The Pure Poet: How Good is the Subjective Credibility and Stylistic Quality of Literary Short Texts Written with an Artificial Intelligence Tool as Compared to Texts Written by Human Authors?

Vivian Emily Gunser[1], Steffen Gottschling[1], Birgit Brucker[1], Sandra Richter[2], Dîlan Canan Çakir[2], and Peter Gerjets[1]

[1] Leibniz-Institut fuer Wissensmedien, Schleichstr. 6, 72076 Tuebingen, Germany
{v.gunser, s.gottschling, b.brucker, p.gerjets}@iwm-tuebingen.de
[2] Deutsches Literaturarchiv Marbach, Schillerhoehe 8-10, 71672 Marbach am Neckar, Germany
{sandra.richter, dilan.cakir,}@dla-marbach.de

Abstract

The application of artificial intelligence (AI) for text generation in creative domains raises questions regarding the credibility of AI-generated content. In two studies, we explored if readers can differentiate between AI-based and human-written texts (generated based on the first line of texts and poems of classic authors) and how the stylistic qualities of these texts are rated. Participants read 9 AI-based continuations and either 9 human-written continuations (Study 1, *N*=120) or 9 original continuations (Study 2, *N*=302). Participants' task was to decide whether a continuation was written with an AI-tool or not, to indicate their confidence in each decision, and to assess the stylistic text quality. Results showed that participants generally had low accuracy for differentiating between text types but were overconfident in their decisions. Regarding the assessment of stylistic quality, AI-continuations were perceived as less well-written, inspiring, fascinating, interesting, and aesthetic than both human-written and original continuations.

Keywords: Cognition, Artificial Intelligence, Literature, NLP, GPT-2

Introduction

Artificial intelligence is increasingly used to provide support in creative domains such as the composition of emotional film trailers (Smith et al., 2017) or the ideation in fashion design (Jeon et al., 2021). As part of this trend, advanced tools for human-AI co-creative processes have been developed in recent years. For instance, in a visual arts context, an empathic AI-tool has been developed that provides help in portrait drawing by means of embodied conversational interaction (Yalçın, Abukhodair & DiPaola, 2020). Another example from the field of music composition is an AI-tool enabling computational melodic harmonization (CHAMELEON) that has been developed by Zacharakis et al. (2021). When evaluating this tool with experienced and inexperienced music composers engaging in human-AI co-creative processes it turned out that this tool was particularly helpful for less experienced students to better express their ideas.

In this paper we will focus on using AI-tools in an even more complex creative domain, namely the production of literary texts such as short stories or poems. This domain can be seen as providing harder challenges than music composition or drawing due to the complexity of its underlying semantic structure and the embodied grounding of the symbols used to express it. Creativity tools in fields such as music or visual arts of course also need to pick up relevant patterns in their respective domains but they would not have to "understand" the symbolic meaning of these patterns in order to be able to play with them in a creative way and to produce novel patterns that would make sense to human recipients. Literary fiction, on the contrary, is based on playing with semantic and formal structures that are generated and understood based on their embodied groundings in the perceptions, feelings and actions of human authors and human readers. These groundings of language meanings in perceptions, feelings and actions are obviously not available to AI-tools so that they would need to navigate a semantic space during text production without really "knowing" what they are writing about. Therefore, it is an important question to investigate how believable or credible literary texts written with the help of AI-tools can be and how they would be perceived aesthetically.

When it comes to AI-based text generation in general, the situation is probably easier for expository texts than for literary texts as they usually describe facts in the external world that can be collected in fact databases for grounding purposes (which is not the case for literary texts describing "inner" facts emanating from the mental life of an author). Accordingly, for expository text generation there are already some successful examples providing evidence that AI-tools can use big databases of facts to automatically produce credible expository texts. Accordingly, based on recent advances in natural language processing (NLP), more and more AI-generated text sources have become available online and more and more applications for creating such texts continue to be developed and refined. For instance, in a study by Graefe et al. (2018) 986 participants read computer-written news articles about sports and financial topics and rated these texts even as more credible and higher in journalistic expertise (but more difficult with regard to readability) than comparable human-written articles from popular German websites for sports (i.e., sport1.de) and financial topics. For the computer-written news articles in this study, an application for natural language generation was used that allowed for the automatic creation of ready-to-publish expository texts based on large databases for different topics such as soccer games, stock exchange market reports,

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

or weather forecasts (Haarmann & Sikorski, 2015). Beyond descriptive expository texts, AI-models have also been developed for writing tasks in conversational contexts such as writing emails (Buschek, Zürn & Eiband, 2021), using bots to communicate via text chat with customers (McKee & Porter, 2020) or providing health consultation (Wang et al. 2021).

In all of the use-cases mentioned above it is of course not only important from a cognitive science perspective to understand how and how well human readers are (still) able to distinguish AI-generated from human-generated texts but also to analyze how AI-based texts are perceived and evaluated. However, these questions might be most interesting in the use-case of writing literary texts due to the important role of an embodied grounding of these texts in the perceptions, feelings and actions of human authors and human readers. Accordingly, one might doubt whether AI-tools will be able to write about these types of experiences - experiences that they are themselves not capable of having - in a way that is perceived as credible and aesthetically appealing to human readers. Empirically, findings about the perception and evaluation of AI-based non-expository texts are quite mixed, depending on the concrete scenario and technology used. For instance, Bringsjord and Ferrucci (1999) reported about BRUTUS, a story telling machine, that it was not yet capable of producing full stories to compete with human writers on creativity ratings. Clark et al., (2018) on the other hand reported that participants found it helpful and fun to use an AI-system to give them suggestions and ideas for writing short stories and slogans.

Moreover, for the automatic generation of natural language to write poetry based on images, even a Turing Test was passed (Liu et al., 2018). This Test (Turing, 1950) is a standard procedure commonly used to examine whether a computer-generated content or behavior can be identified as such by humans when compared to a human-generated content or behavior. In the study of Liu et al. the Turing Test was conducted by asking literature experts and literature novices to choose the human-written poems from a set of mixed human-written and AI-generated poems based on images. All participants ended up with high confusion rates (40-57 %) with the expert group being slightly superior to the novice group at identifying the AI-generated poems.

Similar results were obtained recently by Köbis and Mossink (2021) who used GPT-2 (Generative Pretrained Transformer 2 Model) in their Turing-Test study, which demonstrated that literature novices could not differentiate AI-generated poetry from purely human-written poetry (written by untrained writers in their first study). The correct origin of poems was identified by participants with an average accuracy of 50.21%, indicating no significant deviation from chance level. In this study they additionally asked participants, before reading all poems, how confident they were that they could distinguish between the AI-generated and human-written poems.

Results indicated that participants were rather overconfident (69.33%) compared to their actual performance (50.21%). Thus, participants were quite convinced that they could distinguish between AI-generated and human-written texts, but they were actually not able to do so. Moreover, participants' performance beliefs were not significantly correlated with their actual performance in detecting the correct origin of the texts in a regression analysis. Despite their random performance, participants nevertheless showed a preference for human-written poetry.

In a second study, Köbis and Mossink used professional poems of Maya Angelou and Hermann Hesse and compared them to AI-generated poems. Participants had to identify the correct origin of poems in two different conditions: In the human-in-the-loop (HITL) condition the best GPT-2 poems were preselected for presentation by human raters whereas in the human-out-of-the-loop (HOTL) condition randomly sampled GPT-2 generated poems were presented. The results showed, in sum, that overall, participants were able to detect the correct origin of the poems better than chance levels. In the HOTL condition accuracy levels of the participants were higher than in the HITL condition. The accuracy rates in the HOTL condition differed significantly from chance level whereas the accuracy rates in the HITL condition did not differ significantly from chance level. Again participants had to indicate their confidence after reading each text. In this study 38.91% of the participants showed hints for overconfidence. A linear regression revealed for this second study that participants' performance beliefs and their accuracy levels in detecting AI texts correlated significantly (positive). Again, participants preferred overall the original human-written poems of Maya Angelou and Hermann Hesse to AI-generated poems. In sum, people preferred human-written texts generated by both untrained writers as well as classic authors. For human-written texts from untrained writers Köbis and Mossink found a confusion rate of 50%, whereas for original texts from professional authors the confusion rates were a little bit lower. With regard to their abilities to distinguish between human-written and AI-generated poems, participants were overconfident in the first, but not in the second study.

It has to be noted that Köbis and Mossink did not investigate the perception and evaluation of AI-generated poems in greater detail, for instance with regard to the evaluation of their stylistic quality as an indicator of the aesthetic perception of poems. Moreover, they confined their studies to poems from only two classic authors, both from the 20th century. In our own studies we addressed these issues by first investigating not only the identification of AI poetry but also its aesthetic perception in terms of how participants evaluated the stylistic quality of all texts. Moreover, to generalize the findings of Koebis and Mossink we extended the set of classic authors investigated from two to four authors from different historical epochs. For each author, several unfamiliar texts (narrative texts or poems) were selected as stimulus materials. The final text set used in our studies comprised 18 poems and narrative texts written by different classic authors (i.e., Franz Kafka, Friedrich Hölderlin, Robert Gernhardt, and Paul Celan) as materials.

Furthermore, we did not only compare the original texts from the classic authors (for which one cannot ensure that they are completely unfamiliar to participants) with AI-based texts, but additionally let authors with a literature-specific professional background (instead of untrained writers) create purely human-written plausible text continuations for the first few lines of each text as comparison materials.

The main research questions addressed in our two pre-registered studies are: Do people recognize and aesthetically prefer narrative texts and poems written by (a) participants with a literature-specific professional background (Study 1) or (b) original narrative tests and poems written by classic authors (Study 2) as compared to texts generated by a GPT2-based AI-writing-tool? Therefore, we investigate how readers evaluate the different types of texts regarding their stylistic qualities.

We also investigated how accurate readers can distinguish between AI-based text and texts written purely by humans and how confident the readers were in their classification decisions. Besides low levels of accuracy, high levels of confidence in wrong decisions might be a good indicator for the credibility of AI-generated texts.

Study 1

Methods

Participants Study 1 A sample of $N = 120$ participants who were fluent in German was recruited via the online platform Prolific. The sample consisted of 64 females, 53 males and 3 non-binary persons. Participants' age ranged from 18 to 61 years ($M = 26.76$, $SD = 7.20$).

Materials and Procedure The texts used in this study were generated in an exploratory pilot study (Gunser et al., 2021).

In a writing phase of the pilot study, the first few lines of poems were presented to the participants with literature-related professional background (18 unfamiliar poems; two different poems for each participant). The original texts were authored by the writers Franz Kafka, Friedrich Hölderlin, Robert Gernhardt, or Paul Celan. Participants were asked to write their own continuation for the presented lines and then attempt to develop a second continuation to the same lines using an AI (artificial intelligence)-writing-tool based on GPT-2. As a result, in addition to the original continuation of the poem or text two alternative continuations were created: One continuation written by a human (participant with literature background) and another continuation written with the help of the AI(artificial intelligence)-tool (HITL). In the AI-based continuation, human editing was severely limited to 25% of the words.

In the present Study 1, the 18 human-written continuations and the 18 AI-based continuations were presented in an evaluation phase to the participants. A within-subject design was used: Each participant evaluated nine human-written and nine AI-based continuations. The continuations were presented together with the first lines (i.e., the beginning) of the original texts and participants were informed (via color-coding) where the original beginning ended and the continuations started. Which of the 18 texts was presented in the human-written or AI-based version was counterbalanced across participants. The proportion of AI-based and human-written texts was undisclosed to ensure that decisions could not be derived based on previous trials.

For the first evaluation phase, participants were asked to answer the statement "This text was written with the help of an artificial intelligence (AI)" with "yes" or "no" and then "How confident are you in your decision?" with an indication on a six-point-scale from "50% - I guessed" to "100% - very sure". After the first evaluation phase all continuations were presented again and participants were asked to rate on a scale from "1 - I strongly disagree" to "5 - I strongly agree" if each individual continuation was well-written, inspiring, fascinating, interesting, and aesthetic based on their perception.

Results

Classification accuracy (AI-based vs. human-written) Participants identified 59.72% of the human-written continuations correctly as human-written and 57.96% of the AI-based continuations correctly as AI-based. This means, they misclassified 40.28% of human-written continuations falsely as AI-based and 42.04% of the AI-based continuations falsely as human-based (see Table 1). This indicates that participants were not able to perfectly distinguish between texts (poems) produced by humans and texts produced by an AI-tool in Study 1. The one sample t-test was significant ($mu = 1$), $t(119) = -33.83$, $p < .001$, Cohen's $d = –3.09$. Nonetheless, participants were able to distinguish between AI and human continuations above chance level ($mu = 0.5$), $t(119) = 7.27$, $p < .001$, Cohen's $d = 0.66$. On a descriptive level, AI-texts were assumed to be human-texts (42.04%) more often than human-texts were assumed to be AI-texts (40.28%). However, a paired t-test, $t(119) = 0.82$, $p = .412$, was not significant.

Table 1: Classification table for participants differentiation between AI-based and human-written text continuations.

Participants' classification	Actual continuation type	
	Human-written	AI-based
Human-written	59.72%	42.04%
AI-based	40.28%	57.96%

Confidence Measure In Study 1 participants showed no significant difference in their confidence while classifying AI-based ($M = 0.76$, $SD = 0.15$) compared to human-written continuations ($M = 0.77$, $SD = 0.16$), $t(119) = 0.66$, $p = .509$. An exploratory simple regression with normalized classification accuracy (averaged across all texts for each person) as the predictor and average confidence as the

outcome, showed no significant prediction of confidence via accuracy, β = 0.013, p = .088, R^2 = .016 (see Figure 1).

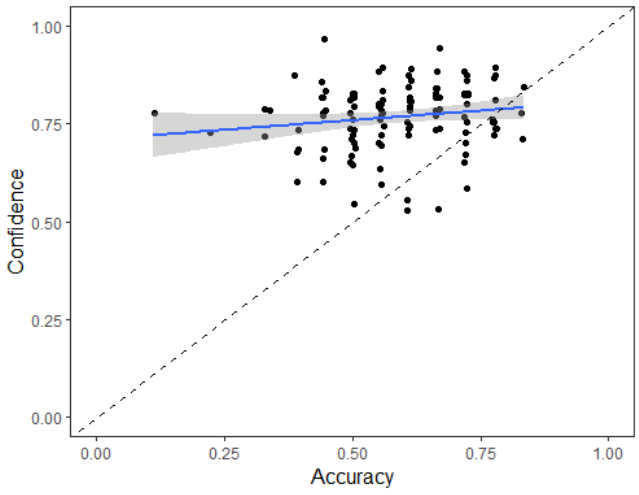

Figure 1: Scatterplot (with regression line) for the association between accuracy and confidence (both averaged across texts) for the classification task in Study 1. The dashed line indicates a perfect match of accuracy and confidence. Points above the line indicate overconfidence while points below the line indicate underconfidence.

Stylistic Quality For the analysis regarding stylistic quality, participants ratings were averaged across all continuations of the same type (AI-based vs. human-written) for each participant to compare the resulting scores in paired t-Tests. Human-written continuations were perceived as more well-written, $t(119) = 7.40$, $p < .001$, Cohen's $d = 0.68$, more inspiring, $t(119) = 4.25$, $p < .001$, Cohen's $d = 0.39$, more fascinating, $t(119) = 5.19$, $p < .001$, Cohen's $d = 0.47$, more interesting, $t(119) = 3.93$, $p < .001$, Cohen's $d = 0.36$, and more aesthetic, $t(119) = 8.01$, $p < .001$, Cohen's $d = 0.73$, than AI-based continuations (see Figure 2).

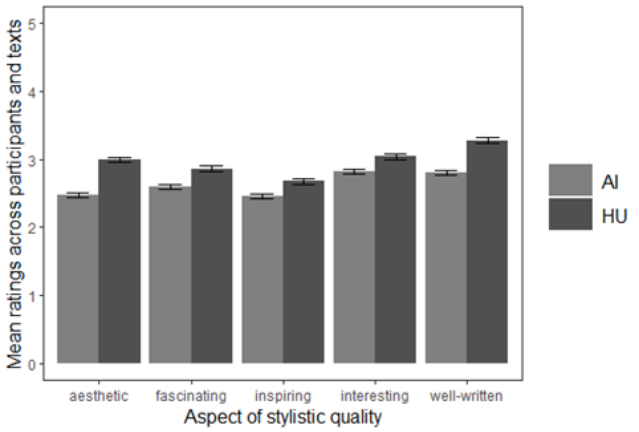

Figure 2: Participants' subjective stylistic quality evaluation of AI-based (AI) and original (OG) texts in Study 1. Error bars indicate standard errors.

Study 2

Methods

Participants Study 2 A sample of N = 302 participants who were fluent in German was recruited via the online platform Prolific. Participants of Study 1 were excluded from participation in Study 2. The sample consisted of 166 females, 129 males and 7 non-binary persons. Their age ranged from 18 to 66 years (M = 27.01, SD = 6.30).

Materials and Procedure The AI-based text continuations were the same as in Study 1. In Study 2 the 18 texts with the AI-based continuation and the 18 original poems of the classic authors were presented to the participants. Apart from this change the materials and procedure were identical to Study 1.

Results

Classification accuracy (AI-based vs. original) Participants identified 66,48% of the original continuations correctly as original and 59,78% of the AI-based continuations correctly as AI-based. This means, they misclassified 33.52% of original continuations falsely as AI-based and 40.22% of the AI-based continuations falsely as original (see Table 2).

Table 2: Classification table for participants differentiation between AI-based and original text continuations.

Participants' classification	Actual continuation type	
	Original	AI-based
Original	66.48%	40.22%
AI-based	33.52%	59.78%

Participants, again, were not able to perfectly distinguish between original (produced by humans) and AI-based continuations (mu = 1), $t(301) = -44.67$, $p < .001$, Cohen's $d = -2.57$. Nonetheless, participants' classification between AI-based and original continuation was still significantly above chance (mu = 0.5), $t(301) = 15.92$, $p < .001$, Cohen's $d = 0.92$.

Moreover, in Study 2 not only numerically, but also statistically AI-based continuations were assumed to be original continuations more often than original continuations were assumed to be AI-continuations, $t(301) = 4.89$, $p < .001$, Cohen's $d = 0.28$.

Confidence Measure In Study 2 the average confidence level of participants for the classification was lower for AI-based continuations (M = 0.74, SD = 0.15) than for the original continuations (M = 0.75, SD = 0.16), $t(301) = -2.92$, $p = .004$, Cohen's $d = -0.17$. An exploratory simple regression with normalized classification accuracy (averaged across all texts for each person) as the predictor and average confidence

as the outcome, showed a significant prediction of confidence via accuracy, β = 0.022, $p < .001$. However, only a small amount of variance in the confidence measure was explained by the actual accuracy, $R^2 = .061$ (see Figure 3).

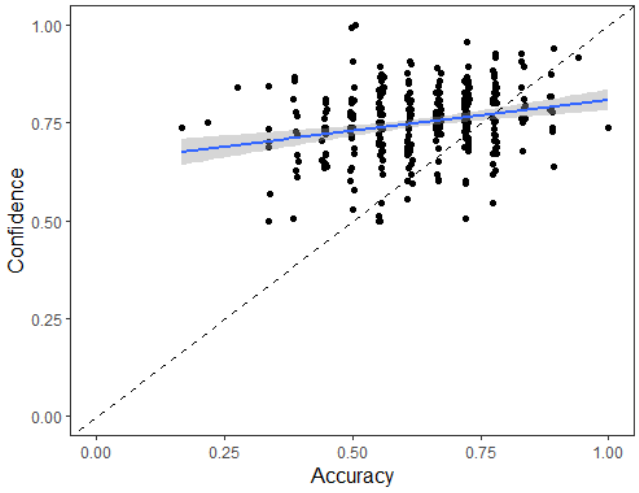

Figure 3: Scatterplot (with regression line) for the association between accuracy and confidence (both averaged across texts) in the classification task in Study 2. The dashed line indicates a perfect match of accuracy and confidence. Points above the line indicate overconfidence while points below the line indicate underconfidence.

Stylistic Quality For the stylistic quality evaluation of AI-based and original (human-written) continuations, we found the same result pattern as in Study 1. Original continuations were perceived as more well-written, $t(301) = 9.65$, $p < .001$, Cohen's $d = 0.56$, more inspiring, $t(301) = 4.21$, $p < .001$, Cohen's $d = 0.24$, more fascinating, $t(301) = 4.64$, $p < .001$, Cohen's $d = 0.27$, more interesting, $t(301) = 5.53$, Cohen' $d = 0.32$, $p < .001$, and more aesthetic, $t(301) = 9.19$, $p < .001$, Cohen's $d = 0.53$, than AI-based continuations (see Figure 4).

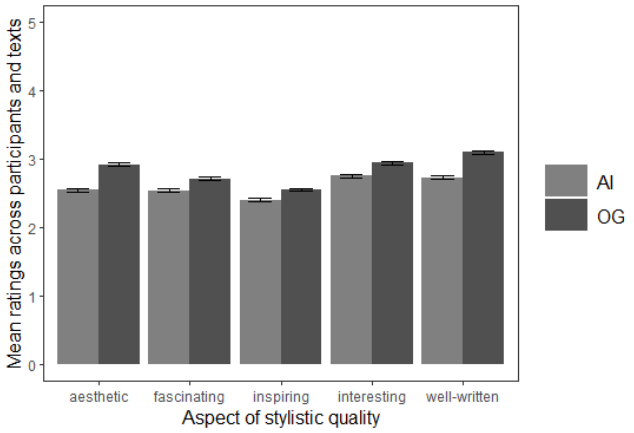

Figure 4: Participants' subjective stylistic quality evaluation of AI-based (AI) and original (OG) texts in Study 2. Error bars indicate standard errors.

Discussion

The first goal of the presented studies was to further investigate to what extent readers are capable of distinguishing between AI-generated texts and poems and texts and poems written by (a) participants with a literature-specific professional background or (b) original narrative texts and poems written by classic authors.

Our results showed that participants felt rather confident in their classification between human-written (both texts written by literature professionals as well as originals written by classic authors) and AI-based texts, but in fact the misclassification rate for both types of texts (human-written and AI-based) was rather high. This indicates that participants were overconfident even though they were not able (actually far away from being able) to perfectly distinguish between the two text types. These findings are partly in line with the findings of Köbis and Mossink (2021) who showed hints of overconfidence in combination with high misclassification rates in their first study but no sign of systematic overconfidence in their second study. Moreover, we also found similar results as Köbis and Mossink regarding the prediction of participants' confidence by the accuracy of participants' decision: Study 1 (AI-based vs. written by literature professionals) accuracy did not predict confidence, whereas in Study 2 (AI-based vs. originals), we found a significant prediction of confidence by accuracy over all the texts. This prediction, however, only explained a small amount of the variance (6%). Further analyses based on the signal detection paradigm used in our study are currently under consideration to gather additional insights on metacognitive levels of participants' confidence.

Compared to already existing research in the field of literature, we examined highly structured and also historical texts of different classic authors. This could provide some explanations on why the classification accuracy in our studies was somewhat higher as in comparable literature: Hölderlin's poems, for example, are written in a sublime style and based on the German language of the 18th-century, like other poems in our corpus. These features could be an indication used by readers to identify human-generated texts. Furthermore, Gernhardt's poems profit from a punchline that presupposes situational knowledge. The AI-based continuations lack such punchlines compared to Robert Gernhardt's original continuations, which makes it even easier for readers to recognize the AI-based texts. Köbis and Mossink (2021), on the other hand, used poems by 20th-century authors Maya Angelou and Herman Hesse that are in contrast for example to Hölderlin, both written in less sublime forms and in contrast for example to Gernhardt do not use punchlines. The lack of these three text features (historical language, sublime style, punchlines) could explain the higher misclassification rate of 50% in Köbis and Mossink's study in comparison to the misclassification rates of around 40% in our studies. Further, Köbis and Mossink used for their studies a GPT-2 model that was specifically trained beforehand on the authors Jane Campion, Roald Dahl, Robert Frost, and William Blake, whereas we decided to use

the basic GPT-2 model in our studies which was not specifically trained beforehand to investigate its possibilities in a co-creation writing scenario to get more generalizable results than deciding to train it with material of specific authors. The pretraining of Köbis and Mossink might have led to higher quality poems written by GPT-2 in their studies, thereby also increasing the possibility of higher misclassification rates.

The second goal of our studies was to gather additional insights in readers' evaluation of stylistic qualities of AI-based compared to texts and poems written by humans (either literature professionals or classic authors). In this regard, previous research (Graefe et al., 2018; Clerwall, 2014; Köbis and Mossink, 2021) showed that AI-generated texts were perceived as boring and were less preferred when compared to human-written texts. Our results corroborate and extend these findings: participants in both of our studies rated texts written by humans (both literature professionals as well as classic authors) as better written, more inspiring, more fascinating, more interesting and more aesthetic than AI-based texts. The fact that medium to strong effects were observed in terms of lower stylistic quality of AI-based texts compared to texts written by humans, further raises the question why readers show relatively low accuracy when differentiating between these text types since one could have assumed that perceived stylistic quality could be used as a potential indicator for correct classification. However, our results speak in the other direction: Despite the fact that participants evaluated the AI-based texts consistently worse than texts written by humans in all five measured dimensions of stylistic quality and across both studies, this did not seem to help participants to achieve better classification rates. Taken together these results lead us to a conclusion beyond the previous literature: the textual input matters because participants may be primed by specific text features.

Based on our findings, we argue that it is even with expert knowledge not easy to write elaborate literature with an AI-writing-tool. Nevertheless, such an AI-writing-tool could be used more in terms of an inspiration for creativity, for example by giving new and potentially unexpected writing prompts. In recent years, such tools that could help people become more creative in writing have increasingly appeared, such as Wordcraft, a human-AI collaborative editor for story writing (Coenen et al., 2021) or CoAuthor as another example presented by Lee, Liang & Yang (2022), a tool for assisting in creative and argumentative writing. In the Human-Computer-Interaction community particularly the opportunities of such tools and large language models are of increasing interest. In our pilot study literature professionals explored the opportunities of the algorithm-based AI-tool to create poems and narrative texts generated in a human-AI co-creative process (e.g., collaborative writing scenario). However, when considering the AI-tools as collaborative writing partners it is important that the use of the AI-tool should constitute an *enhancement* in creativity rather than a *replacement* of the humans in the writing process (cf. human-centered AI; Shneiderman, 2020). It is important to also keep in mind that there are still no legal guidelines who can call themselves the author of such AI-generated literary works.

In addition, our research is important to rank readers' perceptions of the output of AI-tools. This might be helpful to identify and apply optimization possibilities of the respective AI-tools. Certainly, many studies like the one by Köbis and Mossink (2021) or the one presented in this paper are necessary to make a general statement about artificial intelligences that are involved in the process of generating literature.

In the context of the generalizability of AI Oscar Schwartz gives some philosophical thoughts on classic authors in his TED talk: He assumes that computers can write poems, after comparing human-written poems of famous authors and computer-generated poems. Maya Angelou possibly writes more in the style of a computer, perhaps would pass a reverse Turing Test as well. Therefore, we should think about how we define human creativity and whether computers simply reflect our ideas (Schwartz, 2015). Computers are reflecting those ideas without grounding language meanings in perceptions, feelings and actions compared to humans.

If the amount of AI-generated text on the internet increases, the question arises whether future generations will still try to recognize AI-generated content, especially if they have difficulties in doing so (Gunser et al., 2021). Investigating peoples' evaluation of AI-generated content in terms of the stylistic quality of literary short texts is a first promising step that is particularly interesting for future research endeavoring in creative areas such as poetry.

References

Bringsjord, S., & Ferrucci, D. (1999). *Artificial intelligence and literary creativity: Inside the mind of brutus, a storytelling machine.* Mahwah, NJ: Lawrence Erlbaum Associates.

Buschek, D., Zürn, M., & Eiband, M. (2021). The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). New York, NY: Association for Computing Machinery.

Clark, E., Ross, A. S., Tan, C., Ji, Y., & Smith, N. A. (2018). Creative writing with a machine in the loop: Case studies on slogans and stories. *23rd International Conference on Intelligent User Interfaces*, (pp. 329-340). New York, NY: Association for Computing Machinery.

Clerwall, C. (2014). Enter the robot journalist: Users' perceptions of automated content. *Journalism practice*, *8*, 519-531.

Coenen, A., Davis, L., Ippolito, D., Reif, E., & Yuan, A. (2021). Wordcraft: a Human-AI Collaborative Editor for Story Writing. *First Workshop on Bridging Human-Computer Interaction and Natural Language Processing at EACL 2021. Stroudsburg: PA* arXiv:2107.07430.

Graefe, A., Haim, M., Haarmann, B., & Brosius, H. B. (2018). Readers' perception of computer-generated news:

Credibility, expertise, and readability. *Journalism*, 19, 595-610.

Gunser, V. E., Gottschling, S., Brucker, B., Richter, S., & Gerjets, P. (2021, July). Can Users Distinguish Narrative Texts Written by an Artificial Intelligence Writing Tool from Purely Human Text? *International Conference on Human-Computer Interaction* (pp. 520-527). Springer, Cham.

Haarmann, B., & Sikorski, L. (2015). Natural language news generation from big data. *International Journal of Computer and Information Engineering*, 9, 1489-1495.

Jeon, Y., Jin, S., Shih, P.C., & Han, K. (2021). FashionQ: An AI-Driven Creativity Support Tool for Facilitating Ideation in Fashion Design. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, (pp. 1-18). New York, NY: Association for Computing Machinery.

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114, 106553.

Lee, M., Liang, P., & Yang, Q. (2022). CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. *arXiv preprint arXiv:2201.06796.*

Liu, B., Fu, J., Kato, M. P., & Yoshikawa, M. (2018). Beyond narrative description: Generating poetry from images by multi-adversarial training. *Proceedings of the 26th ACM international conference on Multimedia* (pp. 783-791). New York, NY: Association for Computing Machinery.

McKee, H. A., & Porter, J. E. (2020, February). Ethics for AI writing: The importance of rhetorical context. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 110-116). New York, NY: Association for Computing Machinery.

Schwartz, O. (2015). Can a computer write poetry. Retrieved from https://www.ted.com/talks/oscar_schwartz_can_a_computer_write_poetry?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems. *ACM Transactions on Interactive Intelligent Systems*, *10*, (pp. 1-31). New York, NY: Association for Computing Machinery.

Smith, J. R., Joshi, D., Huet, B., Hsu, W., & Cota, J. (2017). Harnessing AI for augmenting creativity: Application to movie trailer creation. *Proceedings of the 25th ACM international conference on Multimedia* (pp. 1799-1808). New York, NY: Association for Computing Machinery.

Turing, A. M. (1950). Mind-a quarterly review of psychology and philosophy. *Computing Machinery and Intelligence*, 59, 433-460.

Wang, L., Mujib, M. I., Williams, J., Demiris, G., & Huh-Yoo, J. (2021). An Evaluation of Generative Pre-Training Model-based Therapy Chatbot for Caregivers. *arXiv preprint arXiv:2107.13115*.

Yalçın, Ö. N., Abukhodair, N., & DiPaola, S. (2020). Empathic AI Painter: A Computational Creativity System with Embodied Conversational Interaction. *Proceedings of the NeurIPS 2019 Competition and Demonstration Track, Vancouver: CA, 123* (pp. 131-141).

Zacharakis, A., Kaliakatsos-Papakostas, M., Kalaitzidou, S., & Cambouropoulos, E.(2021). Evaluating Human-Computer Co-creative Processes in Music: A Case Study on the CHAMELEON Melodic Harmonizer. *Frontiers in Psychology*, 12, 1-16.

A Appendix A

Example of the narrative text "Nachts" written by Franz Kafka: