ChatGPT and Generative AI Guidelines for Addressing Academic Integrity and Augmenting Pre-Existing Chatbots

Daswin De Silva*, Nishan Mills*, Mona El-Ayoubi†, Milos Manic‡, Damminda Alahakoon*
*Centre for Data Analytics and Cognition, La Trobe University, Melbourne, Australia
†Education Services, La Trobe University, Melbourne, Australia
‡Department of Computer Science, Virginia Commonwealth University, Richmond, USA

2023 IEEE International Conference on Industrial Technology (ICIT) | 979-8-3503-3650-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIT58465.2023.10143123

***Abstract*—Chat Generative Pretrained Transformer (ChatGPT) and related Generative AI models are leading a paradigm shift in the acceptance and application of Artificial Intelligence (AI) across all disciplines and industry sectors. Despite the criticisms of an 'intelligence without knowledge or reasoning or the notions of truth', ChatGPT is highly effective at human-like conversation with seemingly sophisticated and useful responses to questions, summarization, classification, extraction and generation tasks. Unlike similar large AI models in the modalities of image, audio and video, text-based conversation is straightforward and familiar to a large audience of regular users of the Internet and smartphone applications. This is further accentuated by the large-scale adoption of 'standard' chatbot technologies for trivial conversations in task-specific automation, across every industry sector. This rare combination of highly effective human-like conversation, familiarity of foundational technology and versatility of intelligent application, has led to several challenges and opportunities in leveraging generative AI. A primary challenge is its impact on the academic integrity of scholarly work, where AI-generated content can be useful and detrimental in both teaching and research. On the other hand, ChatGPT presents a unique opportunity in augmenting pre-existing ('standard') chatbots with human-like conversation for advanced intelligent automation, across all application domains. Although diametrically opposed, the challenge of addressing academic integrity and the opportunity of augmenting pre-existing chatbots are grounded in the conversational AI capabilities of ChatGPT and similar generative AI models. In this paper, we investigate these formative capabilities and present guidelines for leveraging ChatGPT and similar generative AI models.**

***Index Terms*—ChatGPT, Academic Integrity, Generative AI, Intelligent Automation, Chatbot, Artificial Intelligence, Pretrained Language Models, GPT3**

I. Introduction

The evolution of Natural Language Processing (NLP) to Natural Language Understanding (NLU) and Generation (NLG) is signified by the transition of Artificial Intelligence (AI) algorithms for time-series forecasting into next word prediction and text sequence prediction. This line of algorithms have rapidly matured into Pretrained Language Models (PLM). PLM are trained using large-scale text corpora and deep learning algorithms, such as Word2Vec [1], GloVe [2], and Bidirectional Encoder Representations from Transformers (BERT) [3]. The first generation of PLM is composed of skip gram and continuous bag of words (CBoW), where the PLM learns the semantic meanings of words, but fails to recognise context and concepts. Second generation PLM address this gap by learning polysemy, syntactic structures, semantic roles, anaphora, etc, through contextual embeddings, as demonstrated in CoVe [4], ELMo [5], GPT [6], BERT [3]. Trained on 570 GB of data and 175 billion parameters, Generative Pretrained Transformer 3 (GPT-3) has outperformed state of-the-art PLM benchmarks for universal language representation [7]. Surpassing its numerous technical applications in NLP and NLU [8], [9], GPT-3 was finetuned with crowdsourced human feedback using reinforcement learning to build Chat Generative Pretrained Transformer (ChatGPT), the world's first PLM with human-like conversational capabilities [10]. Despite its inherent weaknesses of an intelligence without knowledge, lack of reasoning, and factual errors, the conversational AI capabilities of ChatGPT and similar generative AI models are transforming and/or disrupting every discipline, industry sector and field of study [11]–[14], with reported instances of its role in writing scientific abstracts [15], academic assessments [16], curriculum design [17], and healthcare applications [18], [19].

We begin this article by unpacking ChatGPT and deliberating its conversational AI capabilities, followed by the challenge of addressing academic integrity, and the opportunity of augmenting pre-existing chatbots with human-like conversation for advanced intelligent automation. This challenge and opportunity are contrastive but stem from the same formative conversational AI capabilities of generative AI, which are further investigated to contribute a set of guidelines for leveraging generative AI, such as ChatGPT.

At the time of writing this article, ChatGPT was the flagship generative AI model, reaching a 100 million monthly active users just two months after its launch, the fastest-growing consumer application in human history [20]. However, by the time of submitting the revised version of this paper, Meta AI released a competitor model with public access, Large Language Model Meta AI (LLaMA) [21] and OpenAI released their next version, GPT-4 and ChatGPT-plus [22]. This exponential pace of AI innovation demonstrates the evolving disposition of generative AI towards a generalised AI. Although this article focuses on ChatGPT as an exemplar (and widely-used) generative AI model, the proposed guidelines and contributions of this article are equally applicable across current and forthcoming generative AI models.

II. Unpacking ChatGPT

The conversational AI capabilities of most generative AI models, including ChatGPT, are primarily attributed to the additional layer of Reinforcement Learning from Human Feedback (RLHF) introduced into GPT-3. RLHF [23] utilises human feedback as a measure of performance loss to optimize the language model for generated text, thereby the training the model has already received from a general corpus is fine-tuned to that of complex human needs. Despite the limited documentation on the composition of ChatGPT itself, the enabling RLHF capability is reported in a sibling model to ChatGPT, InstructGPT [10]. InstructGPT deliberates the human labelling process where a large dataset of manually labelled Q&A was specifically constructed for the RLHF task. These labels are classified into ten categories, they are, 1) chat, 2) open Q&A, 3) closed Q&A, 4) text classification, 5) text summarisation, 6) text generation, 7) information extraction, 8) text rewrite, 9) brainstorming, as well as hybrid combinations of the above. It is primarily through these label classifications that the language model responds to questions and conducts a meaningful conversation with a human operator. Within a conversation, these label classifications can be used singularly or as multiples, which is then recognised as a conversational prompt. Even in general use, it is recommended that human operators utilise suitable prompts to communicate with ChatGPT. Therefore, conversational prompts are critical in obtaining the most meaningful response. Despite its brief existence, frequent users of ChatGPT have defined this as 'prompt engineering' or the creation of prompt ensembles when the required information/conversation is multi-dimensional. The composition of conversational AI capabilities and prompt engineering in the context of ChatGPT and the human operator, is illustrated in Figure 1. The conversational prompts from left to right are in increasing order of complexity, where Q&A and chat are the simpler formulations, and brainstorming, rewrite and information extraction are complex.

III. Addressing Academic Integrity

Academic integrity is a subjective term with diverse interpretations, ranging from student acts of plagiarism and contract cheating to the values, behaviour and conduct of staff [24], [25]. In this paper, we adopt a broad reference to teaching, learning, research and service undertakings by all stakeholders (staff and students) within a tertiary education setting. This spans across the authenticity of coursework assessment submissions, assessor feedback for such submissions, coursework design, authorship and originality of research articles, grant applications and related research writing.

In both teaching and research, brainstorming should be the first point of engagement with ChatGPT and similar generative AI models. This human-machine collaboration in brainstorming for new ideas, strategies, methods and the critical review of pre-existing content signals a paradigm shift of all academic environments. This shift is grounded in the history of rational thought and academic traditions of the Socratic method for systematic questioning, inductive reasoning, and universal definitions that is systematic, disciplined and granular [26]. More functionally, ChatGPT can be leveraged in academic research to reduce time to publication, improve research novelty and innovation, and make research findings more accessible and equitable [11], while in teaching, it adds value as a learning support tool in the classroom, a bench-marking tool for assessments and a training tool for new staff. Despite its demonstrable 'intelligence' in numerous application settings, its single point of failure is the absence of 'knowledge' or not truly knowing the real-world meaning, context and relationships of the text responses that are generated. If unchecked/unverified by the human operator, ChatGPT can produce factual inaccuracies, logical fallacies, bias and plagiarism. This anomalous behavior is popularly referred to as 'AI hallucinations' or 'stochastic parroting'. ChatGPT replicates chunks of text quickly and easily. The challenge of academic integrity lies therein, where assessment can be plagiarised as the human operator (student, researcher, professional) can present ChatGPT output as their own. Furthermore, there is a braided challenge as ChatGPT recreates versions of others knowledge without any accurate reference to the original source. At first glance, generated content may appear plausible as it is validated statistically by its training sources, however, it has no mechanism for substantiating the source of its 'expression', thus it becomes a solipsistic trap, blurring fact from fiction. A key concern is the reproduction of ChatGPT into assessment submissions and research articles. The use of generative AI tool by learners to respond to prevailing conventional assessment practices built on low order understanding (multiple choice, simple programming, true/false quizzes) of knowledge and application, pose high risk to academic integrity as responses can be plagiarised.

While there are detection tools being developed to recognise AI-generated text, it is a far more effective approach to invest time in transforming assessment and developing strategies to integrate authentic assessments activities in student learning. Authentic approaches support in the mitigation of risk and limit academic integrity misconduct. The emphasis of authentic assessment on solving real-life problems, where knowledge is dynamic, built around processes of discovery [27], support and improve learning through reflective, critical, and collaborative practice [28]. However, there is a vital need to mainstream authentic assessment practices beyond the domain of theory, exemplars, and outlier practice. Reengineering learning and teaching strategies to focus on the process of learning in contrast to the product of learning, provides opportunities for active reconstruction of curricula.

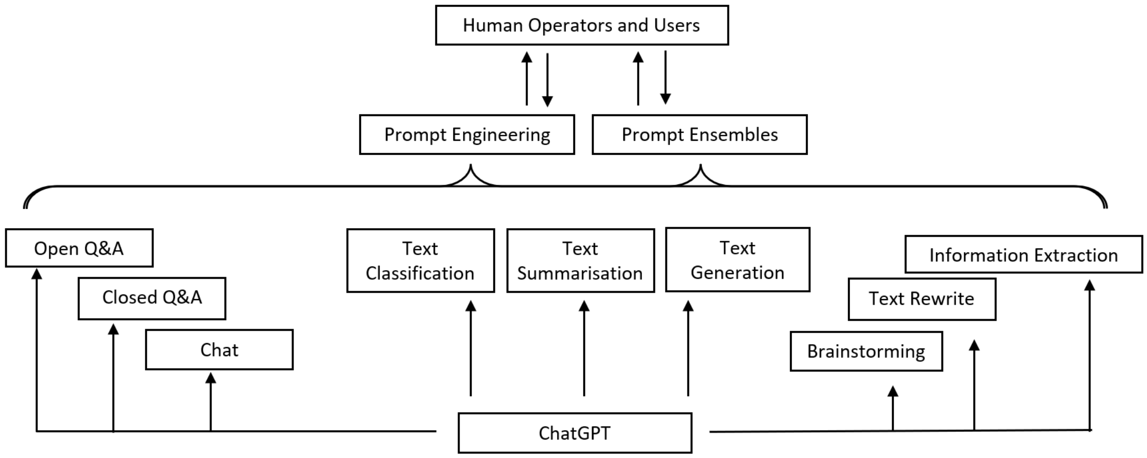

Fig. 1. Conversational AI Capabilities of ChatGPT and Human Operator Interface

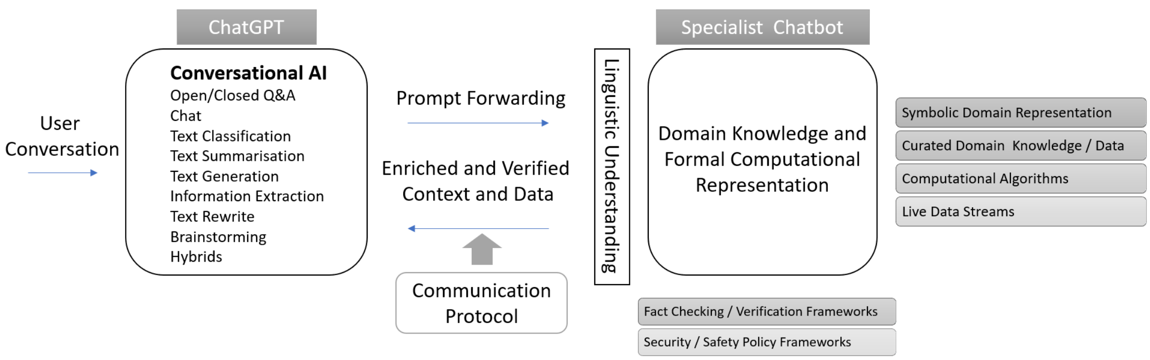

Fig. 2. Functional framework for augmenting a pre-existing chatbot with ChatGPT capabilities

Such constructivist approaches position the learning process at the core of the student experience and is a vehicle for active, meaningful, and engaged learning [29]. This requires closer attention to the design of authentic assessment activities and pedagogical approaches that focus on scaffolding critical higher order thinking to build student capabilities that emphasize critical thinking, complex problem solving, logic development, creativity, and collaboration. Such skills are not easily emulated by Generative AI tools and can be embedded in assessment through task sequencing, prompts, feedback, and timely interventions [30] . The engagement of students in the process of learning makes it difficult to directly plagiarise. This necessitates innovative assessment practices that incorporate methodologies to encourage students to leverage ChatGPT as a supporting technology, to build learner capabilities in critical and independent thinking, where students produce knowledge from fact-checking, verification processes, synthesising and critiquing ChatGPT content. To this end, we propose a set of guidelines grouped by the ten (currently known) conversational AI capabilities of ChatGPT (Table 1).

IV. Augmenting Pre-Existing Chatbots

In this section, we focus on the impact and contributory role of ChatGPT and generative AI in augmenting pre-existing chatbot capabilities, across all disciplines and industry domains, not limited to the education sector. These pre-existing chatbots are more formally known as the automation of communicative labour, and they undergone an exponential investment and application in commercial settings for the automation of repetitive tasks and some intelligent conversational capabilities [31], [32]. Numerous chatbots are reported in recent research literature, such as mental health chatbot with cognitive skills for personalised behavioural activation [33], [34], chatbots for real-time monitoring and co-facilitation of patient-centered healthcare [35], emotion awareness in industrial chatbots [36]. At the time of writing, ChatGPT is a state-of-the-art conversational AI that is far superior to any contemporary chatbot. As noted prior, it is not computationally aware, does not have a notion for whether the information being presented is factually correct and cannot consider recent data. These shortcomings mean that it cannot reliably perform in specialist subject domains, again highlighting the importance of the need for the deployment of authentic assessment.

TABLE I
GUIDELINES FOR ADDRESSING ACADEMIC INTEGRITY AND AUGMENTING PRE-EXISTING CHATBOTS BY CHATGPT CAPABILITIES

Capabilities	**Addressing Academic Integrity**	**Augmenting Pre-Existing Chatbots**
Closed Q&A	In its manifestation as Multiple Choice Questions (MCQ), this should be evaluated for its obsoleteness as traditionally these have been designed with limited complexity. However, if it is unavoidable, then the questions should redesigned to include current events and local disciplinary context.	Closed Q&A has direct application in narrowing down the requirements of the end-user and introducing a logical flow into the chatbot intents design where the output from Closed Q&A flows into creative decisions and branches in the subsequent questions.
Chat	Students can use ChatGPT outputs to evaluate function and better understand the assessment question and context. This should be contextualised locally, peer reviewed and discussed in the classroom to dissuade direct replication.	A pre-existing chatbot is unlikely to possess general conversational capabilities, ChatGPT can provide this directly through its Chat function, which is a significant performance improvement in terms of human-like conversation skills.
Open Q&A	Students can use ChatGPT outputs to evaluate responses, identify factual inaccuracies, logical fallacies, bias (AI hallucinations). Students can compare AI outputs through fact checks and verification from peer-reviewed sources. This supports the build of higher order skills.	Integration with a pre-existing chatbot with domain knowledge ensures that the factual inaccuracies, logical fallacies, bias can be significantly reduced, as it operates similar to fact-checking by a human operator.
Classify	This capability is more frequently used in combination with others, while in its singular use it refers to being able to identify prompts such as, emotion, sarcasm, toxicity, erroneous programming code, type of programming code construct; ask for additional context and explanation how the answer was reached with reference to classroom examples	This capability is highly effective for intent classification in a pre-existing chatbot, where the purpose, direction and potential resolution for end-user queries can be directly simplified through integration
Summary	Summarization can be used as a support technique that enables the student to better understand the assessment task through the inclusion and exclusion of topics being assessed. Furthermore, in parallel to a human-written summary of an assessment submission, an AI-written summary can be used to develop critical and analytical thinking capabilities	End-user queries can be efficiently summarised into noun-verb combinations that map on to the next stage of automation, typically, actions and decisions, this would again simplify the chatbot operation and timeliness of outcomes.
Generate	The most effective practice is to use ChatGPT to generate a baseline solution that students can use as a point of reference, to question plausibility, critique responses to prompts and to produce knowledge.	A pre-existing chatbot will benefit from the generation of conversational cues, topics of interest based on past queries and opportunities for prolonged dialogue where the chatbot intention is continuing the conversation in order to generate an operational lead or improved user satisifaction, for example.
Extract	Typically in comprehension style assessments, ensure it is not a singular task of information extraction only, nest the extraction task as part of a summative assessment requiring further analysis that involves other modalities of data, as well as local context, classroom discussion, and recent/current events.	The extraction of implicit meaning would be a significant, novel capability for a pre-existing chatbot to know beyond the immediate intent, this also extends into the extraction of emotion and sentiment, where applicable
Brainstorm	Provides an immediate boost to the baseline levels of knowledge within a classroom or an assessment task. To avert plagiarism, students should be provided an interactive assessment task that involves a viva voce	Dependent on the application domain, brainstorming can increase participation with the end-user or user groups, where the pre-existing chatbot channels the general conversational capabilities of ChatGPT to draw out topics of interest, opinions and new ideas.
Rewrite	This should be typically nested as part of a larger assessment task, customised to local disciplinary context, and involving a viva voce or video presentation with classroom discussion to minimise academic integrity misconduct.	Helpful in extending the inherent capabilities and target audience of a pre-existing chatbot. Conversational cues, questions, answers and intents can be rewritten so that age group and other socio-demographic factors are taken into account.
Hybrid	Hybrid attempts must be deconstructed into their constituent capabilities and addressed using one or more of the above guidelines. A graph representation would be useful to distinguish strategies for addressing academic integrity	Hybrid capabilities are an opportunity to expand the purpose and utility of the pre-existing chatbot across multiple, intersecting operational domains or industry domains

In contrast, contemporary chatbots with limited linguistic capabilities have been developed to address domain specific queries reliably due to their ability to codify questions posed in terms of a formal representation. These formal representations can be translated to computational models and are informed by real time data. This is a prime opportunity for integration, where ChatGPT and the specialist chatbot display diversity which is symbiotic, and can be leveraged through the Socratic Models (SM) framework [37] to interact as an indivisible unit augmenting the pre-existing specialist chatbot. In this environment ChatGPT would be the interface to the user and the specialist chatbot would ”coach” ChatGPT through prompt engineering to be able to handle domain specific data, fact check, safety check and provide real time updates to the data. Figure 2 present a functional framework for this integration between ChatGPT and pre-existing chatbots.

If we consider a conversation between the Chat System and a human user agent in the Prompt Forwarding phase ChatGPT will forward through prompts which require a domain context through to the Specialist Conversational Agent using the communication protocol that has been established between the two chatbots. The communication protocol would define the specific nuances needed to ensure communication between the two chatbots. The Natural Language Understanding layer of the Specialist Conversational Agent will process the prompt forwarded by ChatGPT using its natural language understanding capabilities and codify it in a formal computational representation which is relevant to the domain, this can then be processed by the engine room of the chat bot to generate reliable, accurate and up to date information. The Specialist will draw on its expertise which is signified by 1) Symbolic Domain Representation – Domain information codified in a discoverable connected format, 2) Curated Domain Knowledge / Data – Databases of information and data which encompass the domain knowledge, 3) Computational Algorithms – Computational algorithms and calculations which are codified and, 4) Live Data Streams – Up to date information through Data Streams and APIs. The specialist will then use this information to enrich, fact check, safety check and curate ChatGPT's responses to the user. The Fact Checking / Verification Frameworks function as filters for this integration. ChatGPT in itself has no mechanism to verify or rationalise the conversations it is synthesising, so the fact checking and verification frameworks will monitor text generated by ChatGPT and appropriate either modify or block the propagation of that text to the user, using policies which can Fact Check ChatGPT's assertions. Similarly, in order to align with security and safety policy mechanisms the Security / Safety Policy Frameworks, will hold security policies of the organisation or domain and also safety information which will integrate any domain specific and general guard rails in information sharing. This would also include indications of prompt injection or hacking attacks which might be mounted by malicious users and also general rules around diversity, non-discrimination, fairness and the prevention of harm. In Table 1, we further deliberate this integration in terms of the core capabilities of ChatGPT.

V. Conclusion

ChatGPT and related generative AI models have triggered a paradigm shift in how we use, interact and even perceive technology at work and in our lives. It is also signalling the new challenges and opportunities of AI and large models trained on billions of data. In this paper, we unpacked the functionality of ChatGPT to identify its conversational AI capabilities, and then explored the role of these capabilities in addressing the challenge of academic integrity and the opportunity in augmenting pre-existing chatbots with human-like conversation for advanced intelligent automation. As future work, we intend to formalise these guidelines into an inclusive framework for leveraging generative AI across diverse discipline and industry settings.

References

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[2] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[3] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Tech. Rep.

[4] B. McCann, J. Bradbury, C. Xiong, and R. Socher, "Learned in translation: Contextualized word vectors," *Advances in neural information processing systems*, vol. 30, 2017.

[5] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, "Allennlp: A deep semantic natural language processing platform," *arXiv preprint arXiv:1803.07640*, 2018.

[6] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[8] R. Dale, "Gpt-3: What's it good for?" *Natural Language Engineering*, vol. 27, no. 1, pp. 113–118, 2021.

[9] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, pp. 681–694, 2020.

[10] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *arXiv preprint arXiv:2203.02155*, 2022.

[11] E. A. van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting, "Chatgpt: five priorities for research," *Nature*, vol. 614, no. 7947, pp. 224–226, 2023.

[12] H. H. Thorp, "Chatgpt is fun, but not an author," pp. 313–313, 2023.

[13] A. Flanagin, K. Bibbins-Domingo, M. Berkwits, and S. L. Christiansen, "Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge," *JAMA*, 2023.

[14] A. Borji, "A categorical archive of chatgpt failures," *arXiv preprint arXiv:2302.03494*, 2023.

[15] H. Else, "Abstracts written by chatgpt fool scientists," *Nature*, vol. 613, no. 7944, pp. 423–423, 2023.

[16] F. J. García-Peñalvo, "The perception of artificial intelligence in educational contexts after the launch of chatgpt: Disruption or panic?" 2023.

[17] D. Baidoo-Anu and L. Owusu Ansah, "Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning," *Available at SSRN 4337484*, 2023.

[18] F. Sanmarchi, D. Golinelli, and A. Bucci, "A step-by-step researcher's guide to the use of an ai-based transformer in epidemiology: an exploratory analysis of chatgpt using the strobe checklist for observational studies," *medRxiv*, pp. 2023–02, 2023.

[19] K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. Sabel, J. Ricke *et al.*, "Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports," *arXiv preprint arXiv:2212.14882*, 2022.

[20] Person and K. Hu, "Chatgpt sets record for fastest-growing user base - analyst note," Feb 2023. [Online]. Available: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[22] B. D. Kleinman and Zoe, "Openai announces chatgpt successor gpt-4," Mar 2023. [Online]. Available: https://www.bbc.com/news/technology-64959346

[23] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3008–3021, 2020.

[24] B. Macfarlane, J. Zhang, and A. Pun, "Academic integrity: a review of the literature," *Studies in higher education*, vol. 39, no. 2, pp. 339–358, 2014.

[25] B. E. Whitley Jr and P. Keith-Spiegel, "Academic integrity as an institutional issue," *Ethics & Behavior*, vol. 11, no. 3, pp. 325–342, 2001.

[26] R. Paul and L. Elder, "Critical thinking: The art of socratic questioning," *Journal of developmental education*, vol. 31, no. 1, p. 36, 2007.

[27] H.-M. Huang, "Toward constructivism for adult learners in online learning environments," *British journal of educational technology*, vol. 33, no. 1, pp. 27–37, 2002.

[28] C. E. Hmelo-Silver, R. G. Duncan, and C. A. Chinn, "Scaffolding and achievement in problem-based and inquiry learning: a response to kirschner, sweller, and," *Educational psychologist*, vol. 42, no. 2, pp. 99–107, 2007.

[29] J. H. Askeroth and J. C. Richardson, "Instructor perceptions of quality learning in moocs they teach." *Online Learning*, vol. 23, no. 4, pp. 135–159, 2019.

[30] P. Guo, N. Saab, L. S. Post, and W. Admiraal, "A review of project-based learning in higher education: Student outcomes and measures," *International journal of educational research*, vol. 102, p. 101586, 2020.

[31] A. Rapp, L. Curti, and A. Boldi, "The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots," *International Journal of Human-Computer Studies*, vol. 151, p. 102630, 2021.

[32] D. De Silva and D. Alahakoon, "An artificial intelligence life cycle: From conception to production," *Patterns*, vol. 3, no. 6, p. 100489, 2022.

[33] P. Rathnayaka, N. Mills, D. Burnett, D. De Silva, D. Alahakoon, and R. Gray, "A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring," *Sensors*, vol. 22, no. 10, p. 3653, 2022.

[34] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial," *JMIR mental health*, vol. 4, no. 2, p. e7785, 2017.

[35] A. Adikari, D. De Silva, H. Moraliyage, D. Alahakoon, J. Wong, M. Gancarz, S. Chackochan, B. Park, R. Heo, and Y. Leung, "Empathic conversational agents for real-time monitoring and co-facilitation of patient-centered healthcare," *Future Generation Computer Systems*, vol. 126, pp. 318–329, 2022.

[36] A. Adikari, D. De Silva, D. Alahakoon, and X. Yu, "A cognitive model for emotion awareness in industrial chatbots," in *2019 IEEE 17th international conference on industrial informatics (INDIN)*, vol. 1. IEEE, 2019, pp. 183–186.

[37] A. Zeng, A. Wong, S. Welker, K. Choromanski, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.