AI MDPI

Article

Teaming Up with an AI: Exploring Human–AI Collaboration in a Writing Scenario with ChatGPT

Teresa Luther [1,*], Joachim Kimmerle [1,2] and Ulrike Cress [1,2]

[1] Knowledge Construction Lab, Leibniz-Institut für Wissensmedien, Schleichstraße 6, 72076 Tübingen, Germany; j.kimmerle@iwm-tuebingen.de (J.K.); u.cress@iwm-tuebingen.de (U.C.)
[2] Department of Psychology, Eberhard Karls University Tübingen, Schleichstraße 4, 72076 Tübingen, Germany
[*] Correspondence: t.luther@iwm-tuebingen.de; Tel.: +49-7071-979-240

Abstract: Recent advancements in artificial intelligence (AI) technologies, particularly in generative pre-trained transformer large language models, have significantly enhanced the capabilities of text-generative AI tools—a development that opens new avenues for human–AI collaboration across various domains. However, the dynamics of human interaction with AI-based chatbots, such as ChatGPT, remain largely unexplored. We observed and analyzed how people interact with ChatGPT in a collaborative writing setting to address this research gap. A total of 135 participants took part in this exploratory lab study, which consisted of engaging with ChatGPT to compose a text discussing the prohibition of alcohol in public in relation to a given statement on risky alcohol consumption. During the writing task, all screen activity was logged. In addition to the writing task, further insights on user behavior and experience were gained by applying questionnaires and conducting an additional short interview with a randomly selected subset of 18 participants. Our results reveal high satisfaction with ChatGPT regarding quality aspects, mainly cognitive rather than affect-based trust in ChatGPT's responses, and higher ratings on perceived competence than on warmth. Compared to other types of prompts, mostly content-related prompts for data, facts, and information were sent to ChatGPT. Mixed-method analysis showed that affinity for technology integration and current use of ChatGPT were positively associated with the frequency of complete text requests. Moreover, prompts for complete texts were associated with more copy–paste behavior. These first insights into co-writing with ChatGPT can inform future research on how successful human–AI collaborative writing can be designed.

Keywords: ChatGPT; collaboration; writing; prompting; artificial intelligence

check for updates

Citation: Luther, T.; Kimmerle, J.; Cress, U. Teaming Up with an AI: Exploring Human–AI Collaboration in a Writing Scenario with ChatGPT. *AI* **2024**, *5*, 1357–1376. https://doi.org/10.3390/ai5030065

Academic Editor: Arslan Munir

Received: 17 June 2024
Revised: 29 July 2024
Accepted: 31 July 2024
Published: 5 August 2024

1. Introduction

Artificial intelligence (AI) is increasingly impacting our lives, transforming how we work, learn, and communicate [1]. This development is particularly driven by recent breakthrough developments in generative AI, a subset of AI that consists of algorithms capable of generating new and diverse content, such as images and text [2]. Currently, generative AI tools are gaining popularity, especially for automated text generation (ATG), a trend that is mainly due to the release of ChatGPT by OpenAI in November 2022. This release of ChatGPT for public use was a significant milestone as it not only showcased the potentially game-changing potential of AI, and especially that of large language models (LLM), to a broader audience but also caused widespread public attention with the tool reaching over one million users within one week of its release [3]. As an LLM based on the generative pre-trained transformer (GPT) architecture, ChatGPT has been trained on massive amounts of text data, enabling it to generate human-like text in a conversational manner in response to human prompts [4]. As such, ChatGPT has the capacity to provide detailed responses on various topics quickly, remember the context, and respond to follow-up questions. It should be noted, however, that ChatGPT operates on a probabilistic model predicting the likelihood of each subsequent word based on the previous one in the sequence [5], and that, thus, the objective function of ChatGPT is not a measure of factual accuracy but rather an approximation of linguistic alignment [6].

Since its release, the widespread use of ChatGPT has highlighted the variety of use cases for this technology. In particular, ChatGPT shows remarkable performance in a multitude of natural language processing (NLP) tasks, including generation, classification, and summarization of text, language translation, and question-answering [7]. These advanced NLP capabilities hold great promise for potential applications of ChatGPT within domains such as customer service, education, medicine, and public health [8,9]. Due to their exceptional language generation capabilities and fluency in text generation, AI tools like ChatGPT have been widely recognized as powerful tools for writing.

1.1. ChatGPT as a Tool for Writing

ChatGPT's versatility, adaptability, and ability to mimic different writing styles have made ChatGPT applicable across various domains of writing [10]. In the creative writing content generation domain, ChatGPT can be used for writing poems, stories, songs, and scripts for screenplays [11,12]. In addition, ChatGPT is also increasingly being used in the field of higher education, where it can enhance productivity and learning efficiency [13]. Its applications range from facilitating brainstorming and offering examples of writing styles and structures to providing feedback on students' writing and assisting in essay writing [14]. Moreover, ChatGPT can prove beneficial for students who are writing in a language other than their native language by assisting in translating text from one language to another and providing suggestions for sentence structure and vocabulary [15]. However, the integration of ChatGPT in the context of higher education is not without challenges. Among educators, the tool's capacity to complete written assignments has led to concerns about plagiarism and AI-assisted cheating [16]. As a result, some universities and schools have reacted by imposing bans on AI tools like ChatGPT [17].

In the realm of scientific writing, ChatGPT offers a promising tool for researchers to streamline the scientific writing process [10]. ChatGPT is highly proficient in generating research ideas, aiding in literature review, and drafting research papers [18,19]. Despite its use is not without risks, as it may sometimes generate incorrect content and provide fake references, ChatGPT is increasingly becoming recognized as a powerful writing assistant in the scientific community with the potential to revolutionize academia and scholarly publishing [20]. The tools' functionality of automating the preparation of scholarly manuscripts has led some researchers to explore ChatGPT's full potential in academic writing by involving ChatGPT in their scientific paper writing. For example, King and ChatGPT [21] made use of the interactive nature of ChatGPT and framed their research as a conversation, prompting ChatGPT to generate paragraphs on various topics. Similarly, ChatGPT has been credited as a co-author in a medical perspective research article [22]. The recent rise in scientific publications with ChatGPT listed as a co-author has sparked discussions in the academic community about the appropriateness of crediting AI tools like ChatGPT as co-authors in scholarly publications [23].

Currently, many scientific journals have not issued standardized guidelines for AI-generated content in scientific articles. However, calls for guidelines and policies are growing as the issue of the recognizability of AI-generated text in the academic literature becomes increasingly apparent, and concerns are raised about authorship and scientific misconduct. For instance, it was found that human reviewers correctly identified only 68% of abstracts as generated by ChatGPT, underscoring the potential for AI-generated content to blur the lines of authorship and originality in scientific writing [24]. It has also been observed that authors may hesitate to admit their use of AI tools for various reasons. In their studies, Draxler et al. [25] showed that although human writers do not consider themselves as owners of AI-generated text, they are reluctant to publicly declare AI authorship—a phenomenon discussed as the AI Ghostwriter effect. Researchers are urging journals to disclose the proportion of articles that contain AI-generated content and to publish guidelines for AI use in scientific writing [26,27]. The reaction from academic journals and publishers varies, with some, such as Science and all Springer-Nature journals, precluding AI or LLMs as authors in their statements [23,28], while others, such as Nature and Sage, accept the use of AI as a tool but not as a co-author [29,30]. Despite these variations in journal policies, most journals seem to oppose a strict ban on the use of AI technology to write or edit a manuscript and there appears to be a consensus regarding the importance of disclosing the use of AI technology in accordance with ethical principles of openness, honesty, transparency, and fair allocation of credit in the first place [31]. Some scholars [32] argue that AI tools like ChatGPT currently meet or will, in the future, meet authorship criteria, claiming that AI has transcended its role as merely a tool to increasingly emerge as a collaborative partner of humans in the writing process. As a consequence, the question arises whether we should then also see and recognize ChatGPT as exactly that, a collaborative partner in writing. Importantly, considering that an AI cannot be held responsible for what it does, recognition of AI in scientific writing does not imply a designation of authorship, but rather a disclosure of the use of AI within the scientific article, as highlighted in the recent literature (e.g., ref. [33]). Recently, the importance of disclosing the use of AI in scientific writing has gained visibility, particularly with reports of retractions of scientific articles due to concerns about AI use without declaration.

1.2. Perceptions of ChatGPT

Given the vast popularity and remarkable global attention that ChatGPT gained quickly after its public release, the question arises as to how this new technology is perceived and adopted by the public. Several studies have been conducted with the aim of investigating the public's initial reactions to ChatGPT. For example, Haque et al. [20] analyzed tweets created by early ChatGPT users over three days, starting six days after the release of ChatGPT. Overall, their findings revealed that most tweets were overwhelmingly positive. Notably, early adopters expressed positive sentiments regarding the use of assistance in software development tasks (e.g., debugging and error handling), the generation of entertaining content, and the generation of human-like text. However, negative perceptions were also expressed, mainly regarding the quality of ChatGPT-generated text and regarding the tool's impact on educational aspects. Similar findings of most early users of ChatGPT expressing satisfaction with using the tool were reported in a study where the authors conducted a sentiment analysis of ChatGPT-related tweets over a more extended period of two months after the release date [34]. Mixed perceptions of ChatGPT have also been observed in studies with distinct groups of people, like educators and students [35]. Such positive and negative user perceptions are influenced on the one hand by pragmatic aspects, such as ChatGPT's capability to provide helpful information, and on the other hand by hedonic attributes like entertainment, creativity, and the capacity to impress or surprise users [36].

Several studies and surveys also identified pragmatic aspects as driving factors for using ChatGPT. For example, a semi-structured survey among active adult ChatGPT users in the United States revealed that most respondents used the tool either primarily for information gathering (36.1%) or for problem-solving (22.2%; ref. [37]). In a diary study, the authors found similar results, with most prompts to ChatGPT aimed at information seeking and participants claiming the tool as particularly useful for specific tasks like coding or structured writing whereas rejecting it as a universal tool for all needs [38].

Other factors determining people's acceptance of ChatGPT and user satisfaction with the tool were recently investigated based on established models and theories for predicting the acceptance of technology. The most prominent model in this regard is the Technology Acceptance Model, which suggests that the adoption of technology depends on two critical factors on the part of the users: their perceived usefulness and their perceived ease of use of the technology [39]. Focusing on the first of the two factors and considering technology affinity and information quality, Niu and Mvondo [40] showed that information quality significantly influenced users' perceived usefulness of ChatGPT and their satisfaction with it. In addition, users perceived usefulness and their affinity toward technology significantly impacted their satisfaction with the tool.

The release of ChatGPT has sparked great research interest regarding the tool's capabilities and weaknesses across diverse areas of application and concerning the characteristics of use and users' perceptions of the tool. Currently, most of the findings on these aspects come from surveys. The potential of ChatGPT to assist humans with writing is highlighted by ongoing debates about its integration into education and scientific writing, particularly regarding the question of authorship—whether ChatGPT should be considered merely as a tool or be recognized as a co-author. User studies indicate that people primarily use ChatGPT for assistance with writing tasks. Despite these ongoing debates and increasing usage, there is yet limited research that focuses on how humans interact with ChatGPT while writing.

The requirement for research on the actual writing process has also been highlighted from the perspective of research on computer-supported collaborative learning. Such studies can provide insights that might stimulate further research on the extent to which AI tools can be employed not only for individual learning but also for knowledge construction [41]. Some research is beginning to address this gap by exploring human–AI co-writing in narrative fiction using a prototype system based on the GPT-3 model [42]. However, this specific focus on writing fiction highlights the need for further research on the dynamics of human–ChatGPT co-writing in different writing contexts.

1.3. Research Questions

To the best of our knowledge, there has yet to be a current investigation on human interaction with ChatGPT in the context of a collaborative argumentative writing task. We present an exploratory study addressing this research gap by employing a mixed-method analysis. In general, our goal was to investigate how people interact with ChatGPT when instructed to use the tool for writing an argumentative text on a hypothetical prohibition of alcohol in public. In this study, we aimed to answer the following research questions (RQs):

RQ1: What are the characteristics of user behavior regarding text- and voice-based systems and specifically ChatGPT?

RQ2: How is the behavior during the co-writing process with ChatGPT characterized in terms of prompting and copy–pasting content from ChatGPT?

RQ3: What texts result as products of the co-writing task in terms of length and quality characteristics?

RQ4: What is the degree of user experience satisfaction, trust in ChatGPT's responses, and attribution of human-like characteristics to ChatGPT after the co-writing task?

RQ5: What are the relationships between general user characteristics, the behavior during the co-writing task, text characteristics, and perceptions of ChatGPT after the co-writing task?

2. Materials and Methods

2.1. Participants

The study was conducted with 135 participants (85 females; 48 males; 2 non-binary) between 18 and 71 years (*M*age = 26.86 years; *SD*age = 9.06 years). Participants were recruited through the universities' mailing lists, a local participant recruitment portal, and flyers distributed in local university buildings. The study was conducted in a laboratory and the participants' data was collected anonymously. To be eligible for the study, participants were required to be at least 18 years of age, have German as their native language or have German language skills at a minimum of level C1, and have normal or corrected to normal vision. Participation in the main part of the study lasted about 60 min and was compensated with 12 Euros per person. In total, 18 of the 135 participants were randomly selected to take part in an additional interview of about 10 min duration and were compensated with 15 Euros.

The study was preregistered on the preregistration platform AsPredicted (AsPredicted #132502). Ethics approval for the study was obtained from the institute's local ethics committee (LEK 2023/019) and all research was performed in accordance with the relevant guidelines and regulations. Participants took part voluntarily and provided written informed consent before participation. The sample size, gender, age distribution, and education of participants can be seen in Table 1.

Table 1. Sample size, gender, age distribution, and education of participants.

Gender			Age		Education		
Female	Male	Non-binary	*M* (*SD*)	Range	Elementary school, secondary school, or equivalent certificate	Subject-related entrance qualification, general qualification for university entrance	Completed (college or university of applied sciences) studies
85	48	2	26.86 (9.06)	18–71	1	68	66

2.2. Material and Procedure

Participants were tested in group sessions of up to six participants in a controlled laboratory setting between 3 July and 3 August 2023. The study environment was implemented in a Microsoft Edge browser. For the writing task, ChatGPT (gpt-3.5-turbo model endpoint) was implemented via the official API of OpenAI into the study environment.

After participants filled in the informed consent form, demographic information was collected (gender, age, native language, education level). Then, participants were administered the German version of the Affinity for Technology Interaction (ATI) Scale [43] The unidimensional 9-item ATI Scale was developed as an economic tool to assess ATI as an interaction style rooted in the construct *need for cognition* [43]. Multiple studies confirmed its reliability and validity. Each of the nine items was rated on a 6-point Likert scale: 1 = not true at all, 2 = mostly not true, 3 = rather not true, 4 = rather true, 5 = mostly true, 6 = totally true (Cronbach α = 0.89). Subsequently, participants were presented with a short statement on risky alcohol consumption. They were instructed to discuss alcohol prohibition in public by using ChatGPT to write a text of at least 600 to a maximum of 1000 words. Moreover, they were instructed to imagine while writing that their text would be published in the comment section of a local newspaper and that they should both provide information on the topic and express their personal position. They were informed they would have 40 min to write the text. A word counter and timer were implemented in the writing environment. During the writing task, all screen activities were recorded using Camtasia 9 (TechSmith Cooperation) software. If participants reached a word count over 1000, a warning was displayed on the screen, and the "continue" button was deactivated. With sufficient word count (more than 600 but less than 1000 words) and time remaining, the participants could end the writing task independently. Once the 40 min had elapsed, the participants were automatically forwarded.

Following text creation, the participants answered questions about their general usage behavior related to text- and voice-based dialog systems (type and frequency of use, familiarity, self-assessed competence in using such systems) and about their usage behavior specifically regarding ChatGPT (type and frequency of use). Furthermore, the participants answered questions and completed validated questionnaires about their experience of working with ChatGPT. We used the German version of the Bot Usability Scale (BUS-11; ref. [44]), a translated version of the Human–Computer Trust Questionnaire [45], and a translated version of the Robotic Social Attributes Scale (RoSAS; ref. [46]).

The BUS-11 is a standardized tool comprising 11 items (α = 0.80) distributed across five factors (perceived accessibility to chatbot functions (2 items; α = 0.87), perceived quality of chatbot functions (3 items; α = 0.79), perceived quality of conversation and information provided (4 items; α = 0.77), perceived privacy and security (1 item), time response (1 item)) and judged on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree) to assess user satisfaction after interaction with chatbots [44].

The Human–Computer Trust Questionnaire comprises 25 items (α = 0.93) and measures five constructs underlying perceived trustworthiness with five items each: perceived reliability (α = 0.82), technical competence (α = 0.77), understandability (α = 0.82), faith (α = 0.89), and personal attachment (α = 0.90; ref. [45]). Each item is rated on a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree).

To measure the attribution of anthropomorphic traits to ChatGPT, we used the RoSAS. This scale comprises 18 items, which cover three dimensions: warmth (α = 0.91), competence (α = 0.88), and discomfort (α = 0.80). Each item is rated on a 9-point Likert scale from 1 (not related at all) to 9 (strongly related). We adapted all instructions of the three questionnaires to match the interaction with the chatbot ChatGPT in our study.

Two attention check items (i.e., explicitly instructed response items) were included, one among the items of the ATI and one among the items of the Human–Computer Trust Questionnaire.

Additional queries apart from the questionnaires comprised a question on participants' approach to the writing task, whether they would collaborate again with ChatGPT for a writing task, whether they would collaborate again with ChatGPT for a writing task on the same topic, how competent they rated ChatGPT for several aspects of a writing task, how low/high they rated their own contribution to the text, how they rated their level of knowledge on the topic of alcohol prohibition in public after compared to before collaboratively writing with ChatGPT, whether they learned something about their usage behavior and about how ChatGPT works by collaborating with ChatGPT, and whether they would do anything differently in retrospect when creating the text. Following the main part of the study, an interview of about 10 min duration was conducted with a sub-sample in retrospect on the writing task. See Question S1 for the applied questions in addition to the questionnaires. See the interview protocol in S1 for an overview of the interview protocol. See Images S1–S3 for a visualization of the writing task.

2.3. Statistical Analyses

In addition to the questionnaire data, logfiles contained participants' dialog with ChatGPT (prompts sent to the system and responses by ChatGPT), the final texts created during the writing task, as well as information on the total amount of time spent on the writing task in milliseconds. Data analysis was performed in R (version 4.3.1; ref. [47]). Qualitative text analyses were performed using MAXQDA 2024 [48]. To characterize the quality of the final texts, the German version of the Flesch Reading Ease score (indicative of the readability of the text, with higher scores indicating easier-to-read texts; ref. [49]) and the Type–Token Ratio (indicative of the lexical diversity of the texts) were calculated using an online calculator (https://www.fleschindex.de/, accessed on 30 September 2023) and MAXQDA 2024 [48], respectively. Moreover, the similarity between the final text and the responses of ChatGPT was quantified by an originality score in percent calculated using the plagiarism detection service Turnitin. We used four metrics BLEU (Bilingual Evaluation Understudy; ref. [50]), ROUGE (Recall-Oriented Understudy for Gisting Evaluation; ref. [51]), METEOR (Metric for Evaluation of Translation with Explicit Ordering; ref. [52]), and BERTScore [53], which are common automatic evaluation metrics for text generation, to evaluate whether participants directly copied or paraphrased content from ChatGPT's responses into the final texts. BLEU is a precision-oriented metric, measuring how much the words in the machine-generated text appeared in the human reference text. BLEU is calculated as the ratio of the matching words (in the machine-generated and the human reference text) to the total count of words in the machine-generated text. ROUGE is a recall-oriented metric, measuring how much the words in the human reference text appeared in the machine-generated text. ROUGE is calculated as the ratio of the matching words (in the human reference text and the machine-generated text) to the total count of words in the human reference text. METEOR captures the semantic similarity between the machine-generated text and the human reference text. METEOR is calculated based on word matches between the machine-generated text and the human reference text, using the harmonic mean of the precision/recall and a measure of fragmentation that captures the ordering of the matched words. BERTScore computes semantic textual similarity by matching the words in the machine-generated text and the human reference text via cosine similarity. The metrics were calculated using Python 3.8. For all analyses, results were considered statistically significant at the alpha = 0.05 level ($p < 0.05$, two-tailed).

3. Results

The results of this study concern the following: (a) user characteristics regarding affinity for technology interaction and use of text- and voice-based applications in general and more specifically regarding ChatGPT; (b) the collaborative writing task itself, that is, the analysis of prompts to ChatGPT, the analysis of participants' interaction with ChatGPT, and characteristics of the texts they composed; (c) participants' perceptions of ChatGPT during and after the writing task in terms of satisfaction with using the tool, trust in ChatGPT's responses, attribution of human characteristics to ChatGPT, and rating of ChatGPT's competence for several aspects of a writing task; (d) relations among user characteristics, aspects of the collaborative writing task, and characteristics of the composed texts.

3.1. User Characteristics

The average ATI score was $M = 3.61$ ($SD = 0.91$, Range = 1.56–5.44) on a six-point scale. A two-sample t-test was performed to compare ATI values in male and female participants. There was a significant difference in ATI values between males ($M = 4.08$, $SD = 0.80$) and females ($M = 3.36$, $SD = 0.88$); $t(131) = -4.70$, $p < 0.001$.

Regarding how familiar participants were with chatbots and/or other dialog-oriented applications, participants indicated a mean familiarity of $M = 3.04$ ($SD = 1.25$, Range = 1–5) on a 5-point scale. Regarding how competent they rated themselves in dealing with chatbots and other dialog-oriented applications, participants indicated a mean competence of $M = 3.01$ ($SD = 0.95$, Range = 1–5). Most participants selected ChatGPT when asked which text- and voice-based dialog systems they use ($n = 89$ votes). In total, $n = 31$ participants indicated that they did not use text- or voice-based dialog systems. The indicated areas for usage of text- and voice-based dialog systems are displayed in Figure 1. When asked whether they had ever used ChatGPT or were doing so at the time of the study, 73 participants stated that they were using ChatGPT at the time of the study, whereas 62 participants indicated that they had never used ChatGPT ($n = 31$) or had used ChatGPT in the past, but were not using it at the time of the study ($n = 31$). The indicated areas for usage of ChatGPT are displayed in Figure 2.

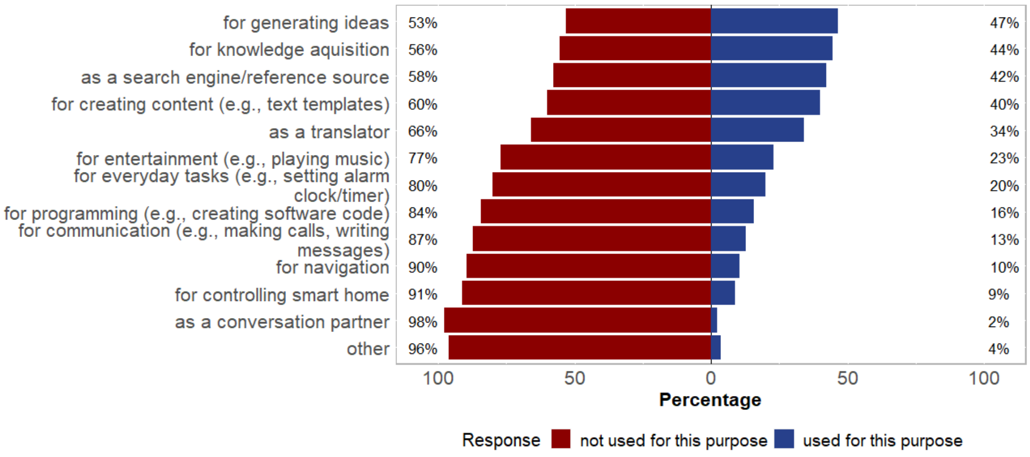

Figure 1. Participants' indicated areas for usage of text- and voice-based dialog systems.

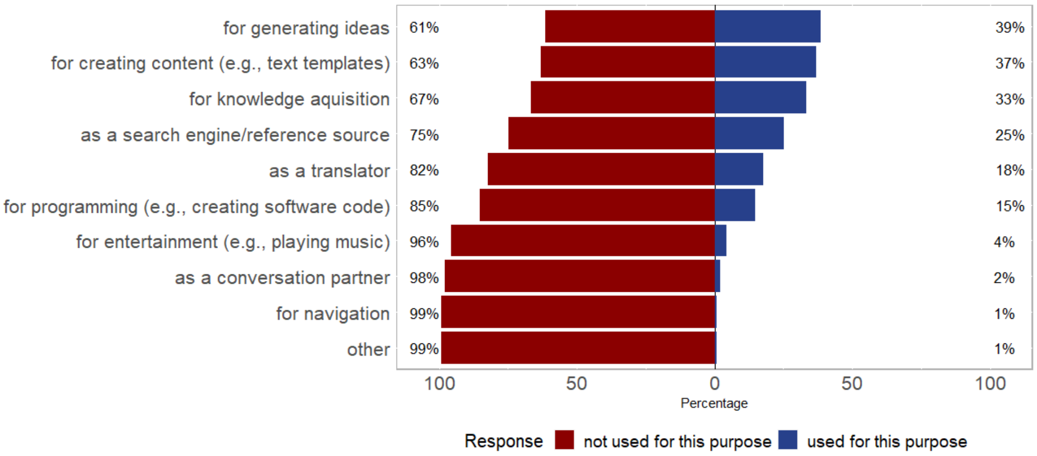

Figure 2. Participants' indicated areas for usage of ChatGPT.

3.2. The Collaborative Writing Task

3.2.1. Prompting Behavior

Of the total of 135 participants, 131 participants used ChatGPT (i.e., sent prompts to ChatGPT) during the collaborative writing task. The four participants not using ChatGPT during the writing task composed the argumentative text entirely by themselves. The number of prompts sent to ChatGPT ranged from 1 prompt to 25 prompts with an average number of $M = 8.15$ ($SD = 4.45$) prompts.

Using an iterative–deductive process, we developed a coding scheme (see Table S1) consisting of six categories based on the following prompt content: (a) *content*, with subcategories sources, individual text sections, definitions, examples and quotations, opinion of ChatGPT, arguments, and data, facts, information; (b) *form*, with subcategories form (linguistic), form (content), and structure; (c) *complete text*; (d) *relationship*; (e) *question/scrutinize*; and (f) *other*. Based on the coding scheme, two independent coders blindly coded the prompts from the $n = 131$ dialogs with ChatGPT. With a 95% segment overlap, the overall average percentage match between the coders was 94.8%, ranging from 82.4% to 97.9% for the individual categories (see Table S2 for the code-specific result table). Random correction was not performed as coders were free to determine the segment boundaries [54]. For a visualization of the prompt frequency in the different top-level categories, see Figure 3, and for a visualization of the prompt frequency in the subcategories of the category *content*, see Figure 4.

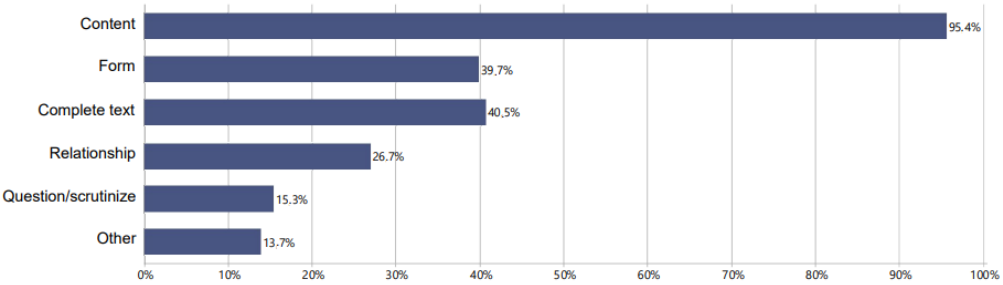

Figure 3. Percentage of prompts in the top-level categories (unit of analysis: $n = 131$ dialogues).

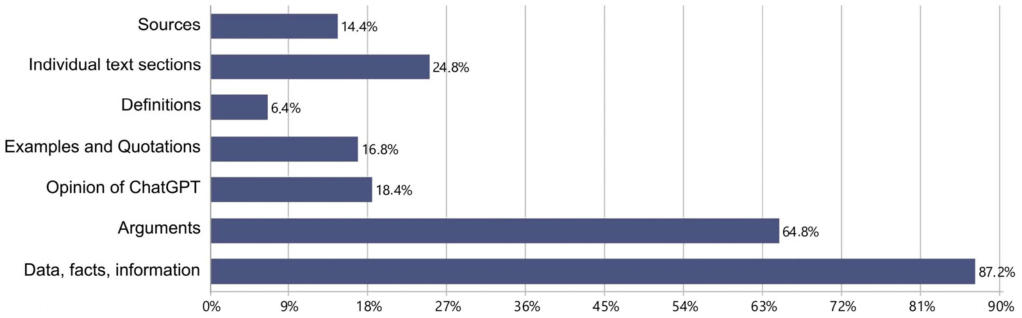

Figure 4. Percentage of prompts in the subcategories of the category *content* (unit of analysis: *n* = 131 dialogues).

3.2.2. Text Creation

On average, participants spent *M* = 36.78 (*SD* = 5.98, Range = 10.47–40.00) minutes on the writing task.

Participants' average self-rated contribution to the final texts was *M* = 48.9% (*SD* = 28.5%, Range = 3–100%). As an indicator of whether and to what degree participants copied text from ChatGPT's responses into the final texts, the similarity of the composed texts with ChatGPT's responses was assessed. The originality scores generated by Turnitin ranged from 0.00–100.00% with an average score of *M* = 42.7% (*SD* = 37.1%).

To further assess participants' interaction with ChatGPT during the writing task in terms of whether they collaborated with the tool or rather just copy–pasted content from ChatGPT into their texts, the metrics BLEU, ROUGE, METEOR, and BERTScore, were calculated. Originally, these metrics are used for automatic evaluation of natural language generation, such as machine translation, and compute a similarity score for each token in the candidate sentence with each token in the reference sentence. To evaluate the syntactic and semantic equivalence of the composed texts and ChatGPT's responses, we compared each participant's composed text (candidate document) with the respective responses of ChatGPT in the dialog (reference document). BLEU compares n-grams of the candidate document with the n-grams in the reference document and counts the number of matches. In our case, a perfect match, that is, a BLEU score of 1.00, would indicate that the text from ChatGPT's responses corresponds exactly to the final text. The average BLEU score was *M* = 0.21 (*SD* = 0.16, Range = 0.00–0.67), indicating a moderate level of overlap between ChatGPT's responses and the final texts. ROUGE is a set of metrics with ROUGE-1 and ROUGE-2 referring to the overlap of unigrams and bigrams, respectively, and with ROUGE-L referring to the longest sequence of words that appear in the same order in both the reference document and the candidate document. For each ROUGE metric, the F1 score is the harmonic mean of precision and recall. We found an average ROUGE-1 F1 score of *M* = 0.46 (*SD* = 0.17, Range = 0.21–0.82), an average ROUGE-2 F1 score of *M* = 0.23 (*SD* = 0.21, Range = 0.01–0.76), and an average ROUGE-L F1 score of *M* = 0.45 (*SD* = 0.17, Range = 0.19–0.82). These results indicate a moderate degree of similarity between ChatGPT's responses and the final texts, with more overlap at the individual word level (ROUGE-1) and in sentence-level structure (ROUGE-L) than at the level of consecutive word pairs or phrases (ROUGE-2). METEOR, as the harmonic mean of unigram precision and recall, with a higher weight on recall, incorporates additional semantic matching based on stems and paraphrasing. The average METEOR score of *M* = 0.26 (*SD* = 0.13, Range = 0.07–0.83) indicates a moderate level of overlap between ChatGPT's responses and the final texts. Using sentence representations from the deep learning model BERT, the BERTScore computes the cosine similarity between contextual embeddings of the words in the candidate and reference sentences. The BERTScore F1 metric provides a balanced measure of precision and recall. We found an average BERTScore F1 of *M* = 0.74 (*SD* = 0.05, Range = 0.63–0.94), which indicates a high similarity between ChatGPT's responses and the final texts in terms of the contextual alignment of their content.

3.2.3. Text Characteristics

The average length of the composed texts was M = 632.56 (SD = 126.15, Range = 272–1013) words. Regarding the quality of the texts, lexical variety, quantified by the Type–Token Ratio, ranged from 0.06 to 0.11 (M = 0.07, SD = 0.009), and text readability, quantified by the German version of the Flesch Reading Ease, ranged from 12.00 to 52.00 (M = 29.46, SD = 9.88). These results indicate an average low lexical variability and a difficult readability (i.e., understandable for academics) of the texts.

3.3. Perception of ChatGPT during and after the Writing Task

The average overall BUS-11 score was M = 75.5% (SD = 10.7%, Range = 20.0–92.7%) and thus above the score of 60% defined by the authors of the scale as a sufficient level of satisfaction. Descriptive statistics of the five BUS-11 factors, the Human–Computer Trust Questionnaire, and the RoSAS are depicted in Table 2.

Table 2. Descriptive statistics of the BUS-11, the Human–Computer Trust Questionnaire, and the RoSAS.

Questionnaire	*M* (*SD*)	Range
BUS-11		
Perceived accessibility to chatbot functions	4.57 (0.75)	1.00–5.00
Perceived quality of chatbot functions	4.30 (0.74)	1.00–5.00
Perceived quality of conversation and information provided	3.59 (0.74)	1.00–5.00
Perceived privacy and security	1.87 (1.02)	1.00–5.00
Time response	3.30 (1.10)	1.00–5.00
Human–Computer Trust Questionnaire	3.85 (0.96)	1.35–6.19
Perceived reliability	4.24 (1.19)	1.00–6.40
Perceived technical competence	4.37 (1.07)	1.00–6.60
Perceived understandability	4.70 (1.26)	1.00–7.00
Faith	2.83 (1.42)	1.00–6.40
Personal attachment	2.47 (1.45)	1.00–6.40
RoSAS		
Warmth	2.46 (1.74)	1.00–7.67
Competence	6.16 (1.72)	1.33–9.00
Discomfort	2.69 (1.42)	1.00–7.33

Most participants (n = 106, 78.5%) indicated that it would be rather or extremely likely that they would collaborate again with ChatGPT for a writing task, and n = 78 (57.8%) of participants indicated that it would be rather or extremely likely that they would collaborate again with ChatGPT for a writing task on the same topic.

Regarding the rated competence of ChatGPT for several aspects of a writing task, the tool was particularly assigned competence for the creation of arguments, research of synonyms, and correction of grammatical and spelling mistakes (see Figure 5). ChatGPT was not assigned a higher level of competence by the participants compared to their own for any of the listed aspects (see Figure 6). Most participants (n = 119, 88.1%) rated their level of knowledge on the topic of alcohol prohibition in public after the writing task as about the same (n = 59) or as rather better (n = 60) compared to before the task, while the other (n = 16, 11.9%) rated their level of knowledge as much better (n = 6) or as rather worse (n = 10).

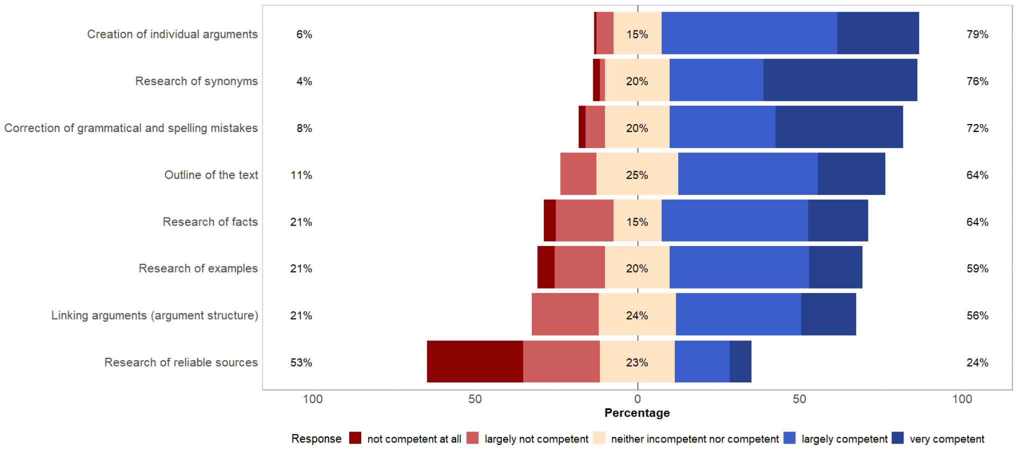

Figure 5. Rated competence of ChatGPT for several aspects of a writing task.

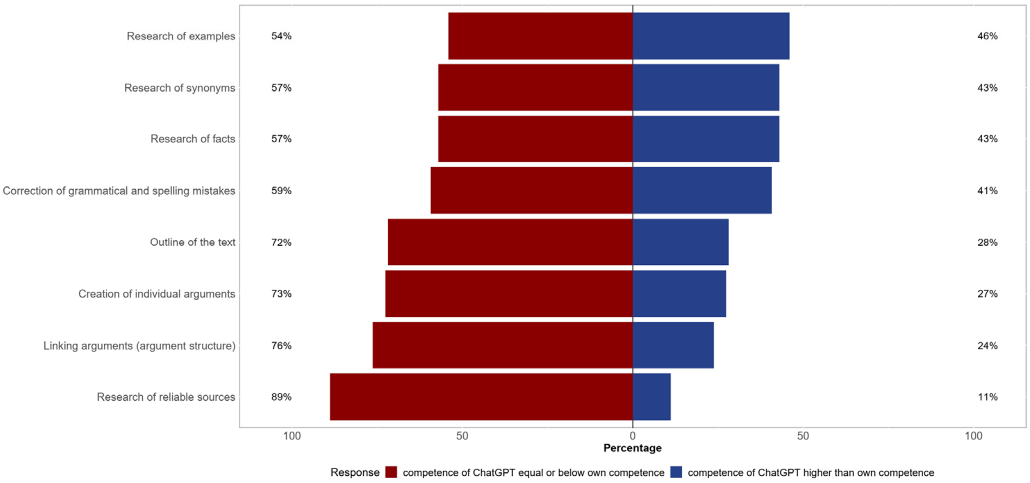

Figure 6. Assigned competence level to ChatGPT for aspects of a writing task compared to participants' own competence.

Regarding the questions on whether participants learned anything about how ChatGPT works and about their user behavior by collaborating with ChatGPT, most participants (n = 104, 77.0%) indicated they had learned somewhat (n = 54), rather much (n = 37) or much (n = 13) to the first question, whereas to the latter question, most participants (n = 98, 72.6%) indicated that they had learned nothing (n = 14), rather little (n = 41) or somewhat (n = 43).

The notes from the 18 conducted interviews provide further information on the participants' approach to the writing task and their experience of ChatGPT while using it. Eight participants mentioned that they adopted an experimental approach and described their approach as using ChatGPT when they encountered knowledge gaps in their own writing, but without having explicitly defined this strategy. In contrast, seven participants reported that they proceeded strategically so that they had an outline for the text in mind and specific questions for ChatGPT. Two participants indicated that they combined strategic and trial-and-error methods. Seven participants stated that ChatGPT completely positively convinced them, whereas six participants indicated that ChatGPT has led to conviction and skepticism among them, and four indicated that ChatGPT had exclusively triggered skepticism. One interviewee did not use ChatGPT during the writing task and thus did not provide information on using a strategy or whether the tool triggered conviction or skepticism.

3.4. Relations among User Characteristics and Characteristics of the Writing Process

Significant positive correlations were found between participants' mean ATI score and frequency of prompts for complete texts (r = 0.19, p = 0.035), frequency of prompts for individual text sections (r = 0.18, p = 0.018), use of anthropomorphic language with ChatGPT (r = 0.20, p = 0.025), prompts concerning form (content) (r = 0.32, p < 0.001), and questioning or scrutinizing the responses of ChatGPT (r = 0.25, p = 0.005). Conversely, there was a significant negative correlation between participants' mean ATI score and the frequency of prompts for data, facts, and information (r = −0.20, p = 0.011).

To assess whether participants' experience in using ChatGPT was associated with the type of prompts they sent during the writing task, we performed Chi-Square tests of independence with all codes binarized (0 = not coded, 1 = coded). There was a significant relationship between participants' experience in using ChatGPT and the frequency of at least one prompt for the complete text, X^2 (2, 131) = 13.38, p = 0.001. Moreover, we found significant relationships between participants' experience in using ChatGPT and the frequency of use of anthropomorphic language with ChatGPT, X^2 (2, 131) = 10.13, p = 0.006, and the frequency of questioning/scrutinizing ChatGPT, X^2 (2, 131) = 6.36, p = 0.042. Regarding prompts for *content*, Chi-Square tests revealed significant relationships between participants' experience in using ChatGPT and frequency of prompts for individual text sections, X^2 (2, 131) = 8.35, p = 0.015, and between participants' experience in using ChatGPT and frequency of prompts for definitions, X^2 (2, 131) = 8.19, p = 0.017. Regarding prompts for *form*, Chi-Square tests revealed a significant relationship between participants' experience in using ChatGPT and the frequency of prompts for content-related form, X^2 (2, 131) = 14.59, p < 0.001. See Figure S3 for an overview of the proportion of the type of prompt coding depending on participants' experience in using ChatGPT.

3.5. Relations among Characteristics of the Writing Process

Correlation analyses between writing time and the frequency of types of prompts to ChatGPT yielded significant results regarding complete text requests (r = −0.25, p = 0.004) and prompts for data, facts, and information (r = 0.38, p < 0.001). The results of the performed correlation tests are displayed in Figure S1.

Correlation analyses between the originality scores (i.e., the similarity of the composed texts with ChatGPT's responses) and the frequency of types of prompts to ChatGPT yielded significant results regarding complete text requests (r = 0.56, p < 0.001), use of anthropomorphic language with ChatGPT (r = 0.25, p = 0.004), prompts for individual text sections (r = 0.30, p < 0.001), prompts for data, facts, information (r = −0.38, p < 0.001), and prompts related to form (content) (r = 0.52, p < 0.001). The results of the performed correlation tests are displayed in Figure S2. Significant positive correlations were found between writing time and participants' self-rated contribution to the text (r = 0.44, p < 0.001). A significant negative correlation was revealed between the similarity of the composed texts with ChatGPT's responses and participants' self-rated contribution to the text (r = −0.73, p < 0.001). Correlation tests between the number of prompts sent to ChatGPT and characteristics of the writing process showed a significant positive correlation with the time spent on the writing task (r = 0.20, p = 0.024) and no significant correlation with the similarity of the composed texts with ChatGPT's responses (r = 0.12, p = 0.170).

3.6. Relations between Writing and Text Characteristics

Regarding associations between characteristics of the writing process and characteristics of the composed texts, a correlation test showed a significant negative correlation between writing time and text length (r = −0.33, p < 0.001). A significant positive correlation was found between writing time and the German version of the Flesch Reading Ease (r = 0.35, p < 0.001). Writing time and Type–Token Ratio were not found to be significantly correlated (r = 0.09, p = 0.326). Significant negative correlations were revealed between the similarity of the composed texts with ChatGPT's responses and Type–Token Ratio (r = −0.32, p < 0.001), and with the German version of the Flesch Reading Ease (r = −0.66,

$p < 0.001$). A significant positive correlation was found between the similarity of the composed texts with ChatGPT's responses and text length ($r = 0.23$, $p = 0.006$).

4. Discussion

4.1. Main Findings

4.1.1. Behavior Regarding Text- and Voice-Based Systems, and ChatGPT

The questionnaire data collected as part of this study revealed that ChatGPT was the most frequently selected tool in the sample, compared to other text- and voice-based dialog systems, reflecting the tool's widespread popularity. Interestingly, newer tools available at the time of data collection, such as Google Bard, released in Europe on 13 July 2023, were not chosen as frequently as ChatGPT, which had been available longer. These results and our finding that most participants in our sample indicated that they currently use ChatGPT suggest that ChatGPT has established a solid user base. This development is also supported by a survey from Statista Consumer Insights, which revealed that most of the respondents who had used ChatGPT during the survey period (23 March 2023 to 5 April 2023) indicated they would use the tool again [55]. However, it must also be noted that in our study, a considerable number of 31 participants indicated that, at some point, they had stopped using ChatGPT. Moreover, 31 participants indicated they did not use text- or voice-based dialog systems. Although the reasons for this finding remain unclear, it is possible that a lack of awareness of the use of such systems played a role, especially considering that we did not provide an explanation of how these systems are defined alongside the question. A lack of awareness about the presence of ATG technologies has previously been demonstrated [56]. The four primary reported uses for text- and voice-based systems, specifically ChatGPT, were generating ideas, knowledge acquisition, as a search engine/reference source, and for creating content (e.g., text templates). Similar to a diary study with a relatively small sample of students and professionals in India [38] and a survey with a more representative sample of ChatGPT users in the United States [37], these findings highlight pragmatic aspects as the driving forces of ChatGPT use. There is much debate concerning the use of ChatGPT for knowledge acquisition. Based on its training, ChatGPT's knowledge base is limited. Often, ChatGPT provides incorrect or biased responses [57], and while in comparison to Google Search, it can serve as a valuable resource for general information, it may not be as effective for specialized guidance (e.g., [58]). Thus, it is advised to consult multiple sources, not just ChatGPT, for accurate understanding [59].

4.1.2. Behavior during Co-Writing with ChatGPT

Concerning the collaborative argumentative writing task, our findings revealed a large range of the number of prompts that were sent to ChatGPT. While some participants opted for one-shot prompting, others sent multiple prompts during the task. Notably, four participants did not send any prompts, completing the task independently of ChatGPT. A closer analysis of the prompts' content showed that they were predominantly content-related, as opposed to, for example, prompts for complete texts. Most of these content-related prompts were queries for data, facts, and information. There were also many prompts for arguments. This pattern of results is consistent with the findings from a diary study, which found that most prompts to ChatGPT were for information seeking [38]. Moreover, the nature of the prompts also reflects the participants' rated competence of ChatGPT for aspects of a writing task, with the highest rating for the capability to create individual arguments. Our findings align with previous research findings indicating that ChatGPT is frequently used for information-seeking tasks due to its conversational interface. However, several studies also highlight significant concerns regarding the accuracy and reliability of information provided by ChatGPT. For example, in the medical domain, ref. [60] found that prompting ChatGPT with 88 questions from the daily routine of radiologists, resulted in correct responses in only about two-thirds, with the remainder of responses containing errors. Moreover, they found the majority of the references provided by ChatGPT to be fabricated, raising concerns about the reliability of its outputs. In another recently conducted study, the authors found that users perceive ChatGPT's responses to be of higher information quality compared to the information obtained from Google Search; however, their findings also revealed pronounced limitations of ChatGPT compared to Google Search in fact-checking tasks [61]. Together, these findings highlight the importance of user awareness regarding the limitations of ChatGPT. While it can be a useful tool for generating ideas and providing general information, it is crucial for users to cross-check its output with reliable sources. Overall, our findings suggest that participants primarily used ChatGPT to obtain content for independent compilation into a text, indicating a preference for the tool's support in constructing arguments and content over obtaining complete texts that may require extensive editing and modification.

The number of prompts to ChatGPT was not associated with the copy–paste of content from ChatGPT's responses to the texts. However, we found significant positive correlations between the copy–paste behavior of participants and certain types of prompts they sent to ChatGPT. More copy–paste behavior was associated with a higher frequency of complete text requests, requests for individual text sections, and requests related to form (content). Furthermore, more copy–paste behavior was associated with higher use of anthropomorphic language in the prompts to ChatGPT. These results suggest that the texts generated by ChatGPT in response to prompts for complete text or text sections were mainly adopted by the participants and therefore not scrutinized further. Such findings raise the question of whether ChatGPT induces social loafing (see [62]), a socio-psychological phenomenon where individuals exert less effort in a group than when working alone. So far, this question has not been addressed in research. Participants might perceive ChatGPT as a capable team member due to its capabilities, leading to a diffusion of responsibility in the way that they might assume that ChatGPT will handle significant portions of the task, reducing their perceived need to contribute. Especially considering that there was a time limit for the writing task in our study, it is plausible that the efficiency and convenience offered by ChatGPT might have encouraged an overreliance on the tool. Notably, we observed that the more complete text requests were sent to ChatGPT, the significantly less time was spent on the writing task. Significantly more time was spent on the writing task; however, more prompts for data, facts and information were sent.

4.1.3. Texts as Products of Co-Writing with ChatGPT

The texts as products of the writing task varied greatly in length. Moreover, we found great variety among the participants regarding the time they spent on the writing task. These results could be explained by differences in participants' prompting (one-shot vs. multiple-shot) and copy–paste behavior. Specifically, we found that the originality of the final texts compared to ChatGPT's responses ranged from 0 to 100% among the participants. More fine-grained analysis of equivalence on the semantic and syntactic levels showed overlap, particularly on the individual word level as well as regarding sentence structure and context. The text of one participant contained more than the declared maximum of 1000 words, probably due to the text leading to a warning message. Then, the participant waited until the time for the task had elapsed without changing the word count.

4.1.4. Perception of ChatGPT after Co-Writing: Satisfaction, Trust, and Human-Likeness

Regarding the participants' perception of ChatGPT, we found a generally high level of satisfaction with ChatGPT after the writing task. The high proportion of 79% of participants who indicated that they would rather or very likely collaborate with ChatGPT again for a writing task underlines this finding. The participants particularly expressed positive sentiment concerning the accessibility of ChatGPT's functions, the quality of its functions, and the quality of conversation and information provided by it. Participants were significantly less satisfied with the privacy and security aspects, reflecting the main concerns related to ChatGPT that are also expressed by the public [63]. Based on previous research findings [40], the high satisfaction ratings in the questionnaire could also be partly attributable to the moderate technology affinity among the sample. However, the interviews show a slightly different picture: About the same number of participants stated that ChatGPT ultimately convinced them while writing, as well as that ChatGPT led to both conviction and skepticism. The reasons for skepticism included that the content generated seemed strange or even incorrect and that specific questions were not possible. Similar to the validation studies of the ATI Scale, we found a gender effect, with men indicating a higher ATI. Besides this finding relying on self-report, it should also be considered in light of the findings of a recent study, which showed that technology-related education eliminated the gender effect in LLM technology adoption [25].

Participants' trust in ChatGPT's responses and their tendency to anthropomorphize, that is, attribute human characteristics to ChatGPT, relate to similar aspects of ChatGPT, for which they indicated high levels of satisfaction. Concerning trust, our findings revealed higher levels of cognition-based trust in ChatGPT, reflecting beliefs in its consistent functioning, task accuracy, and understandable functionality, compared to affect-based trust, reflecting beliefs in its functioning in untested situations and the users' preference for the system. These findings indicate that while participants were confident in ChatGPT's current capabilities and understood its functioning, they expressed concerns regarding the tool's performance in novel situations and indicated a lower preference for the system. ChatGPT was perceived as competent by the participants but received considerably lower ratings regarding the perceptual dimensions of warmth and discomfort. Taking into account recent findings of significantly worse task performance associated with the use of ChatGPT compared to the use of a search engine [64], it becomes clear that people may be biased in their assessment of ChatGPT's functional capabilities and have difficulties in objectively judging ChatGPT's functional capabilities, which poses the risk of overreliance on the tool and can result in an unreflected acceptance of ChatGPT responses. In the same vein, it has been found that a further threat to adequate credibility judgments is posed by the conversational nature of ChatGPT [65]. Our findings are consistent with McKee et al.'s [66] work that deals with the spontaneous emergence of perceived warmth and competence in impressions of prominent examples of AI technology. By analyzing participants' responses along several perceptual dimensions, they found that participants frequently discussed AI systems in terms of ability, a facet of competence according to Abele et al. [67], and morality and sociality, both facets of warmth [66]. Overall, they found significantly more competence-related content than warmth-related content in the participants' impressions of AI systems. McKee et al. [66] evaluated judgments on AI systems other than ChatGPT. Although warmth and competence judgments may depend very much on the particular AI system [68], it is interesting that our study yielded similar results regarding the AI system ChatGPT.

4.1.5. Relationships between User Characteristics, Behavior during Co-Writing, Text Characteristics and Perceptions of ChatGPT

Participants' ATI and their use of ChatGPT were found to be significantly related to their interaction with ChatGPT during the writing task. We observed moderate positive correlations between ATI and the frequency of complete text requests, requests for individual text sections, requests related to form (content), the frequency of questioning or scrutinizing ChatGPT, and the frequency of the use of anthropomorphic language in the dialog with ChatGPT. Compared to participants who indicated not currently using ChatGPT, there was a higher proportion of participants among those who indicated currently using ChatGPT who at least once sent a complete text request, sent a request for text sections, used anthropomorphic language with ChatGPT, questioned or scrutinized ChatGPT. Considering that the use of LLMs has been found to be positively associated with technology-related education [25], and given that affinity for technology has been related to knowledge concerning new technologies [69], our findings could be explained by the participants indicating higher ATI and current use of ChatGPT also having more knowledge about the capabilities of new technological systems overall, and ChatGPT specifically. Thus, compared to participants with lower ATI and no current use of ChatGPT, they might have been aware that ChatGPT is able to generate complete texts, which is reflected in their higher frequency of prompting for complete texts and text sections. The description of this observation is not meant to be evaluative, and the elicitation of entire texts does not necessarily imply a particularly reasonable approach. On the contrary, this approach can also be quite problematic. However, the findings show that these people were at least aware that the tool has these capabilities. Moreover, these participants used more anthropomorphic language in their dialogs with ChatGPT, which could be due to them experiencing the writing with ChatGPT more like having a natural conversation. The moderate negative correlation between participants' ATI and the frequency of prompting for data, facts, and information indicates that participants with a lower ATI used ChatGPT more like a Google Search-like tool.

Regarding quality aspects of the texts, we found a moderate negative correlation between the similarity of the composed texts with ChatGPT's responses and the lexical variety of the texts. Moreover, there was a high negative correlation between the similarity of the composed texts with ChatGPT's responses and the readability of the texts (Flesch Reading Ease). The finding of more content adoption from ChatGPT's responses associated with a lower lexical variety of the texts might be due to the fact that ChatGPT has been found to particularly excel in the English language [70], while in our study, the participants interacted with ChatGPT in German. Another reason for the negative correlation could be the fact that our sample mainly consisted of highly educated participants, most likely with a relatively large vocabulary, which is also reflected in the higher lexical variety of the texts. These findings indicate that ChatGPT generates text that includes more repetitive (than unique) words and text with a more complex sentence structure, making the text more challenging to read.

Based on our findings regarding the originality of the texts and the rating of the participants' contribution to the texts, it can be concluded that the participants had a relatively good insight into the proportion of text they had contributed to the texts and the proportion of text that ChatGPT had generated. Considering the ongoing debate on AI authorship, these findings do not suggest an AI Ghostwriter effect in our sample, as reported in a study by Draxler et al. [25]. Based on our findings, we would expect that when asked whether they would have credited co-authorship to ChatGPT, most participants would probably have answered this question with "yes" (for discussions on the perception of human vs. AI authorship see also [71–73]).

4.2. Limitations

While the present study provides valuable insights into human–AI collaborative writing, several limitations should be considered. First, we conducted the study under laboratory conditions and thus were limited to recruiting in the immediate vicinity. Regarding gender, age, and education, there was an over-representation of young, highly educated participants who identified as female. As a result, caution needs to be taken in generalizing our findings to a broader population. A recent study showed that younger people are more inclined to use LLMs [25]. Consequently, it might be interesting to examine collaborative writing with ChatGPT among a wider range of participants, including different age groups and education levels. Furthermore, our study focused on writing in the context of argumentative writing, which differs in many aspects from other kinds of writing, such as the writing of medical reports, which could be investigated in future studies. In addition to the type of writing task, future studies should also vary the writing topic. Due to the novelty of the research field, there are still few validated questionnaires for assessing user perception and behavior, specifically regarding (text-generative) AI tools. Therefore, for this exploratory study, we relied heavily on open-ended questions, and it should be a task of future research to work on validated questionnaires in this field. Lastly, we implemented GPT-3.5 in our study. Since the study was conducted, OpenAI continued to work on ChatGPT, and future studies should consider implementing newly released versions.

5. Conclusions

The current study represents one of the pioneering attempts to investigate the dynamics of people's interaction with ChatGPT in a collaborative writing setting. Even though ChatGPT is increasingly used in different areas of writing and has become a focus of current debates in education and research due to its ability to act as a writing assistant, the actual process of co-writing with ChatGPT has not yet been the focus of research. Prior works have started to address this research gap by studying human–AI co-writing in the context of narrative fiction [42] and by examining users' perceptions of co-writing with ChatGPT through surveys (e.g., [37]). By employing an exploratory, multi-method approach, our study sheds light on (a) the users' approach and their behavior during the actual co-writing with ChatGPT, (b) correlations between their approach and behavior during co-writing and their user characteristics and characteristics of the text as the product of co-writing, and (c) the users' perception of ChatGPT in general and related to co-writing.

Our study demonstrates that ChatGPT was very popular several months after release and that pragmatic aspects, such as ChatGPT's capability of generating content, were driving factors for using it. Our sample was for the most part eager to collaborate with ChatGPT during the writing task and indicated high satisfaction with ChatGPT after writing. Despite ChatGPT's capability to generate comprehensive texts within seconds, our findings show that people seem to prefer to compile information into a text themselves and to integrate ChatGPT into the writing process mainly as a source of data, facts, and information, rather than as a source of ready-made texts. These findings could open new avenues for future research to assess potential relationships with a sense of personal responsibility for the final text as the product of collaboration. Differences in interaction behavior with ChatGPT during co-writing were evident in our study. We observed a large variability in the number and type of prompts sent to ChatGPT, as well as in the extent to which content from ChatGPT's responses was copied and pasted into the final texts. Factors that might explain these variabilities in how ChatGPT is integrated into the collaborative writing process are people's experiences with ChatGPT and their general affinity for interacting with technology. A higher frequency of prompting ChatGPT for complete texts was associated with more copy–paste of content from ChatGPT's responses to the texts, indicating this type of prompting leads to more unchecked use of ChatGPT. In light of these findings, future research should investigate both prompting and copy–paste behavior and their effects on collaboration in more detail to address the questions of how ChatGPT and other AI tools can best be integrated into the writing process and how successful collaboration can be fostered in this context.

Considering the continuous evolution and release of new AI tools, along with their increasing intrusion into fields like education and science, the emphasis for education institutions and society as a whole should shift toward nurturing critical thinking skills rather than advocating for a broad ban on these technologies. While these skills are probably essential for adapting to technological advancements, they might also play a crucial role in fostering successful collaboration between humans and AI technology.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ai5030065/s1, Questions S1: Applied Questions Additional to the Questionnaires; Image S1: Visualization of the Writing Task; Image S2: Visualization of the Writing Task; Image S3: Visualization of the Writing Task; Table S1: Used Coding Scheme for the Prompts; Figure S1: Results of the correlation tests between the writing time and the different types of prompts sent to ChatGPT; Figure S2: Results of the correlation tests between the similarity of the composed texts with ChatGPT's responses (originality score) and the different types of prompts sent to ChatGPT; Figure S3: Proportion of type of prompt coding depending on participants' usage of ChatGPT.

Author Contributions: Conceptualization: T.L. and J.K.; methodology: T.L. and J.K.; formal analysis: T.L.; investigation: T.L.; data curation: T.L.; writing—original draft preparation: T.L.; writing—review and editing: J.K. and U.C.; visualization: T.L.; supervision: J.K.; project administration: J.K.;

funding acquisition: J.K. and U.C. All authors have read and agreed to the published version of the manuscript.

Funding: The research reported here was funded by budget resources of the Leibniz-Institut für Wissensmedien.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Ethics Committee of the Leibniz-Institut für Wissensmedien (protocol code: LEK 2023/019; date of approval: 16 May 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original data presented in the study are openly available on OSF at https://osf.io/7rxq2/, accessed on 1 August 2024.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Long, D.; Magerko, B. What is AI Literacy? Competencies and Design Considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–16.
2. McKinsey Consultant. What Is Generative AI? 2023. Available online: https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai (accessed on 13 June 2024).
3. Altman, S. [@sama] ChatGPT Launched on Wednesday. Today It Crossed 1 Million Users! 2022. Available online: https://x.com/sama/status/1599668808285028353?lang=en (accessed on 13 June 2024).
4. OpenAI. ChatGPT. 2023. Available online: https://openai.com/blog/chatgpt (accessed on 13 June 2024).
5. Titus, L.M. Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy. *Cogn. Syst. Res.* **2024**, *83*, 101174. [CrossRef]
6. Tan, T.F.; Thirunavukarasu, A.J.; Campbell, J.P.; Keane, P.A.; Pasquale, L.R.; Abramoff, M.D.; Kalpathy-Cramer, J.; Lum, F.; Kim, J.E.; Baxter, S.L.; et al. Generative Artificial Intelligence through ChatGPT and Other Large Language Models in Ophthalmology: Clinical Applications and Challenges. *Ophthalmol. Sci.* **2023**, *3*, 100394. [CrossRef] [PubMed]
7. Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.L.; Tang, Y. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA J. Autom.* **2023**, *10*, 1122–1136. [CrossRef]
8. Howard, A.; Hope, W.; Gerada, A. ChatGPT and antimicrobial advice: The end of the consulting infection doctor? *Lancet Infect. Dis.* **2023**, *23*, 405–406. [CrossRef] [PubMed]
9. Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
10. Huang, J.; Tan, M. The role of ChatGPT in scientific communication: Writing better scientific review articles. *Am. J. Cancer Res.* **2023**, *13*, 1148–1154.
11. Lucy, L.; Bamman, D. Gender and representation bias in GPT-3 generated stories. In Proceedings of the Third Workshop on Narrative Understanding, Virtual, 11 June 2021; Association for Computational Linguistics: Kerrville, TX, USA, 2021; pp. 48–55.
12. Atlas, S. ChatGPT for Higher Education and Professional Development: A Guide to Conversational AI. Available online: https://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1547&context=cba_facpubs (accessed on 20 January 2024).
13. Fauzi, F.; Tuhuteru, L.; Sampe, F.; Ausat, A.M.A.; Hatta, H.R. Analysing the role of ChatGPT in improving student productivity in higher education. *J. Educ.* **2023**, *5*, 14886–14891. [CrossRef]
14. Su, Y.; Lin, Y.; Lai, C. Collaborating with ChatGPT in argumentative writing classrooms. *Assess. Writ.* **2023**, *57*, 100752. [CrossRef]
15. Lund, B.D.; Wang, T. Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Libr. Hi Tech. News* **2023**, *40*, 26–29. [CrossRef]
16. Cotton, D.R.E.; Cotton, P.A.; Shipway, J.R. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innov. Educ. Teach. Int.* **2024**, *61*, 228–239. [CrossRef]
17. Grassini, S. Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Educ. Sci.* **2023**, *13*, 692. [CrossRef]
18. Dowling, M.; Lucey, B. ChatGPT for (Finance) research: The Bananarama Conjecture. *Financ. Res. Lett.* **2023**, *53*, 103662. [CrossRef]
19. Macdonald, C.; Adeloye, D.; Sheikh, A.; Rudan, I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. *J. Glob. Health* **2023**, *13*, 01003. [CrossRef] [PubMed]
20. Haque, M.U.; Dharmadasa, I.; Sworna, Z.T.; Rajapakse, R.N.; Ahmad, H. "I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data. *arXiv* **2022**, arXiv:2212.05856. [CrossRef]

21. King, M.R.; ChatGPT. A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cell. Mol. Bioeng.* **2023**, *16*, 1–2. [CrossRef] [PubMed]
22. ChatGPT Generative Pre-trained Transformer; Zhavoronkov, A. Rapamycin in the context of Pascal's Wager: Generative pre-trained transformer perspective. *Oncoscience* **2022**, *9*, 82–84. [CrossRef] [PubMed]
23. Stokel-Walker, C. ChatGPT listed as author on research papers: Many scientists disapprove. *Nature* **2023**, *613*, 620–621. [CrossRef] [PubMed]
24. Else, H. Abstracts written by ChatGPT fool scientists. *Nature* **2023**, *613*, 423. [CrossRef] [PubMed]
25. Draxler, F.; Buschek, D.; Tavast, M.; Hämäläinen, P.; Schmidt, A.; Kulshrestha, J.; Welsch, R. Gender, age, and technology education influence the adoption and appropriation of LLMs. *arXiv* **2023**, arXiv:2310.06556. [CrossRef]
26. Aczel, B.; Wagenmakers, E.-J. Transparency guidance for ChatGPT usage in scientific writing. *PsyArXiv* **2023**. [CrossRef]
27. Tang, G. Letter to editor: Academic journals should clarify the proportion of NLP-generated content in papers. *Account. Res.* **2023**, 1–2. [CrossRef]
28. Thorp, H.H. ChatGPT is fun, but not an author. *Science* **2023**, *379*, 313. [CrossRef]
29. Nature. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature* **2023**, *613*, 612. [CrossRef] [PubMed]
30. Sage Publishing. ChatGPT and Generative AI. Available online: https://au.sagepub.com/en-gb/oce/chatgpt-and-generative-ai (accessed on 13 June 2024).
31. Shamoo, A.E.; Resnik, D.B. *Responsible Conduct of Research*; Oxford University Press: Oxford, UK, 2009.
32. Polonsky, M.J.; Rotman, J.D. Should Artificial Intelligent Agents be Your Co-author? Arguments in Favour, Informed by ChatGPT. *Australas. Mark. J.* **2023**, *31*, 91–96. [CrossRef]
33. Hosseini, M.; Resnik, D.B.; Holmes, K. The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. *Res. Ethics* **2023**, *19*, 449–465. [CrossRef]
34. Korkmaz, A.; Aktürk, C.; Talan, T. Analyzing the user's sentiments of ChatGPT using twitter data. *Iraqi J. Comput. Sci. Math.* **2023**, *4*, 202–214. [CrossRef]
35. Limna, P.; Kraiwanit, T.; Jangjarat, K.; Klayklung, P.; Chocksathaporn, P. The use of ChatGPT in the digital era: Perspectives on chatbot implementation. *J. Appl. Learn. Teach.* **2023**, *6*, 64–74.
36. Skjuve, M.; Følstad, A.; Brandtzaeg, P.B. The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. In *Proceedings of the 5th International Conference on Conversational User Interfaces, Eindhoven, The Netherlands, 19–21 July 2023*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–10.
37. Choudhury, A.; Shamszare, H. Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *J. Med. Internet Res.* **2023**, *25*, e47184. [CrossRef]
38. Dixit, A.; Jain, R. Chat of the Town: Gathering User Perception about ChatGPT. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4502004 (accessed on 13 June 2024).
39. Davis, F.D. Technology acceptance model: TAM. *Al-Suqri MN Al-Aufi AS Inf. Seek. Behav. Technol. Adopt.* **1989**, *205*, 219.
40. Niu, B.; Mvondo, G.F.N. I Am ChatGPT, the ultimate AI Chatbot! Investigating the determinants of users' loyalty and ethical usage concerns of ChatGPT. *J. Retail. Consum. Serv.* **2024**, *76*, 103562. [CrossRef]
41. Cress, U.; Kimmerle, J. Co-constructing knowledge with generative AI tools: Reflections from a CSCL perspective. *Int. J. Comput.-Support. Collab. Learn.* **2023**, *18*, 607–614. [CrossRef]
42. Ghajargar, M.; Bardzell, J.; Lagerkvist, L. A Redhead Walks into a Bar: Experiences of Writing Fiction with Artificial Intelligence. In Proceedings of the 25th International Academic Mindtrek Conference, Tampere, Finland, 16–18 November 2022; pp. 230–241.
43. Franke, T.; Attig, C.; Wessel, D. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *Int. J. Hum.–Comput. Interact.* **2019**, *35*, 456–467. [CrossRef]
44. Borsci, S.; Schmettow, M.; Malizia, A.; Chamberlain, A.; van der Velde, F. A confirmatory factorial analysis of the Chatbot Usability Scale: A multilanguage validation. *Pers. Ubiquitous Comput.* **2023**, *27*, 317–330. [CrossRef]
45. Madsen, M.; Gregor, S. Measuring human-computer trust. In Proceedings of the 11th Australasian Conference on Information Systems, Brisbane, Australia, 6–8 December 2000; pp. 6–8.
46. Carpinella, C.M.; Wyman, A.B.; Perez, M.A.; Stroessner, S.J. The Robotic Social Attributes Scale (RoSAS): Development and Validation. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, Vienna, Austria, 6–9 March 2017; pp. 254–262.
47. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
48. VERBI Software. MAXQDA 2022 Berlin, Germany: VERBI Software. Available online: https://www.maxqda.com/ (accessed on 30 January 2024).
49. Flesch, R. *How to Write Plain English: Let's Start with the Formula*; University of Canterbury: Christchurch, New Zealand, 1979.
50. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
51. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Kerrville, TX, USA, 2004; pp. 74–81.

52. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
53. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.
54. Rädiker, S.; Kuckartz, U. Intercoder-Übereinstimmung analysieren. In *Analyse qualitativer Daten mit MAXQDA: Text, Audio und Video*; Springer: Wiesbaden, Germany, 2019; pp. 287–303.
55. Brandt, M. ChatGPT Gefällt den Nutzer:innen. Available online: https://de.statista.com/infografik/29840/umfrage-zur-nutzung-von-ki-anwendungen-in-deutschland/ (accessed on 13 June 2024).
56. Lermann Henestrosa, A.; Kimmerle, J. Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany. *Behav. Sci.* **2024**, *14*, 353. [CrossRef] [PubMed]
57. Zhou, J.; Ke, P.; Qiu, X.; Huang, M.; Zhang, J. ChatGPT: Potential, prospects, and limitations. *Front. Inf. Technol. Electron. Eng.* **2023**, 1–6. [CrossRef]
58. Ayoub, N.F.; Lee, Y.-J.; Grimm, D.; Divi, V. Head-to-Head Comparison of ChatGPT Versus Google Search for Medical Knowledge Acquisition. *Otolaryngol.–Head Neck Surg.* **2024**, *170*, 1484–1491. [CrossRef] [PubMed]
59. Mogavi, R.H.; Deng, C.; Kim, J.J.; Zhou, P.; Kwon, Y.D.; Metwally, A.H.S.; Tlili, A.; Bassanelli, S.; Bucchiarone, A.; Gujar, S.; et al. ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. *Comput. Hum. Behav. Artif. Hum.* **2024**, *2*, 100027. [CrossRef]
60. Wagner, M.W.; Ertl-Wagner, B.B. Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Can. Assoc. Radiol. J.* **2023**, *75*, 08465371231171125. [CrossRef]
61. Xu, R.; Feng, Y.; Chen, H. Chatgpt vs. google: A comparative study of search performance and user experience. *arXiv* **2023**, arXiv:2307.01135. [CrossRef]
62. Latané, B.; Williams, K.; Harkins, S. Many hands make light the work: The causes and consequences of social loafing. *J. Pers. Soc. Psychol.* **1979**, *37*, 822–832. [CrossRef]
63. Alawida, M.; Mejri, S.; Mehmood, A.; Chikhaoui, B.; Isaac Abiodun, O. A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity. *Information* **2023**, *14*, 462. [CrossRef]
64. Krupp, L.; Steinert, S.; Kiefer-Emmanouilidis, M.; Avila, K.E.; Lukowicz, P.; Kuhn, J.; Küchemann, S.; Karolus, J. Unreflected acceptance–investigating the negative consequences of chatgpt-assisted problem solving in physics education. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*; IOS Press: Amsterdam, The Netherlands, 2024; pp. 199–212.
65. Anderl, C.; Klein, S.H.; Sarigül, B.; Schneider, F.M.; Han, J.; Fiedler, P.L.; Utz, S. Conversational presentation mode increases credibility judgements during information search with ChatGPT. *Sci. Rep.* **2024**, *14*, 17127. [CrossRef] [PubMed]
66. McKee, K.R.; Bai, X.; Fiske, S.T. Humans perceive warmth and competence in artificial intelligence. *iScience* **2023**, *26*, 107256. [CrossRef] [PubMed]
67. Abele, A.E.; Ellemers, N.; Fiske, S.T.; Koch, A.; Yzerbyt, V. Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychol. Rev.* **2021**, *128*, 290–314. [CrossRef]
68. Theophilou, E.; Koyutürk, C.; Yavari, M.; Bursic, S.; Donabauer, G.; Telari, A.; Testa, A.; Boiano, R.; Hernandez-Leo, D.; Ruskov, M.; et al. Learning to Prompt in the Classroom to Understand AI Limits: A Pilot Study. In Proceedings of the AIxIA 2023—Advances in Artificial Intelligence, Rome, Italy, 6–9 November 2023; Springer: Cham, Switzerland, 2023; pp. 481–496.
69. Backhaus, J.; Huth, K.; Entwistle, A.; Homayounfar, K.; Koenig, S. Digital Affinity in Medical Students Influences Learning Outcome: A Cluster Analytical Design Comparing Vodcast with Traditional Lecture. *J. Surg. Educ.* **2019**, *76*, 711–719. [CrossRef] [PubMed]
70. Urchs, S.; Thurner, V.; Aßenmacher, M.; Heumann, C.; Thiemichen, S. How Prevalent is Gender Bias in ChatGPT?-Exploring German and English ChatGPT Responses. *arXiv* **2023**, arXiv:2310.03031.
71. Proksch, S.; Schühle, J.; Streeb, E.; Weymann, F.; Luther, T.; Kimmerle, J. The impact of text topic and assumed human vs. AI authorship on competence and quality assessment. *Front. Artif. Intell.* **2024**, *7*, 1412710. [CrossRef]
72. Lermann Henestrosa, A.; Greving, H.; Kimmerle, J. Automated journalism: The effects of AI authorship and evaluative information on the perception of a science journalism article. *Comput. Hum. Behav.* **2023**, *138*, 107445. [CrossRef]
73. Lermann Henestrosa, A.; Kimmerle, J. The Effects of Assumed AI vs. Human Authorship on the Perception of a GPT-Generated Text. 2024. Available online: https://osf.io/preprints/psyarxiv/wrusc (accessed on 13 June 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.