Enhancing Human-Computer Interaction through AI: A Study on ChatGPT in Educational Environments

Dhruval Kenal Kothari
School of Computer Science and Engineering
Nanyang Technological University
Singapore
dhruval001@e.ntu.edu.sg

Owen Noel Newton Fernando
School of Computer Science and Engineering
Nanyang Technological University
Singapore
ofernando@ntu.edu.sg

***Abstract*— This research investigates the potential of Chat Generative Pre-Trained Transformer (ChatGPT) in Human-Computer Interaction (HCI) within educational contexts. Examining the intersection of Artificial Intelligence (AI) and HCI, the study emphasizes ChatGPT's ability to provide personalized and immediate responses, enhancing student engagement and understanding. A literature review reveals ChatGPT's applications in higher education, while highlighting challenges related to critical thinking in the field of HCI. This paper then outlines our objectives, focusing on answering student queries and generating Multiple-Choice Questions (MCQs) for revision. Experimental results demonstrate the superiority of Custom GPTs, emphasizing the importance of context-specific tailoring. The discussion addresses limitations, proposing future work on model fine-tuning, optimization, and human testing. In conclusion, this research contributes insights into leveraging ChatGPT in HCI education, highlighting its potential for personalized and effective learning experiences.**

Keywords*— *ChatGPT, human computer interaction (HCI), Custom GPT, education, artificial intelligence (AI)

I. Introduction

Recently, the combination of AI and HCI has paved the way for key advancements in various domains. ChatGPT is an innovative language model developed by OpenAI that stands out as a promising tool bridging the gap between AI and education. With its capacity to comprehend context and generate human-like answers, ChatGPT holds the potential to revolutionize the way students engage with educational materials. By providing quick and personalised responses, ChatGPT can directly target the needs of a student individually, offering immediate feedback, and help with understanding complicated concepts. This makes it a promising tool that promotes a student's active participation and cognitive advancement by adapting to their own pace of learning [6]. Popular applications include Medical students consulting MedChatbot [19] for medical knowledge, restaurants receiving allergy information based on users' allergens [20] and healthcare staff using Mandy [21], a chatbot system to automate patient intake process. Generating human-like responses adds a conversational dimension to the learning experience, enhancing student engagement and understanding.

In our context, we focus specifically on HCI. HCI is a multidisciplinary field of study that focuses on the design and use of computer technology, emphasizing on the interaction between people (users) and computers [4]. It involves understanding how users can design interfaces that are intuitive, efficient, and provide a positive user experience [1]. Due to its qualitative nature, HCI thrives in the realm where there is no evident right or wrong answer, and decisions can be made accurately if backed by clear explanations.

In this paper we evaluate the use of Large Language Models (LLMs), such as ChatGPT, within the field of HCI as an influential educational tool for university students. A key concern when it comes to using such LLMs in this context is correctness of responses. This is due to the qualitative nature of HCI, deeming two responses practically accurate if reasoned for correctly. Classic NLP metrics like BLEU score [17] are hence, no longer reliable in this context. We aim to find a way to accurately verify the accuracy of LLM generated solutions for any given input, but verifying problems with a specific context of domain through methods such as translation validation [18] is a challenge in itself, making a general verifier with high accuracy in answering HCI questions even harder. Hence, in this paper we also generate a new evaluation method, consisting of multiple choice questions answered by different LLM models, on questions generated off different topics. Our research objectives are as follows:

RQ1: To test and compare the accuracy of different language models in answering prompts and questions related to HCI, shedding light on their potential usage in university educational settings.

RQ2: To develop a self-improving loop capable of learning from incorrect responses. The goal is to create a system that can autonomously expand its knowledge base over time, contributing to the continuous improvement of its performance in addressing HCI-related prompts and questions.

RQ3: Examine the effectiveness of the self-improving loop in enhancing the language model's performance over time, with emphasis on its adaptability to HCI specific context and terminology.

II. Literature Review

ChatGPT exhibits potential in higher education, helping with writing improvement by generating texts, summarizing information, and detecting grammar errors [7]. The model's application extends to medical education as well, where it demonstrates accuracy in clinical decision-making [8]. Such LLMs are used widely in many different fields especially in the form of chatbots. Education chatbots each teach topics such as psychology (Freudbot), English practice (CSIEC) [22], medical knowledge (MedChatbot) [19] and self-assessment and reflection (CALMsystem) [23]. Other educational applications include the use of a chatbot to answer university related FAQs [24]. Business chatbots supplement customer service for business and e-commerce [25], help with

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00100

task completion [27] and improve user experience [26]. LLMs can also be trained to answer domain specific queries such as information on La Liga, a Spanish football league [28] and museum artifacts (Art-bots).

However, critical thinking and originality of ideas are essential components of genuine learning, which might pose to be a challenge for LLMs like ChatGPT which is trained on substantial amounts of text data [9]. Although ChatGPT is highly trained on large amounts of data, it differs fundamentally from human information processing. The model generates each new token sequentially based on previously generated tokens within the same sequence [13]. This means that it lacks introspective capabilities and cannot self-correct or perform sanity checks on its outputs [14]. Moreover, ChatGPT lacks deep reasoning abilities and is unable to utilize tools such as the internet for more latest information [14]. Various strategies can improve ChatGPT's performance. For example, prompt engineering encourages more accurate responses by spreading reasoning across multiple prompts, a concept known as "chain of thought prompting" [15]. Methods like self-consistency [16] and backtracking [14] allows the model to review its output.

Most large language models possess the ability to accomplish novel tasks with minimal examples, even if they are not directly associated with their initial training [11]. This expands the applications of such models indefinitely, allowing them to be used in endless fields with little to no context.

Specific to the field of HCI, research has been done to establish that ChatGPT is a valid tool that can perform meaningful content analysis and is able to accurately understand content that can be used for HCI research [12]. However, the application of ChatGPT in an educational setting for HCI demands further exploration. In the unique context of HCI, where answers are not strictly right or wrong due to its qualitative nature, the challenge lies in defining appropriate contexts for evaluation. HCI's qualitative nuances require a careful consideration of context to prevent confusion among students leveraging this technology for educational purposes. While existing research establishes ChatGPT's validity, an emphasis on contextual relevance becomes pivotal in educational HCI applications.

III. Methodology

The tech stack we use is Python using the ChatGPT API. Preliminary testing was first done on the ChatGPT playground on the OpenAI website before running multiple prompts via the Python script. As ChatGPT is trained on large datasets of text data, various biases and inaccuracies in this data can be reflected in its responses [2]. Certain prompts received responses that were completely incorrect or contextually wrong. Hence, certain experiments were repeated immediately on receiving outlier responses.

Custom GPTs are custom versions of ChatGPT that are tailored for a specific purpose [3]. In our case, we are creating a custom GPT to answer HCI related questions, passing in textbook and materials as context when initializing our GPT.

To conduct our experiments, we first need to generate MCQ questions. The following steps are followed to automatically generate a question bank that will be used in our experiments:

- Generating knowledge base
- Parsing content and generating question bank
- Data processing

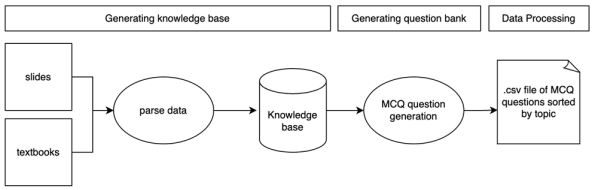

Fig. 1. Architecture diagram showing the processes to generate the knowledge base and form .csv file of MCQ questions

A. Generating knowledge base

This process involves using a diverse array of content materials, carefully curated to ensure complete coverage of HCI concepts. We source information from various channels, including textbooks and PowerPoint slides created by experienced HCI professors. The use of multiple content formats serves 2 main purposes. First, we use condensed and summarized versions of HCI content from professor-prepared slides that captures the essence of the subject matter. Secondly, we tap into the breadth and depth of HCI offered by textbooks, allowing full coverage of all content. Hence, passing in such varied forms of content ensures that our knowledge base is diverse and covers all nuances of HCI in question generation.

B. Parsing content and generating question bank

Next, we split all our content into 10 main topics, following the topic breakdown in HCI textbooks [1]. Our goal is to create a question bank that not only exhibits diversity but also ensure thorough coverage of all the content. We then generate 10 MCQs for each topic, resulting in a total of 100 questions. This approach guarantees a balanced representation of content since certain topics may have varying levels of emphasis within textbooks. By segmenting and generating questions for each topic, we establish a uniform distribution of content that covers the entire range of HCI concepts.

For each question generated, we generate 4 options as well as the correct answer. All of this is done using an automated python script that taps on the ChatGPT API (gpt-3.5-turbo model). The main technology involved here is text summarization, content parsing and question generation. This automated approach streamlines the process while using the power of natural language processing.

C. Data processing

The results from our Python script and the ChatGPT API are in textual format. To ensure seamless handling of this data for our experiments, we transform this text data into a CSV file, providing a tabular structure that enables efficient organization and manipulation of the data for the subsequent experimental flow.

D. Accuracy metrics

Given the qualitative nature of HCI, nuances and contextual understanding plays an essential role in answering questions. As there is no clear-cut right or wrong answer, there arises a challenge to devise a custom accuracy metric tailored to the specific requirements of comparing different Large Language Models (LLMs) in answering HCI questions. Hence, we have produced a custom metric that acknowledges the inherent subjectivity of HCI while providing a concrete

and measurable framework for model evaluation. Model accuracy is calculated using the following formula:

$$Model\ Accuracy\ (\%) = \frac{Number\ of\ correct\ responses\ by\ model}{Total\ number\ of\ questions} * 100$$

The primary focus is passing generated MCQs back into different models and subsequently comparing the number of correct responses each model produces. This is, in some sense, a MCQ examination taken by each model.

We then use these incorrect questions to further "train" the model by collating all such questions into a document and uploading it back as context.

IV. Experimental Setup

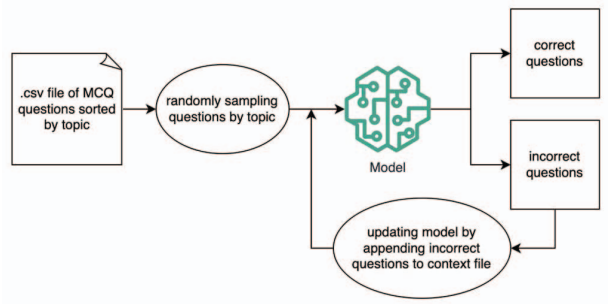

Fig. 2. Architecture diagram showing experimental procedure to test accuracy of model

A. Question sampling and order randomization

For each of the 10 topics, we randomly sample 3 questions. This yields a set of 30 questions, forming the basis for assessing the LLM. To eliminate biases and ensure comprehensive evaluation, we then randomize the order of the selected questions. Hence, we now have varied sequences of questions for each experimental iteration.

B. Choice of Models and evaluation

We employ the use of several different models, each pre-trained on different data, with a different number of parameters and developed by different corporations. This gives us a wide variety of models to test on. We then pass through the sample set of 30 questions into each model. We repeat this process 5 times to account for variability. For each iteration we calculate the accuracy score using the method mentioned above. We then calculate the average of these scores, giving us the overall accuracy score for each model. This experimental workflow ensures a thorough and unbiased evaluation of the LLM model's performance on our curated MCQs. The repetition of the process adds a layer of robustness to our assessment, accurately evaluating the performance of each model.

V. Results and Analysis

TABLE I. Accuracy of Models on Experiments

Model	Accuracy
llama70b-v2-chat	25.16129032258064%
gpt-3.5-turbo-16k	63.13725490196079%
gpt-3.5-turbo	63.52941176470588%
tulu-2-dpo-70b	64.13125398810162%
meta-llama/Llama-2-70b-chat-hf	70.88235294117646%
qwen-14b-chat	72.94117647058825%
mixtral-8x7b-instruct-v0.1	74.11764705882354%
Custom GPT	83.52941176470588%

As expected, we see that Custom GPTs have the best performance in our experiments, as their responses are more contextualized to the books and slides used to generate the MCQs. Other ChatGPT models, such as the gpt-3.5-turbo performs poorly, due to lack of fine-tuning to the task at hand as well as failure to align with the specific context of the generated MCQs. Pre-training the model on a larger set of data also does not reap any benefit as seen from the similar accuracy scores of gpt-3.5-turbo and gpt-3.5-turbo-16k, where the latter is trained on a significantly larger dataset of text data.

VI. Discussion

The superiority of Custom GPT highlights the importance of tailoring language models to the specific context of the task. Fine-tuning or training on domain-specific data significantly enhances performance as we see in our experiments.

Although Custom GPT shows remarkable performance, continuous optimization and fine-tuning could boost accuracy further. Exploring hyperparameter tuning and additional training iterations might give even better results. For example, the following is an incorrect response by the Custom GPT model. The correct answer should be Option A, according to the content this question was generated from.

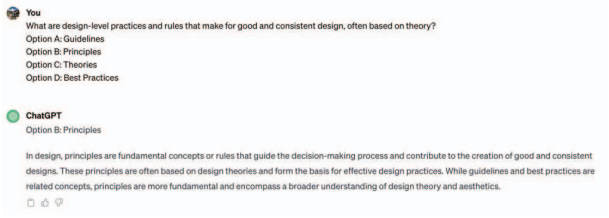

Fig. 3. Example of incorrect response to MCQ question by ChatGPT

However, in our next prompt we ask the same question again, this time passing in the incorrect response as context.

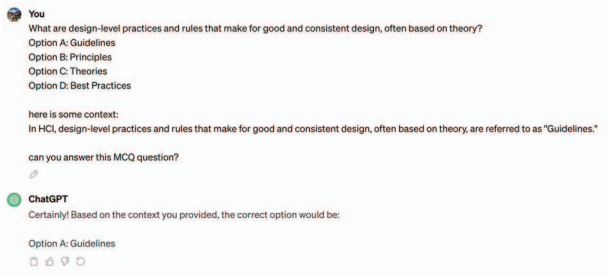

Fig. 4. Example of correct response to previously asked question with additional context

We see that the model recognizes this updated context and now provides the correct response. Hence, it is beneficial to have a self-fulfilling loop of training for the model and allows it to learn from the mistakes made, significantly improving accuracy. As CustomGPTs can take in files as input, we can append incorrect answers to a file and incorporate this data during subsequent training sessions. By constantly updating this file with instances of incorrect responses, we provide the model with a continuous stream of learning material, enhancing its ability to deliver accurate results.

VII. Future work

A. Limitations

While this paper provides valuable insights on the accuracy of LLMs in the field of HCI, it encounters various challenges associated with the use of ChatGPT such as narrow scope and scale of data collection, which focuses on a summarized subset of HCI content. Given limited computation power and resources, experiments were only run for 5 iterations per model. As seen in the discussion section, collating incorrect responses, and passing them back into the model as context reaps correct answer. Multiple iterations of this would hence significantly increase accuracy of custom GPTs in answering HCI questions.

B. Fine tuning and optimization

We can explore further fine-tuning of our models to enhance accuracy. One example is to increase the number of training iterations for our experiments or generate a larger variety of questions from a more diverse knowledge base.

C. Human testing and validation

Another improvement is to conduct our experiments using human participants to evaluate the practical usability and acceptance of AI-generated content in educational settings. Human participants can provide feedback on the generated MCQs and answer them accordingly, together with commenting on the accuracy of the expected answers from the models. Their perspective on correctness contributes to the improvement of the model's knowledge base, addressing any discrepancies that may be overlooked in the training phase. Participants will also help to evaluate the utility of explanations accompanying each option. This feedback helps to refine the quality of information provided to leaners, ensuring that the options are not only accurate but also presented in a manner that is conducive to effective learning.

VIII. Conclusion

In conclusion, this research contributes valuable insights into leveraging ChatGPT in HCI education, specifically using Custom GPTs- showcasing its potential for personalized and effective learning experiences. As AI continues to play an increasingly prominent role in education, the findings of this study pave the way for further advancements and refinements in the integration of AI tools for educational enhancement.

References

[1] Benyon, D. (2014). Designing Interactive Systems: A Comprehensive Guide to HCI and Interaction Design (3rd ed.). Pearson.

[2] Kalla, D., Smith, N., Samaah, F., & Kuraku, S. (2023). Study and Analysis of Chat GPT and its Impact on Different Fields of Study. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4402499

[3] (N.d.). Retrieved from https://openai.com/blog/introducing-gpts

[4] A. Sears, The Human-Computer Interaction Handbook. 2002.

[5] Montenegro-Rueda, M., Fernández-Cerero, J., Fernández-Batanero, J. M., & López-Meneses, E. (2023). Impact of the Implementation of ChatGPT in Education: A Systematic Review. Retrieved from https://www.mdpi.com/2073-431X/12/8/153

[6] Sánchez, O. V. G. (2023). Uso y percepción de ChatGPT en la educación superior. Retrieved from https://riti.es/index.php/riti/article/view/261

[7] Atlas, S. (2023). ChatGPT for Higher Education and Professional Development: A Guide to Conversational AI. Retrieved from https://digitalcommons.uri.edu/cba_facpubs/548/

[8] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., … Tseng, V. (2023). *PLOS Digital Health*, *2*(2). doi:10.1371/journal.pdig.0000198

[9] Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. Retrieved from https://digitallibrary.aau.ac.ae/handle/123456789/980

[10] Reza Hadi Mogavi (2023). ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. Retrieved from https://www.sciencedirect.com/science/article/pii/S2949882123000270

[11] Beltagy, I., Cohan, A., IV, R. L., Min, S., & Singh, S. (2022). Zero- and Few-Shot NLP with Pretrained Language Models. Retrieved from https://aclanthology.org/2022.acl-tutorials.6/

[12] Tabone, W., & de Winter, J. (2023). *Royal Society Open Science*, *10*(9). doi:10.1098/rsos.231053

[13] Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., … Sun, L. (2023). A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. Retrieved from https://arxiv.org/abs/2302.09419

[14] Long, J. (2023). Large Language Model Guided Tree-of-Thought. Retrieved from https://arxiv.org/abs/2305.08291

[15] Amer-Yahia, S., Bonifati, A., Chen, L., Li, G., Shim, K., Xu, J., & Yang, X. (2023). From Large Language Models to Databases and Back: A discussion on research and education. Retrieved from https://arxiv.org/abs/2306.01388

[16] Kim, Z. M., Taylor, D. E., & Kang, D. (2023). "Is the Pope Catholic?" Applying Chain-of-Thought Reasoning to Understanding Conversational Implicatures. Retrieved from https://arxiv.org/abs/2305.13826

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318, 2002.

[18] N. P. Lopes, J. Lee, C.-K. Hur, Z. Liu, and J. Regehr. Alive2: bounded translation validation for llvm. In Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, pages 65–79, 2021.

[19] H. Kazi, B. S. Chowdhry, and Z. Memon, "MedChatBot: An UMLS based Chatbot for Medical Students," International Journal of Computer Applications, vol. 55, no. 17, pp. 1–5, 2012.

[20] P. P.-T. Hsu, J. Zhao, K. Liao, T. Liu, and C. Wang, "AllergyBot," Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17, pp. 74–79, 2017.

[21] L. Ni, C. Lu, N. Liu, J. Liu, J. L. S. o. K. Sciences, Systems, U. 2017, J. Liu, J. L. S. o. K. Sciences, Systems, U. 2017, J. Liu, J. L. S. o. K. Sciences, Systems, U. 2017, and J. Liu, "MANDY: Towards a Smart Primary Care Chatbot Application," Springer, pp. 38–52, 2017.

[22] J. Jia, "CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning," KnowledgeBased Systems, vol. 22, no. 4, pp. 249–255, 2009.

[23] T. F. D. M, "Estimaci on de la densidad con n ucleo variable," Knowledge-Based Systems, 2014.

[24] B. R. Ranoliya, N. Raghuwanshi, and S. Singh, "Chatbot for university related FAQs," 2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017, vol. 2017-Janua, pp. 1525– 1530, 2017.

[25] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, "A new chatbot for customer service on social media," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, 2017, pp. 3506– 3510.

[26] S. Gupta, D. Borkar, C. De Mello, and S. Patil, "An ecommerce website based chatbot," International Journal of Computer Science and Information Technologies, vol. 6, no. 2, pp. 1483–1485, 2015.

[27] B. Behera, "Chappie-a semi-automatic intelligent chatbot," 2016. Retrieved from https://www.cse.iitb.ac.in

[28] C. Segura, J. Luque, and M. R. Costa-juss, "Chatbol, a chatbot for the Spanish La Liga," International Workshop on Spoken Dialog System Technology 2018 (IWSDS), pp. 1– 12, 2018.