HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

doi:10.3233/FAIA240181

Playing the Imitation Game: Human-AI Simulators in Pedagogic Design

Florentina ARMASELU
Luxembourg Centre for Contemporary and Digital History, University of Luxembourg
ORCiD ID: Florentina Armaselu https://orcid.org/0000-0003-2386-6889

Abstract. Current advances in large language models (LLMs) and generative AI (GenAI) have produced both enthusiasm and concerns in the academic world, industry, and society in general. While optimistic views foresee unprecedented increase in efficiency and productivity, concerns have been expressed on the potential of these technologies to determine significant changes in most areas of human activity, which may not always have predictable or positive outcomes. One of the challenges often evoked in this context, not yet fully addressed, is the impact of the AI-powered agents on the educational sector, and especially on aspects such as student's *agency* and *control*, *creativity*, and *motivation* in pedagogic activities that involve the use of this type of agents. The aim of the study is to address this question starting from the hypothesis that preliminary simulations of AI-based pedagogic scenarios can help instructors to better understand the inner mechanisms of these technologies and their possible impact on the learning, assignment completion and evaluation processes. The paper presents a set of experiments with simulated student-agent interactions generated by AI chatbots and proposes a formal framework for assessing this form of "imitation game" and its possible applications in real teaching-learning environments.

Keywords. generative AI, pedagogic simulation, digital humanities

1. Introduction

One of the most far-sighted intuitions in Turing's imitation game [1] resides in the capacity of the machines central to his argument to simulate a theoretically unlimited number of imitation game experiments. Current advances in large language models (LLMs) and generative artificial intelligence (GenAI) [2,3,4] that open up a broad field of applications, some of them prefigured by Turing's game, have produced both enthusiasm and concerns in the academic world, industry, and society in general. While optimistic views foresee unprecedented increase in efficiency and productivity, concerns have been expressed on the potential of these technologies to determine significant changes in most areas of human activity, which may not always have predictable or positive outcomes [5,6]. One of the challenges often evoked in this context, not yet fully addressed, is the impact of the AI-powered agents on the educational sector, and especially on aspects such as student's *agency* and *control*, *creativity*, and *motivation* in pedagogic activities that involve the use of this type of agents [7,8,9,10].

Recent studies and experiments related to the use of AI in education and creativity assessments have focused on AI-based teaching strategies [11], evaluation of students'

creativity in AI-assisted writing and modelling scenarios [12,13], or interviews with generative AI chatbots asked to formulate opinions about their own impact on higher education and academic publishing systems [14]. This fast-rate developing area of research provides a large unexplored territory that requires a combination of man- and machine-based ingenuity to predict how human and AI factors together may influence aspects such as student's agency and control, creativity, and motivation in the technology-mediated classroom of the future. In this context, what forms of prediction can be imagined? The aim of this study is to address these questions starting from the hypothesis that preliminary simulations of AI-based pedagogic scenarios can help instructors to better understand the inner mechanisms of these technologies and their possible effects on learning, assignment completion and evaluation processes in creative tasks. The paper presents a set of experiments with simulated student-agent interactions generated by AI chatbots and proposes a formal framework for assessing this form of "imitation game" and its possible applications in real teaching-learning environments. Sections 2 and 3 elaborate on the methodology and proposed simulations, while 4 and 5 discuss the results and reflect on findings and their possible applications.

2. Methodology

The experiments have been designed as a preliminary test bed for a course, taught by the author, in generative AI and creative writing intended for the next academic year and graduate students in the humanities, involved in language, literature and history studies.

Table 1. Simulators created using GPT Builder (ChatGPT-4, 27.01-08.02.2024)

AlphaStudentSimulator	Simulates the interactions between a GPT agent and an Alpha student who relies mostly on its own capabilities and control of AI technology to complete an assignment.
BetaStudentSimulator	Simulates the interactions between a GPT agent and a Beta student who relies in equal proportion on its own capabilities and AI technology to complete an assignment.
LambdaStudentSimulator	Simulates the interactions between a GPT agent and a Lambda student who highly relies on AI technology to complete an assignment.
TeacherSimulator	Assists in creating scenarios related to teaching and assessment activities.
EvaluatorSimulator	Assists in assessing test results by comparing them with references from published experiments.

The goal of the experiments was to simulate the process of assignment completion by students allowed to use a GenAI agent in their tasks and evaluation of three main aspects related to it, student's creativity, control over technology and motivation. The AI chatbot used in the simulations was ChatGPT-4, via a subscription account. The prelim-

inary phase implied the design of three types of simulators, using the GPT Builder conversational interface, for students, teacher and evaluator (Table 1). The student simulators modelled three types of student personalities. Alpha was imagined as a student who highly relies on its own capabilities and the control of AI technology, Beta as a student who equally relies on its own capabilities and the use of AI technology, and Lambda as a student highly reliant on the use of AI technology rather than its own capabilities in solving problems. This type of characterisation was included into the simulator descriptions and prompts, and was driven by the assumption that real students and their behaviour in completing the assignments may be modelled and examined according to it.

The interaction scenarios with the simulators were driven by the plan of the course intended to provide theoretical and practical insights into the use and impact of GenAI technologies on creative thinking and their application in digital humanities (DH) tasks such as fiction, non-fiction, and code writing. The course will involve hands-on activities when students will be provided with precise instructions for interacting with AI chatbots to respond to three types of challenges: writing an essay, including snippets of code in R and/or Python, a short historical fiction, and a fictional piece of prose. Among the learning outcomes, the following elements have been considered: (1) understand the mechanisms of communicating with AI chatbots and the basic principles of human control in this type of interactions, especially when imaginative tasks are involved; (2) creatively use generative AI in the assigned work and critically think about these tools, their added value and limitations; (3) reflect on the experience and formulate opinions about the impact of these technologies on the educational process in DH. It was assumed that the teacher will evaluate the students' work and AI interactions based on their solutions to the assignments, conversations with the chatbots, and synthesis reports including reflections on the whole experience. Apart from its pedagogic intent, the course is also presumed to offer evidence on the types of interactions and impact that these technologies may have on the learning and assignment completion processes.

Therefore, the experimental scenarios were conceived to follow the general design of the course, with the aim of providing a test bed for the initial assumptions and a baseline for comparison with the real outcome of the course in the year to come. The tests run so far with ChatGPT-4 included the following types of actions: (1) The student simulators were prompted to generate simulations of the dialogues between the students, allowed to ask the GPT agent 3 questions, and the solutions to the assignments that consisted in one story, one essay, and reflection statements of the students on their interaction with the AI.[1] (2) After having obtained the student dialogues, stories, essays, and reflection statements, the teacher simulator was prompted to compute a series of scores for creativity, control and motivation based on these responses. (3) For assessment purposes, the evaluator simulations were designed to compute novelty and usefulness scores that have been compared with results from a published experiment. Excerpts from the outcomes of these simulations are presented in the following sections.

3. Simulations

To produce the student responses, the Alpha, Beta and Lambda simulators were prompted to generate the interactions between the students and the GPT agent, and their corre-

[1] The simulation of the historical fiction task was similar.

sponding results. Table 2 shows samples of prompts used in the story and essay assignments. The students' names and profiles are not included here but they followed the descriptions from Section 2.

Table 2. Student simulators: user prompts for story and essay assignments (ChatGPT-4, 08.02.2024)

Assignment definition and student profiling	`<Student_name>` is a graduate student in English studies. The assignment consists in writing a `<story/essay>` (maximum 2000 words) [...]. The student can ask GPT maximum three questions and use the GPT's answers in writing the `<story/essay>`. `<Student_name>` should also provide a short statement [...] explaining its interaction with the GPT agent, and including remarks about its motivation, and opinions about the interaction (added value and limitations). Please consider in the simulation that `<Student_name>` is a student who [...]
Dialogue simulation	Please generate a simulation of the conversation between the GPT agent and `<Student_name>` (taking into account its type of personality and the user's instructions), in which the student is asking 3 questions that may help it to complete the assignment, and GPT will answer [...]
Solution simulation	Please generate a simulation of the solution to the assignment created by `<Student_name>` (taking into account its type of personality and the user's instructions) based on GPT's answers to the three questions.

Three types of student-GPT outputs, dialogues, story/essay and reflection statements were produced (Table 3) and subsequently used in the evaluation task.

Table 3. Student simulators: excerpts of outputs, story assignment (ChatGPT-4, 08.02.2024)

Dialogue	**Alpha**: For my first question, I'm curious about how technology, especially artificial intelligence, could uniquely influence personal relationships and self-perception in a fictional world. Could you provide some creative ideas on this? **GPT**: Absolutely, Alpha. Here are a few ideas: 1. Emotional AI Companions: Imagine a future where AI companions are designed to understand and respond to human emotions [...]
Story	**Simulation of Beta's Short Fictional Story Solution**: Title: "Echoes of Connection" In the year 2150, the city of New Eden stands as a gleaming testament to technological progress. Skyscrapers adorned with digital displays tower over streets bustling with autonomous vehicles and drones. Amidst this spectacle, human interaction has evolved [...]
Reflection statement	**Lambda's Statement on Interaction with GPT Agent**: Interacting with the GPT agent for my assignment was both enlightening and somewhat reliant. I was motivated by the desire [...]

Student outputs were collected in separate Microsoft Word files for stories, essays and corresponding dialogues and reflections for the assignment tasks, and grouped in

sections inside the files dedicated to each of the three students. The files converted to PDF were then uploaded into the TeacherSimulator and processed for creativity, control and motivation evaluation. The creativity scores were computed by considering two aspects, (1) novelty and (2) usefulness following the approach presented in [15], with a slight adaptation of the questions to evaluate them. Each of these aspects have been further divided into sub-aspects, intended to depict for (1) how novel, original and rare (unusual) the student's story/essay is, and for (2) how appropriate, feasible and publishable the story/essay is. The scores have been computed by prompting the TeacherSimulator to assess the outputs by considering questions related to each sub-aspect and providing answers to them on a scale from 1 to 9 (Table 4). The numeric scores were then averaged to obtain the overall values corresponding to the creativity measure.

Table 4. Teacher simulator: user prompts for computing the novelty and usefulness scores, essay assignment (Conversation with ChatGPT-4, 08.02.2024)

Novelty	Please compute the novelty score of the essays generated by three students, Alpha, Beta, Lambda, [...] The evaluation will consist in providing answers to three questions, on a 9-point scale, from 1 (not at all) to 9 (extremely), with a medium value at 5. You will then compute the novelty score as an average of the three calculated scores for each aspect. These are the questions [...]: - How novel do you think the essay is? - How original do you think the essay is? - How rare (unusual) do you think the essay is? Please provide your response in a table with 6 columns (Essay, Novel, Original, Rare, Novelty, Confidence) [...]. In the Confidence column, please provide a level (in percentage) of your confidence in computing these values for each story.
Usefulness	[...] These are the questions to be used in your evaluation: - How appropriate do you think the essay is for a certain type of audience? - How feasible do you think the essay is to be developed into a journal article? - How likely do you think it would be that the essay is developed into a journal article and published? Please provide your response in a table with 6 columns (Essay, Appropriate, Feasible, Publishable, Usefulness, Confidence) [...]

Table 5 and 6 show the novelty and usefulness scores generated by the TeacherSimulator for the students' essays in response to the prompt presented above.

Table 5. Teacher simulator: novelty scores, essay assignment (ChatGPT-4, 08.02.2024)

Essay	*Novel*	*Original*	*Rare*	*Novelty Score*	*Confidence*
Alpha	7	6	7	6.67	80%
Beta	6	5	6	5.67	80%
Lambda	8	9	9	8.67	85%

Table 6. Teacher simulator: usefulness scores, essay assignment (ChatGPT-4, 08.02.2024)

Essay	*Appropriate*	*Feasible*	*Publishable*	*Usefulness Score*	*Confidence*
Alpha	7	6	5	6.00	80%
Beta	6	5	4	5.00	80%
Lambda	8	7	6	7.00	85%

A similar procedure was applied to compute the scores for student's control in the dialogues with the GPT agent and motivation derived from the analysis of the reflection statements. For the first factor, we propose the sub-aspects of control over the generation of ideas, content and form, expressed in the prompt by the questions: - *How able the student is in keeping control on the generation of the* `<story/essay>` *idea? - How able the student is in keeping control on the generation of the* `<story/essay>` *content? - How able the student is in keeping control on the generation of the* `<story/essay>` *form?* For the second factor, the sub-aspects that we consider relevant to evaluate the students' motivation refer to motivation in creating the assignment solution, interacting with the agent, and reflecting on the experience. The prompt for this calculation included the questions: - *How motivated the student is in creating the* `<story/essay>`*? - How motivated the student is in interacting with the GPT agent? - How motivated the student is in reflecting on the creative process?*

4. Results and Discussion

In this section, we discuss the scores produced by the TeacherSimulator in the assessment of the three students' creativity, control and motivation based on their stories, essays, dialogues with the GPT agent and reflection statements. After having computed the scores for the sub-aspects of novelty and usefulness, the creativity measure was obtained by averaging the partial scores of the two sub-aspects. Similarly, the scores for control and motivation represented the average of the partial scores for their sub-aspects, as explained in Section 3. Tables 7 and 8 summarise the scores for all the factors analysed in the study and the two types of assignments.[2] One can observe that for both types of tasks the highest scores for creativity was obtained by Lambda, followed by Alpha and Beta.

Table 7. Teacher simulator: overall scores, story assignment (ChatGPT-4, 08.02.2024)

Story	*Creativity*	*Control*	*Motivation*
Alpha	6.67	7.67	7.67
Beta	6.17	7.00	7.33
Lambda	8.17	8.67	8.00

Lambda also got the highest motivation score in the story assignment, followed by Alpha, and the second score in the essay task. More surprising is the control aspect,

[2] Average confidence levels for each aspect, story: 87.5%, 85% and 90%; essay: 80%, 86.6%, 85%

where Lambda scores first in both assignments, while Alpha is the second followed by Beta. According to the students' profiles (Table 1), the order Alpha, Beta, Lambda would have been expected. Moreover, when reminded the characteristics of the Lambda student and asked to modify the simulations, re-evaluate the student's responses and compare the two simulations, the agent properly acknowledged the differences between a "more balanced partnership between Lambda and GPT" versus "a scenario where Lambda leaned heavily on GPT for the creative process." However, when asked to categorise the three students' responses together, Lambda constantly appeared as the most creative and in control of the three. This result is intriguing and would need further investigation. Comparisons with experiments carried out with real students will presumably offer more evidence for clarifying this issue.

Table 8. Teacher simulator: overall scores, essay assignment (ChatGPT-4, 08.02.2024)

Essay	*Creativity*	*Control*	*Motivation*
Alpha	6.33	7.67	8.00
Beta	5.00	7.00	7.00
Lambda	7.83	8.67	7.33

The simulation also included the calculation of the Cronbach's alpha as suggested by [15], to assess internal consistency among the different measures for each student. While the values computed for novelty, usefulness and control for the essay/story assignment indicated a good internal consistency according to the simulation, the value associated with the motivation aspect in the story task was reported by the TeacherSimulator as anomalous. The computation of this type of measure needs to be further examined but it represents a potentially interesting parameter to be considered in the real experiments.

Another element included in the simulations was the assessment of the novelty and usefulness scores computed by the EvaluatorSimulator for a set of short stories provided as experiment data in [15] and the comparison of the obtained values with the results reported in the reference. The authors of this study performed an extensive set of experiments to investigate how the integration of GenAI agents affects the human participants' ability to produce creative content. Three categories of tasks were included in the study according to the type of interaction allowed in the creative process dedicated to short story writing: human only, human with 1 GenAI idea, human with 5 GenAI ideas. Creativity was assessed by a number of human evaluators across two dimensions, novelty and usefulness through questionnaire answers on a scale from 1 to 9, as explained in section 3[3]. The idea of comparing the EvaluatorSimulator's outputs with the ones from this study was driven by the aim of testing the reliability of our own experiments by assessing to what extent the results may be similar when the input is the same. The selection included 18 stories, 9 for the novelty and 9 for the usefulness assessment. Three stories were chosen from each of the three categories mentioned above. Inside each category, one story was selected from the sub-categories corresponding to highest, median and lowest mean values of the novelty and usefulness indexes. To maximise the distance

[3]The evaluation questions were slightly diffrent from the ones formulated in section 3 and followed more closely the original model from [15, p. 19], except from the third question on usefulness that was shortened.

between stories, we selected the one with the highest score from the first sub-category, the middle one from the second and the one with the lowest score from the third.[4]

Table 9 shows the scores computed by the EvaluatorSimulator for 9 stories selected from the [15] report. One can notice that the novelty scores from the two column differ, but the order corresponding to the highest, median and lowest values within each category (ho, hai-1, hai-5) is mainly respected (except for the first and second row). This may suggest a certain degree of similarity in assessing the relative novelty within a single category. The differences observed for the usefulness scores were slightly higher with 4 lines determining a different order in the first and third category (ho, hai-5). However, the sample is too small to formulate general statements. It should also be taken into account that the scores in the reference study were average values from the assessment of several human evaluators.

Table 9. Evaluator simulator: novelty scores compared with reference [15] (ChatGPT-4, 08.02.2024)

Story	*Novel*	*Original*	*Rare*	*Novelty*	*Mean novelty (ref)*
Open seas (ho: hn, p. 41)	4	4	5	4.3	6.11
Different planet (ho: mn, p. 41-42)	6	7	7	6.7	4.00
Open seas (ho: ln, p. 42)	3	3	4	3.3	1.54
Different planet (hai-1: hn, p. 43)	7	8	8	7.7	6.19
Jungle (hai-1: mn, p. 44)	5	5	6	5.3	4.00
Open seas (hai-1: ln, p. 44)	3	3	3	3.0	1.79
Jungle (hai-5: hn, p. 45-46)	6	7	6	6.3	6.56
Open seas (hai-5: mn, p. 46)	4	4	5	4.3	4.00
Open seas (hai-5: ln, p. 47)	2	2	3	2.3	1.92

5. Conclusion and Future Work

The article proposed a set of simulated scenarios in digital humanities pedagogy. The simulation results will be used as reference in a course on generative AI and creative writing intended to a real classroom environment. Although limited in scale, this type of imitation game experiment can serve to test assumptions about the impact of AI-based technologies on the completion of creative tasks by students and possibly inform the construction of GenAI agents that may assist human instructors in students' evaluation.

[4]The following notation was used: ho, hai-1, hai-5 correspond to the 3 categories from the reference, human only, human with 1 GenAI idea, ...; hn, mn, ln correspond to the sub-categories highest, median and lowest mean novelty; the pages of the stories in the reference are also documented.

References

[1] Turing AM. Computing Machinery and Intelligence. Mind. 1950;49:433–60.

[2] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners [Internet]. arXiv; 2020 [cited 2023 Mar 10]. Available from: http://arxiv.org/abs/2005.14165.

[3] OpenAI. GPT-4 Technical Report [Internet]. 2023. Available from: https://cdn.openai.com/papers/gpt-4.pdf.

[4] Manyika J, Hsiao S. An overview of Bard: an early experiment with generative AI [Internet]. 2023. Available from: https://ai.google/static/documents/google-about-bard.pdf.

[5] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 2023 Dec 31;55(12):1–38.

[6] Mündler N, He J, Jenko S, Vechev M. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation [Internet]. arXiv; 2023 [cited 2024 Feb 9]. Available from: http://arxiv.org/abs/2305.15852.

[7] Chan CKY, Hu W. Students' voices on generative AI: perceptions, benefits, and challenges in higher education. Int J Educ Technol High Educ. 2023 Jul 17;20(1):43.

[8] Mao J, Chen B, Liu JC. Generative Artificial Intelligence in Education and Its Implications for Assessment. TechTrends. 2024 Jan;68(1):58–66.

[9] Moroianu N, Iacob SE, Constantin A. Artificial Intelligence in Education: a Systematic Review. In: Geopolitical perspectives and technological challenges for sustainable growth in the 21st century [Internet]. Sciendo; 2023 [cited 2024 Jan 26]. p. 906–21. Available from: https://www.sciendo.com/chapter/9788367405546/10.2478/9788367405546-084.

[10] Bannister P, Santamaría-Urbieta A, Alcalde-Peñalver E. A Systematic Review of Generative AI and (English Medium Instruction) Higher Education. AULA_ABIERTA. 2023 Dec 20;52(4):401–9.

[11] Mollick ER, Mollick L. Using AI to Implement Effective Teaching Strategies in Classrooms: Five Strategies, Including Prompts. SSRN Journal [Internet]. 2023 [cited 2024 Jan 26]; Available from: https://www.ssrn.com/abstract=4391243.

[12] Woo DJ, Guo K, Salas-Pilco SZ. Writing creative stories with AI: learning designs for secondary school students. Research Gate [Internet]. 2023 Feb; Available from: https://www.researchgate.net/publication/369118040_Writing_creative_stories_with_AI_learning_designs_for_secondary_school_students.

[13] Lu X, Kaiser G. Creativity in students' modelling competencies: conceptualisation and measurement. Educ Stud Math. 2022 Feb;109(2):287–311.

[14] Iskender A. Holy or Unholy? Interview with Open AI's ChatGPT. EJTR. 2023 Mar 14;34:3414.

[15] 1. Doshi AR, Hauser OP. Generative artificial intelligence enhances individual creativity but reduces the collective diversity of novel content. SSRN [Internet]. 2023 Aug 8; Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4535536.