Analyzing people`s awareness of the difference between human text and artificial intelligence generated texts(ChatGPT 4.0).

Sultan Daniyarov, Diyar Zhumataev

Introduction

Background information

Artificial Intelligence (AI) has recently become one of the most essential parts in people's lives, allowing individuals to use AI`s multifunctionality in various areas, from education to creative industries. One of the most famous, advanced features of AI is its ability to generate texts, the function that is expected to boost an individual's productivity by allowing them to create essays, articles, and even stories in a matter of a second.

Problem statement

However, this feature is also being used in ways that raise concerns. For example, there are increasingly more situations when students would present AI-generated work as their own, or even some cases when AI-generated texts are used to mislead or deceive people by creating fake content that appears to be written by a real person.

Therefore, It is crucial to understand how well people can recognize AI generated texts. Understanding the level of AI recognition among individuals would explain what impact factors, like age, gender, or even AI familiarity have on these recognition levels.

Purpose

The aim of this research is to explore the general awareness of AI-generated texts among people. We speculate factors such as specialization, age, gender or even AI familiarity will have a notable impact on one`s Ai recognition level. We also think that people`s AI recognition level would vary in accordance with the text`s style, more precisely, we hypothesize that people would have a better performance when the style of the text is closer to casual message exchange(e.g Whatsapp), whereas the worst performance would be when participants are exposed to formal texts. Such findings in these fields are expected because casual message exchange is an area that all people experience every day, whereas the formal language is the main language, using which AI generates its texts.

Significance of the research

Conducting research papers on this topic is extremely crucial, because it has the potential to help people adapt to the increasing use of AI-generated content. By identifying how well people can recognize texts created by AI technologies, the findings of this research may help to develop some key strategies, or even tools to potentially raise the level of AI recognition among people, thereby lowering the possibility of being deceived by AI.

Literature review

Introduction

The proliferation of artificial intelligence (AI) technologies has led to the development of highly sophisticated systems capable of generating content that closely mimics human behavior across various domains, including natural language processing, visual art, and music composition. As the AI systems become more advanced, the ability of humans to accurately

differentiate between AI-generated and human-generated content has emerged as a critical area of scholarly inquiry. Nevertheless, a distinct effect of how participants` background influences the ability to differentiate bots from humans was not deeply studied. This research paper explores the ability of 11-64 year old people across the territory of Kazakhstan to identify a bot from humans, by correlating their responses in dependence of their age, lifestyle and occupation. By investigating these findings, this study aims to figure out whether people's cultural, age, gender background could influence their ability to differentiate artificial intelligence (AI)-generated texts from the human-created ones in various text styles. This research aims to highlight the growing interest in using artificial intelligence for an interaction with humans and lift an awareness among the people that AI may be used this way.

Identification of the research gaps

Although the current studies confirm that AI can be distinguished from human beings by requesting certain types of questions, many inconclusive details, like how participants` age, demographic affects their ability to determine an AI are still unknown. For instance, in a study of Wang H., et al researchers determined several areas that people are comfortable and that AI finds complicated to work with by examining 10 participants with the age distribution from 10 to 50 years old. A different approach was taken in the research of Jannai, D., et al: researchers used a gamified approach, in which participants had to distinguish AI from humans in small chat correspondences. This research statistically determines the best strategies used in this game by analyzing over 32000 conversations of unknown participants. Hence, researchers generally did not consider statistics about participants` background in their studies. Thus, the lack of data on age and demographic effects on the ability to differentiate artificial intelligence from human underscores the need to do further exploration into the topic and find the most effective and long-term way to distinguish AI from humans

In a study by Stock-Homburg, R.M. (2023), participants were asked to differentiate between innovations created by humans and those generated by AI, using a Turing Test approach. The findings showed that while participants could sometimes tell the difference, their judgments were often swayed by personal biases and the inherently subjective nature of evaluating innovation. The study also highlighted some limitations, including a lack of diversity among participants and an inadequate exploration of these biases. Additionally, the research didn’t account for how transparency in AI processes or the specific context of the innovations might influence perceptions.

In 2024 Jones, C.R., & Bergen, B.K research, a Turing test was conducted in which participants tried to distinguish GPT-4 from human-generated text. Most participants had difficulty identifying which responses were coming from the AI. However, the study only included brief exchanges, leaving a gap in understanding how well GPT-4 can support human responses in longer, more complex conversations.

In the paperwork of Hamada, M., et al, researchers studied how artificial intelligence (AI) could improve business efficiency and effectiveness for enterprises in Kazakhstan. They found that AI significantly enhanced operational processes. However, the study mainly focused on large enterprises, leaving a gap in understanding AI's impact on small and medium-sized businesses in the region.

In a 2024 study by Stratchan, J.W., researchers tested theory of mind on large language models and humans, assessing their ability to understand the thoughts and intentions of others.

The results showed that although the models showed some level of this understanding, they still lagged behind humans. However, research has focused primarily on simple scenarios, leaving a gap in understanding how these models work in more complex real-world situations.

Awareness and Perception of AI in Text Generation (percentage/numbers)

Understanding people's awareness and perception of AI in text generation is a key factor in determining efficient strategies of distinguishing AI from humans. In a qualitative study examining different AI models` features, Wang, H., et al (2023) proposed a new framework, called FLAIR for detecting AI in a single question. In order to identify efficient questions for FlAIR, researchers examined people and different AI models by giving them a series of questions. As an outcome of the research, it was reported that bots are not good at symbolic manipulation, randomness, and graphical understanding, the areas in which humans had scored perfectly. In a similar study of Jannai, D., et al (2023), researchers tested human capabilities to detect an AI. Overall, the research shows that the overall probability of correctly guessing whether the partner is a bot or human is 68%. The study also proposes strategies people used to guess correctly: Grammatical errors and typos, personal questions, hard requests, and etc. AI was also found out to be easily detectable if subjected to questions on Faux pas in a paper work **of** Stratchan, J.W., et al. In contrast, in the same study, Ai was found to be strong in questions on Irony, hunting and strange stories. Additionally, in a study conducted by Jones, C.R., & Bergen, B.K. (2024), the probability of correctly guessing if the partner is a bot or not was shown to depend on what AI models were used. For example: GPT-4 achieved a pass rate of 54%, outperforming GPT-3.5 (50%) and the ELIZA baseline (22%). These researches mainly show that people's awareness of AI text generation still prevails AI`s trials to disguise itself as humans, as people still were able to see a difference between a bot partner and a human partner.

In the study of Stock-Homburg, R.M., et al(2023), researchers conducted a Turing test to examine the balance between human and artificial intelligence innovation. Participants were presented with a series of ideas and asked to identify which ones were created by humans and which were created by artificial intelligence. The study found that 52% of respondents accurately distinguished between human and AI-generated ideas. These results highlight the challenge of differentiating AI and human contributions to creative fields.

In the paperwork of Hamada, M., et al (2021), researchers explored how artificial intelligence could enhance business processes and effectiveness for enterprises in Kazakhstan. They found that 42% of Kazakhstani retailers were already integrating AI technologies into their operations, with another 35% planning to do so within the next five years. This trend indicates that by 2024, 77% of Kazakhstani retailers are expected to adopt AI solutions. These findings highlight the growing role of AI in improving efficiency and effectiveness in the retail sector.

Conclusion

By identifying the main areas in which AI could be effortlessly distinguished from humans and statistical evidence of people correctly guessing if the partner is an AI or not , this review shows that AI has a clear difference from humans and that human`s awareness of AI text generation is still on a high level.

However, this review also identifies an essential research gap: the need to take into account people's background to correctly identify AI`s chances to disguise as human. Tackling this gap is crucial to create new and optimal strategies to distinguish an AI from humans.

Methodology

Participants

Participants will be selected randomly through online platforms and social media to ensure that the survey will have a diverse group of respondents in terms of age, gender. The approximate sample size of the questionnaire will be about 90-100 individuals from the central Asia region.

Materials (Text selection)

The texts used in the survey will be carefully examined to create a fair mix between AI and human texts. AI texts will be generated via ChatGPT-4, and human texts will be selected from various online platforms. In accordance with the sections of the survey, Both human and ChatGPT-4 generated text will contain some possible errors (e.g., spelling, grammar, punctuation) to examine maximum potential of AI`s text disguise. Also, to make a survey more passable and comfortable for respondents, hence, increasing their seriousness when they would respond to the questions, we decided to shorten the time spent on each question by setting the following limitations to the texts:

1) the texts will have a maximum of 3 sentences
2) Language which is used to generate text is russian, since our audience is russian speaking

Research Design

This study will be conducted using a survey as a main source of data regarding people`s AI recognition level. The survey will be conducted through Google Forms, allowing for a wide-distribution, adequate sample size, and ease of data collection.

Data Collection

The survey will be created on Google Forms platform, allowing responses to be automatically collected and stored for analysis in tables and diagrams. The survey will remain open for a set period to obtain an adequate sample size.

Survey Structure

The survey will consist of five main parts:

In the first part, respondents will be asked to provide basic demographic details such as age, specialization, gender type and experience with AI technologies, which will help in a further analysis.

From part 2 to part 4 of the questionnaire**,** respondents will be presented with a series of questions, each of which would have both AI and human created texts. In each question, participants will have to determine which text was created by AI and human respectively.

In part 2, participants will be presented 2 questions that were created to resemble everyday communication via online messengers(e.g. Whatsapp). The texts within this part of the survey will contain the most amount of errors.

In part 3, participants will also be presented with 2 questions, however, the texts will be created and selected to resemble a post in social media(e.g reddit). The texts within this part of the questionnaire will be designed to contain less amount of errors than in part 2 of the survey.

In part 4 of the survey, respondents will face formal language texts(e.g formal letters). The texts in this part will be designed to have no possible errors (e.g., spelling, grammar, punctuation).

In part 5 of the questionnaire, respondents will be asked to share their methods of identifying the AI-generated content from human created content.

Data Analysis

Data will be analyzed using statistics to determine the overall success rate of identifying AI-generated texts. Additionally, demographic information will also be a subject of analysis to determine if certain factors (e.g., age, education level) have an impact on the respondents' ability to identify AI-generated texts. Also, methods applied by participants to identify AI will also be analyzed to find the most common and efficient strategy to identify whether the text was created by a human or not.

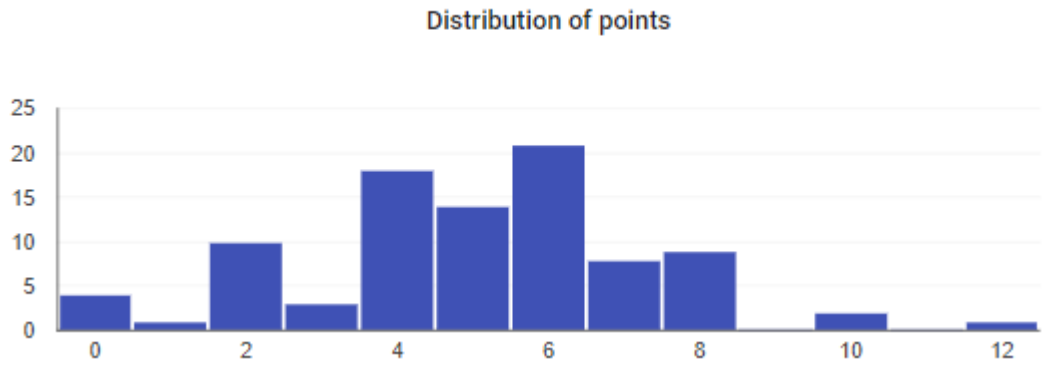
Results & Discussion

The findings from the survey responses are summarized in this section, revealing the level of awareness and recognition accuracy of AI-generated texts; some analysis will also be provided in this section.

General Perception/Familiarity of respondents with AI/chatGPT 4

According to the questionnaire responses, Figure 1 shows that participants were able to recognize AI-generated text with an accuracy of about 41.92% , which is notably lower than the findings in the Jones, C.R., & Bergen, B.K. (2024) study, in which the probability of correctly guessing whether the partner is a bot or not for specifically GPT-4 model was 54% . We suggest that such a difference may appear because of two possible reasons. One possible reason is that the studies were in different geographical regions, suggesting that individuals` location is also one of the factors influencing people's recognition level of AI-generated texts. Another possible reason is that the previous study focused on a specific age group, so differences in age might also explain why the recognition rates vary between studies.

Figure 1



Note. Distribution of points.

Gender factor

Interestingly, AI recognition rate showed a negligible dependence on people's gender type(5.5 %). Only 41.38% of women scored above the overall mean score of 41.92%, whereas the same statistics for male participants were 46.88%, showing almost no difference between two categories.

Age factor

As it can be seen on the trend line of figure 2 that shows how the number of points respondents get depends on the age of the participant, a clear dependency between two variables exists. The trend line given in figure 2 depicts that the lower the participant`s age the better he or she would perform in identifying AI-generated texts from human created ones. We suggest that such an outcome occurred because younger people are generally more acquainted with AI technologies than older ones.

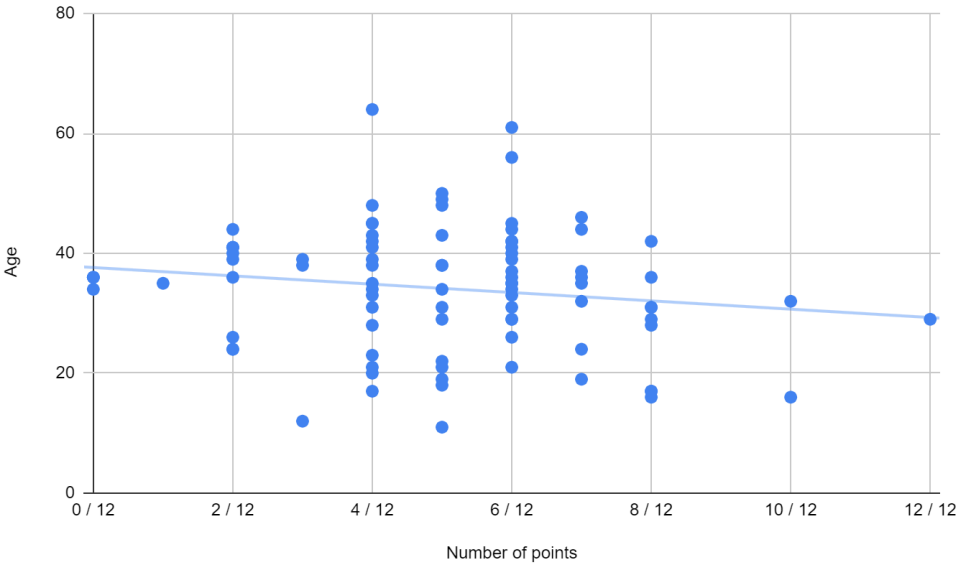


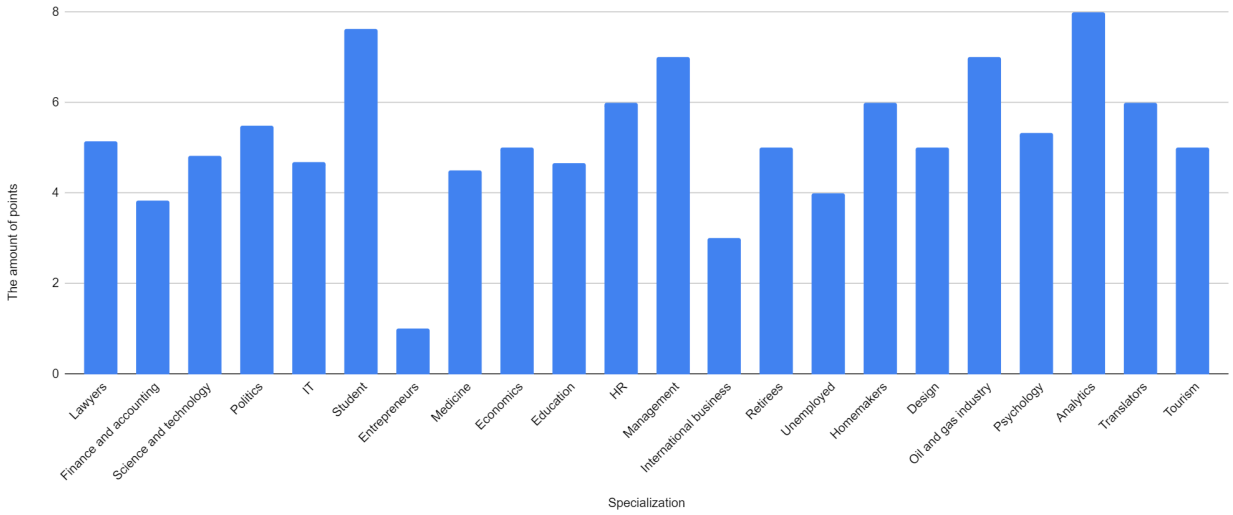
Figure 2

Note. Correlation between the Age group (y-axis) of participants and Number of points (x-axis) they received on the Test.

Specialization

As it can be seen on Fig. 7, respondents` specialization has slight or no impact on the amount of points earned by them, since the majority of specialization earn around the same amount of points.

Figure 7

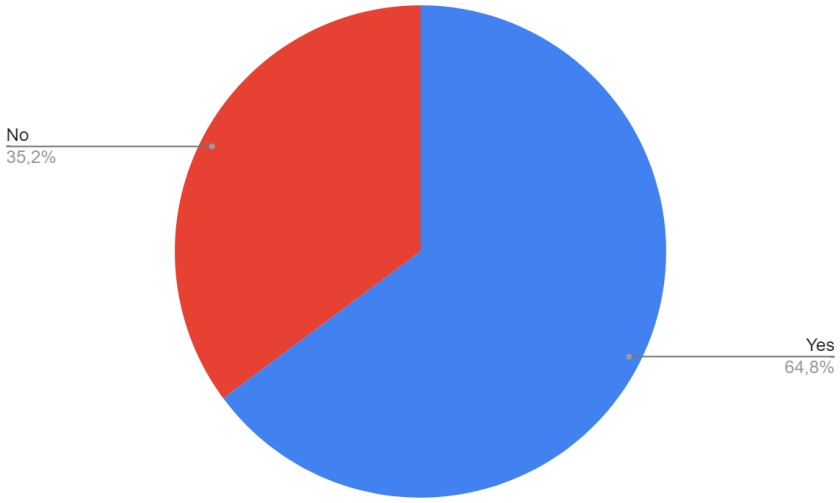


Note. Correlation between the Specialization (occupation) type of the respondents and the Amount of Points they received on the Test.

The general familiarity with AI

The data from Fig.5 shows respondents`s own evaluation of their familiarity with AI technologies. Surprisingly, we found no difference between AI recognition rate (Fig. 6) of people who think that they are not familiar with AI and people who think that they are familiar with AI.

Figure 5



Note. Respondent`s evaluation of their AI familiarity.

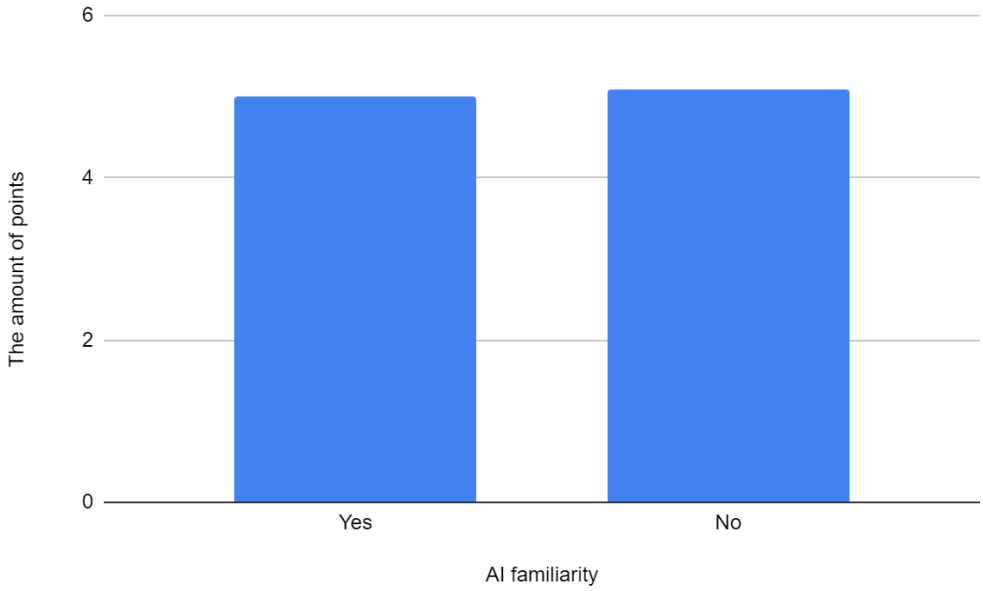


Figure 6

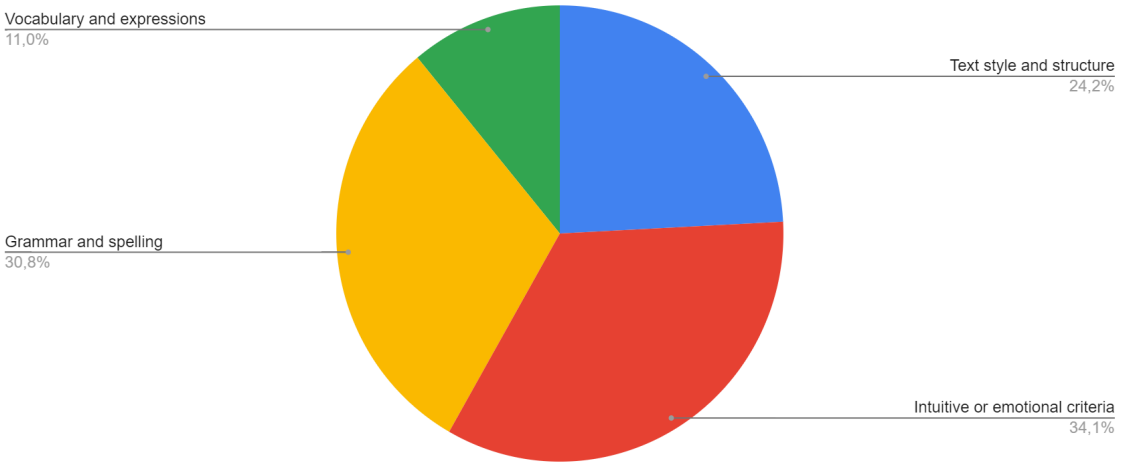
Note. The correlation between the amount of points respondents received on a test (y-axis) and their AI familiarity (x-axis).

Methods used for identification

There were various methods that respondents would use for AI text identification purposes. Among the applied methods (Fig.3) were “Intuitive or emotional criteria” (34,1%), “Text style and structure” (24,2%) and “Grammar and spelling” (30,8%), “Vocabulary and expressions” (11,0%). We speculate that the most applied method “Intuitive or emotional criteria” obtained such a wide usage in our survey because according to the data from Fig.5, there is a huge number of people who do not know what AI technologies are, thus, are not aware of the basic principles of AI text generating.

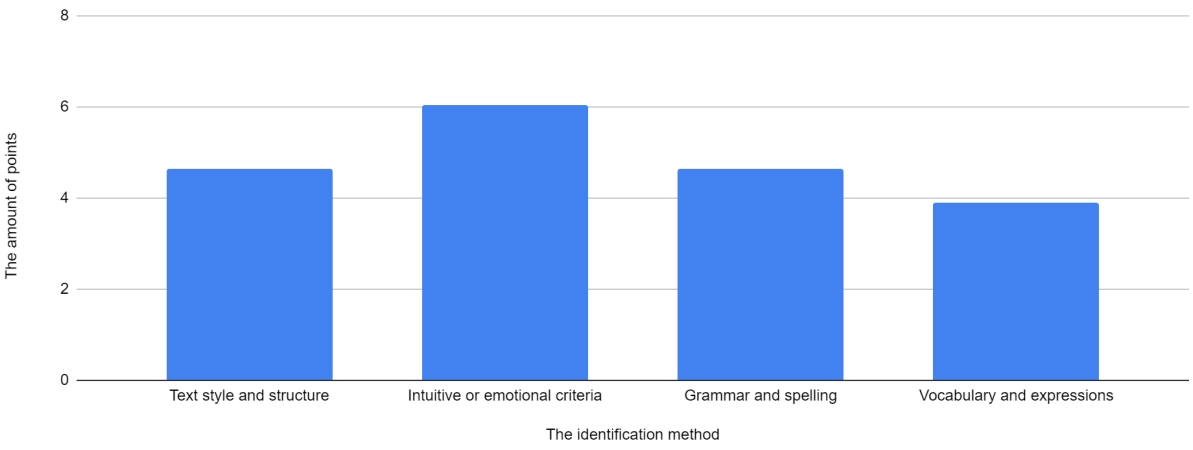
Interestingly, those who used the “Intuitive or emotional criteria” method achieved the best overall score among others (Fig. 4). The possible reason for that is our approach to generate more complex, human-like text using AI, since we purposefully gave ChatGPT-4 a task to make human-like texts that may contain several mistakes. Therefore, this result may show that there is a connection between the success of the method and the method used to generate questions.

Figure 3. The proportion of methods used



Note. The proportion of methods used.

Figure 4



Note. Correlation between the number of points participants received on the test (y-axis) and the identification method they used (x-axis).

Sections of the Survey

Surprisingly, our hypothesis about correct answer distribution across three sections of the survey was wrong.

The first section of the questionnaire was the casual message exchange sector, which is similar to communication through various messengers (e.g. WhatsApp, Instagram, etc), and it was hypothesized to accumulate the most amount of correct answer choices. By the end of the survey, however, this section accumulated only 158 out of 364 correct answer choices, taking the second place among all three sections.

The second section of the survey was the post in the social media sector that contained questions similar to those that are asked on websites for opinion sharing (e.g Reddit). According to our hypothesis, this section was supposed to take second place of correct responses on the Test. Nevertheless, the least amount of correct answers was observed in this section (135 out of 364 correct answers).

The third section contained formal communication texts, which was hypothesized to have the least amount of correct answer choices. However, it was the most correctly answered sector (165 out of 364 correct answers).

Overall, these results are worrisome, since, as it was found out, both familiarity with an AI and the level of AI text recognition is low, creating a notable possibility that individuals would not be able to distinguish AI from humans. This may potentially hold unpleasant consequences for these individuals, because more scams through AI usage are reported from day to day in various fields(e.g. education, finances, creative industries).

Conclusion

Key findings

Our research identified several key findings that may be of a high interest for those who would foster the following development of this topic. First, not all factors showed a clear impact on people`s AI recognition. For example, while age has a clear impact on people's perception of AI created texts, slight or no impact of factors such as general familiarity with AI and gender were found on people`s AI recognition level. Some key methods of how to distinguish human texts from AI-generated texts were found out, and among them the most successful strategy was “Intuitive or emotional criteria”. Thus, we believe that AI-generated texts could be distinguished from human texts by understanding how either AI or humans develop their ideas and topics in their texts.

Implications

Given findings may be implemented into special apps or websites that would automatically identify whether the text was written by a human or not. As well as implementing them in apps, this research`s findings may be used in a further comparison between the findings

of the other researches, revealing more how each participant`s characteristic may influence their level of AI perception.

Final remarks

Finally, it is highly recommended by us to examine further how each factor impacts people's ability to identify whether the text was created by humans or not, maintaining focus on Diverse Demographics to understand more about how awareness of AI-generated content varies across different segments of the population.

References

Wang, H., Luo, X., Wang, W., & Yan, X. (2023). Bot or Human? Detecting ChatGPT Imposters with A Single Question. *ArXiv, abs/2305.06424*.

Jones, C.R., & Bergen, B.K. (2024). People cannot distinguish GPT-4 from a human in a Turing test. *ArXiv*, abs/2405.08007.

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023). Human or Not? A Gamified Approach to the Turing Test. *ArXiv*, abs/2305.20010.

Strachan, J.W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M.S., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8, 1285 - 1295.

Stock-Homburg, R.M. (2023). The Tightrope Between Human and AI-Generated Innovation: A Turing Test. *SSRN Electronic Journal*.

Hamada, M., Temirkhanova, D., Serikbay, D., Salybekov, S., & Omarbek, S. (2021). Artificial Intelligence to Improve the Business Efficiency and Effectiveness for Enterprises in Kazakhstan. *SAR Journal - Science and Research*.