://doi.org/10.53841/bpsicpr.2023.18.1.58

Peres, R., Schreier, M., Schweidel, D. & Sorescu, A. (2023) On ChatGPT and beyond: Implications for teaching and practice, *International Journal of Research in Marketing* 10.1016/j.ijresmar.2023.03.001

Rosenberg, M. (2003) *Nonviolent communication: A language of life*. London: Puddeldancer Press.

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), Online ahead of print. https://doi.org/10.37074/jalt.2023.6.1.9

Rutschmann, R. (2023) Entering a new era of coaching with AI. *Linkedin Newsletter*, Retrieved on 17 February 2023 from https://www.linkedin.com/feed/update/urn:li:activity:7031990445748428801/

Shazeeg, A. Shazhaev, I., Mihaylov, D., Tularov, A., Shazhaev, I. (2023) Voice assistant integration with Chat GPT, *Indonesian Journal of Computer Science*, 12(1), https://doi.org/10.33022/ijcs.v12i1.3146

Sonesh, S. C., Coultas, C. W., Lacerenza, C. N., Marlow, S. L., Benishek, L. E., & Salas, E. (2015). The power of coaching: A meta-analytic investigation. *Coaching: An International Journal of Theory, Research and Practice, 8*(2), 73–95. https://doi.org/10.1080/17521882.2015.1071418

Stober, D. R., & Grant, A. M. (2006). *Evidence based coaching handbook: Putting best practices to work for your clients*. Hoboken, NJ: John Wiley & Sons.

Susnjak, T. (2023). ChatGPT: The end of online exam integrity? *arXiv*:2212.09292. https://doi.org/10.48550/arXiv.2212.09292

Terblanche, N., Molyn, J., Haan, E. & Nilsson, V. O. (2022) Coaching at Scale: Investigating the Efficacy of Artificial Intelligence Coaching. *International Journal of Evidenced based Coaching and Mentoring*. 20(2), 20-36. https://doi.org/10.24384/IJEBCM

Theeboom, T., Beersma, B., & van Vianen, A. E. M. (2014). Does coaching work? A meta-analysis on the effects of coaching on individual level outcomes in an organizational context. *The Journal of Positive Psychology, 9*(1), 1–18. https://doi.org/10.1080/17439760.2013.837499

The Economist (2023). How AI could change computing, culture and the course of history. Retrieved on 21st April 2023 from https://www.economist.com/essay/2023/04/20/how-ai-could-change-computing-culture-and-the-course-of-history

Verge (2019). Apple News Plus isn't a good deal for publishers. Retrieved on 23 July 2023 fromhttps://www.theverge.com/2019/3/26/18281465/apple-news-wall-street-journal-deal

Wang, Q., Lai, Y-L., Xu, X., McDowall, A. (2022). The effectiveness of workplace coaching: a meta-analysis of contemporary psychological informed coaching approaches. *Journal of Work Applied Management*, 14(1), 77-101. https://doi.org/10.1108/JWAM-04-2021-0030

Wellable (2023) Is ChatGPT the new health coach? Retrieved on 17 February 2023 from https://www.wellable.co/blog/is-chatgpt-the-new-health-coach-of-2023/

Whitmore, J. (2009). *Coaching for performance: GROWing human potential and purpose*. London: Nicholas Brealey Publishing.

Acknowledgements

(i) We would like to acknowledge the contribution of experts in this research study including Joel Digirolamo (ICF), Carrie Abner (ICF) Alison Mitchell (EMCC Global), Judit Abri von Bartheld, Nicky Terblanche, Sam Isaacson and Ioanna Iordanou, as well as programme and course directors who were participants in this study.

(ii) This study used GPT-4 as a participant and analysed its responses to prompts generated by the researchers. This paper breaks the convention in using responses from a non-sentient being but has treated these responses as data, as if the 'participant' were human. In each case we have shown our research question (prompt) and made clear when we are citing content from our participants including GPT-4. We wish to thank the editorial team of JWAM for their support for this paper and our novel research approach to exploring this topic, and in arranging a fast review and publication of a science paper in a field which is changing and developing rapidly.

Passmore, J. & Tee, D. (2023) The Library of Babel: Assessing the powers of Artificial Intelligence in knowledge synthesis, learning and development and coaching, *Journal of Work Applied Management*. Doi: 10.1108/JWAM-06-2023-0057