Identifying Whether a Short Essay was written by a University Student or ChatGPT

Christopher Saarna
Asia University, Japan

www.ijte.net

To cite this article:

Saarna, C. (2024). Identifying whether a short essay was written by a university student or ChatGPT. *International Journal of Technology in Education (IJTE)*, *7*(3), 611-633. https://doi.org/10.46328/ijte.773

International Journal of Technology in Education
2024, Vol. 7, No. 3, 611-633 https://doi.org/10.46328/ijte.773

Identifying Whether a Short Essay was written by a University Student or ChatGPT

Christopher Saarna

Article Info

Article History

Received:
01 February 2024
Accepted:
24 May 2024

Keywords

ChatGPT
Academic dishonesty
EFL
Short essay
Detection tools

Abstract

This study seeks to clarify whether teachers are able to distinguish between essays written by English L2 students or generated by ChatGPT. 47 instructors who hold experience teaching English to native speakers of Japanese in universities or other higher education institutions were tested on whether they could identify between human written essays and ChatGPT generated essays. The ICNALE written corpus (Ishikawa, 2013) was used to find and randomly select the essays of four Japanese university students' written work who studied English at roughly CEFR A2 level. The AI chatbot, ChatGPT, was used to generate four essays utilizing prompts which directed the chatbot to mimic grammar mistakes common to nonnative speakers of English. Teachers were requested to identify which of the eight essays they believed to be human written or ChatGPT generated. On average, the teachers were able to identify 54.25% of items accurately. This result is slightly better than random chance, and implies that most teachers cannot make an accurate assessment on a ChatGPT generated essay when ChatGPT is prompted to make grammar mistakes.

Introduction

It is no secret that AI chatbots have taken the world by storm in the past year. In just two months, many people went from not knowing what a chatbot is, to using chatbots regularly, with ChatGPT having over 100 million active users just two months after public release (Hu, 2023). Included among those who are starting to get acquainted with chatbots are the teachers and students within academic institutions. However, in the increasingly technologically advanced world of today, educators face many challenges associated with teaching in the age of AI chatbots. Among the challenges listed by ChatGPT itself are - the potential for plagiarism, misinformation, over-dependency, privacy, and the need for training and familiarity (OpenAI, 2024). This paper sets out to tackle one of the primary concerns many educators have about ChatGPT, which is, the identification of writing produced by ChatGPT which could be used to misrepresent students' assignments and compositions.

Literature Review

AI Chatbots and ChatGPT

Some form of AI chatbot has been around since the 1960s when ELIZA, the forefather of modern chatbots, was

created by Joseph Weizenbaum (Zemcik, 2019; Berry, 2023). Eliza was a simple conversational agent capable of asking questions to and answering the user. Later chatbots such as PARRY developed in 1972, and Dr. Sabaitso developed in 1991 introduced new functions such as inducing more complicated discussions and pioneering audio output by synthesizing the speech of the AI chatbot (Zemcik, 2019).

However, it was only in November 2022 when OpenAI made ChatGPT-3 available to the public that AI chatbots entered the public consciousness (Berry, 2023). ChatGPT is an AI chatbot trained on Large Language Models (LLMs). This includes a vast library of real conversations, journals, books, compositions, and almost any type of text publicly available online (Berry, 2023). ChatGPT is capable of producing text extremely quickly, efficiently, and with a very natural tone making it the most advanced and impressive AI chatbot currently available (Rudolph, et al., 2023).

EFL Essay Writing and Academic Dishonesty

With the ability to create lengthy amounts of human-like text within seconds, there has been a large amount of public discourse regarding ChatGPT and its future implications on many different facets of society (Rudolph, et al., 2023). While ChatGPT has many positive attributes such as the wide array of functions and knowledge that can be made available at our fingertips, there are also some potential pitfalls which should be addressed. Unfortunately, the current author is unqualified to make accurate assessments about the majority of fields and areas of society that ChatGPT and AI chatbots may affect in the near future. However, one area that educators in the field of language teaching in particular should observe closely is that of teaching writing.

Traditionally, the most common form of writing assessment for ESL and EFL students has been the short essay. According to Cheng & Wang (2007) many instructors reported assessment by long essays, student journals and portfolios, and editing of sentences and paragraphs. However, short essays were the most common form of writing assessment and were utilized by 91.9% of Canadian ESL instructors, 86% of Hong Kong EFL instructors, and 88.6% of Beijing EFL instructors (Cheng & Wang, 2007).

Despite the common utilization of short essay writing in EFL and ESL environments, teachers should be aware of the potential for academic dishonesty in writing in many academic societies. Even prior to recent great advancements in technology, academic dishonesty has always been prevalent in academia. McCabe and Trevino (1996) surveyed 6000 students from 31 US college campuses and found that 84% of students have engaged in an act of academic dishonesty on a written assignment at least once in their academic careers. The most popular forms of academic dishonesty included copying material without footnoting (54%), plagiarism (26%), falsified bibliography (29%), and collaboration on individual work (49%). It was also found that rates of academic dishonesty were committed at remarkably similar rates in 1963 and 1996 (McCabe & Trevino, 1996).

In Japan rates of academic dishonesty as assessed by a self-reported, anonymous survey conducted at nine Japanese universities were found to be slightly less than in the US with 64% of Japanese students admitting to acts of academic dishonesty (Carlson, 2020). Carlson found that the most prevalent types of academic dishonesty

among Japan university students were accepting a group grade without doing adequate work (31.64%), copy and pasting translations from Google translate or DeepL directly into an assignment (31.25%), and copying material without citations (19.92%).

Student Usage of ChatGPT

According to a survey of 1,223 college students conducted by the website intelligent.com in May 2023, 30% of college students in the US acknowledged using ChatGPT to aid in their academic assignments. 46% of students admitted to using ChatGPT somewhat frequently or very frequently for their academic studies. Among all subjects, ChatGPT was most commonly used to help students with English assignments (Intelligent.com, 2023).

Additionally, an anonymous survey of 6,300 German students found that 63.4% of students have used AI tools such as ChatGPT for academic purposes (von Garrel & Mayer, 2023). A study of 200 Vietnamese university students showed that 72.5% of students had experience using ChatGPT for academic purposes (Ngo, 2023). Finally, a survey of 71 Hungarian high school students found that 57.7% of students used ChatGPT every day, 28.2% reported using it a few times a week, and 14.1% of students used ChatGPT less than once a week (Forman, et al., 2023). Many students in these surveys stated that using chatbots and ChatGPT for academic purposes was useful for saving time on assignments, helpful with idea generation, provided better understanding of complex subjects, and was convenient for translation into different languages (Forman et al., 2023; von Garrel and Mayer, 2023; Ngo, 2023). With so many students engaging in frequent use of ChatGPT and a prevalent culture of academic dishonesty, it can be theorized that a significant portion of students may find it tempting to use ChatGPT in ways that may be considered inappropriate.

High Number of Students for Part-time Teachers

Another point of consideration that must be accounted for is the high number of students in many university classes. This is especially true for teachers who teach many university courses across a range of campuses. According to OECD (2014) data, Japan has one of the highest average classroom sizes in the OECD. Elementary school classrooms average 27 students per class and middle schools average 32 students per class in Japan.

Little modern data could be found on university class sizes in Japan. However, one study conducted at Nagoya University of Foreign Studies found that the average EFL class had 15-25 students, although the majority of students believed the ideal class size to be fewer than 12 students per class (Mortali, 2023). Another study found that the average class size for EFL communication classes at Takushoku University was 25 students (Arai, 2017). With an average of 15-25 students per class, a university teacher with a workload of ten classes could expect to receive assignments from between 150 and 250 students. In the author's personal experience teaching at six different universities, most EFL university classes in Japan have between 20 and 35 students.

Strong (2023) states that, according to OECD estimates, about 40% of university faculty in 38 countries are considered part-time contractors. These rates increase to 50% in America and 60% of faculty members in Japan

(Strong, 2023). Additionally, part-time teachers in Japan face more insecure situations with regards to job security and generally pay higher gross amounts for insurance and pension because of the lack of benefit provisions from their workplaces (Nagatomo, 2014). This could lead to many part-time teachers taking on a high number of classes and thus a large number of students in order to make ends meet.

Indeed, the current author informally surveyed eight part-time teachers he was acquainted with and asked them roughly how many university students each taught. It was found that the eight part-time teachers claimed to teach on average 316 university students per semester. This is on top of other administrative duties and part-time teaching at non-university institutions of learning. Therefore, it can be reasonably surmised that most part-time teachers don't have the luxury of looking through assignments and essays thoroughly. This not only leads to a lack of time for giving corrective feedback on written work but may also lead to difficulty in detection of academic dishonesty. Despite this lack of time for providing feedback, 91.7% of Canadian and Chinese university teachers gave feedback on written assignments according to Cheng & Wang (2007). The majority of teachers were able to provide this feedback within one or two weeks. However, providing written feedback to hundreds of students across a multitude of campuses and academic programs would seem a challenging and time consuming endeavor for even the most proficient part-time instructors.

Detection of Academic Dishonesty and Time Constraints

Providing feedback to hundreds of students would presumably take up a large portion of time for instructors. For an average part time instructor in Japan who might teach 316 students on a part time basis, dedicating even a minimal time of four minutes per short essay to provide feedback would result in 21 hours of extra work. Additionally, teachers in this age of AI chatbots must deal with identification and detection issues related to increasingly sophisticated methods of academic dishonesty (Cotton et al., 2024). This poses the question of whether instructors are up to the task of spending the time and mental resources necessary to provide students with adequate amounts of feedback, keeping up to date with technological advances in AI chatbots, and checking for potential academic dishonesty in written essays. In this paper, the author decided to focus on identification issues concerning ChatGPT generated essays.

Research Question

Research Question 1 - Can teachers with EFL experience in Japanese higher education identify which of eight essays were generated by ChatGPT and which were written by CEFR A2 level Japanese university students?
Research Question 2 - Did teaching experience and position have an effect on identification ability of ChatGPT vs Japanese university student written essays?

Methods

ICNALE Corpus

For this project, the author chose to utilize the publicly available International Corpus Network of Asian Learners

of English (ICNALE) first organized by Dr. Shin Ishikawa in 2013 which has continued development to this day. The ICNALE database is a collection of several thousand essays, spoken monologues, and spoken dialogs which can be used to research written work and spoken communication produced by Asian learners of English at universities in ten different Asian countries. One of the key advantages of using this database is that it utilizes uniformity for each task. For written essays, all students were required to express their opinions in at least 200 words on the two following topics: *It is important for college students to have a part-time job* and *Smoking should be completely banned at all the restaurants in the country*. The students were given 20-40 minutes to write each short essay on Microsoft Word. Students were allowed the use of a spell checker but were prohibited from using any other references when writing the essays. Additionally, all participants in the ICNALE database have been classified into approximate CEFR levels via implementation of L2 vocabulary size tests, TOEIC test scores, and TOEFL test scores (Ishikawa, 2013).

For this particular research paper, the author decided to filter for essays written by Japanese university students studying English at approximately a CEFR level of A2. This produced a total of 154 written essays. A random number generator sourced from randomnumbergenerator.com was subsequently used to select four authentic essays composed by Japanese learners of English at an A2 CEFR level. These essays will henceforth be referred to as the *Human Essays* in this research paper.

AI Generated Essays

The author used ChatGPT to generate four essays that were designed to imitate English L2 university students' writing. The four essays generated by ChatGPT will henceforth be known as the *ChatGPT Essays* when referred to in this paper. The four prompts used to generate the ChatGPT responses were as follows:

- *Please write a 250 word essay arguing against "It is important for college students to have a part-time job". Please write as if you are a non native English speaking university student and make mistakes typical of a low level student but not spelling mistakes*
- *Please write a 250 word essay arguing against "Smoking should be completely banned at all the restaurants in the country". Please write as if you are a non native English speaking university student and make mistakes typical of a low level student but not spelling mistakes. (ChatGPT writes)*
  Can you try again but make more mistakes
- *Smoking should be probihited in all restaurants in the country というテーマで250語のエッセイを書いてください。英語を母国語としない大学生のつもりで、低レベルの学生にありがちなミスをしないように書いてください（スペルミスは不可）。(ChatGPT writes)*

もう一度書いてください。でも、もっと間違いがあって、とても低いレベルの英語で。(ChatGPT writes again)

もう一度書いてください。でも、もっと間違いがあって、とても下手な留学生見たに書いてください。

- *Please write a 250 word essay on the topic "It is important for college students to have a part-time job". Please write as if you are a not native English speaking university student and make mistakes typical of*

a low level student but not spelling mistakes

The author put all eight essays into OpenAI's AI detector application coined ZeroGPT and found that the ChatGPT generated essays exhibited 93%, 100%, 92%, and 97% likelihood of being human written. All human written essays from the ICNALE database exhibited a 100% chance of being human written. This is in stark contrast to Gao et al. (2023) who experienced great success in detecting Chatgpt generated essays with the GPT-2 Output Detector tool. The author also put all eight essays into OpenAI's GPT-2 Output Detector tool and found that all eight essays received a 99.9% rating of *REAL,* indicating that the detection tools provided by OpenAI were insufficient in detecting essays when ChatGPT was prompted to make grammatical errors (see Appendix A for the eight essay samples used in the identification items for this study.)

Survey and Dissemination to University Teachers

The author then used the survey platform Jotform.com and placed the four human and four ChatGPT essays inside in random order, again using a random number generator. The reason why Jotform.com was preferred over other platforms such as surveymonkey.com and Google forms is because it was much easier for the author to implement a time limit on Jotform.com. This timer was applied as an added obstacle for teachers in order to simulate the condition of reviewing a number of similarly themed essays while under the burden of time constraints. The author decided to give a limit of 32 minutes (approximately four minutes per essay) plus one extra minute to give details about the participants' teaching history in Japanese institutions of higher education. This consisted of information about current work position, years of experience teaching non-native speakers of English, and native language. In total 33 minutes were allotted to answer the three work experience related questions and eight Human or ChatGPT identification items. Please see Appendix A to view the survey in full.

Participants

Starting on January 19th, 2024 the author began to contact current and former colleagues via email. The author also decided to enlist the help of teachers on the private Facebook group *Online Teachers Japan.* Participants were asked to read the following instructions before clicking on the link and beginning the survey.

My project is basically on whether teachers can distinguish between essays written by ChatGPT and Japanese L2 English learners.

Anyone who has taught in a Japanese university or institution of higher learning may participate. No emails or personal information will be collected beyond general work experience so your privacy will be respected. However you may want to take note of your answers or the date and time when you finish because you may forget your answers before I publicly post the survey results.

The purpose of the present research is to determine whether university instructors have the ability to identify whether a short English essay was written by a human university student (native language

Japanese) or by AI in a timely manner. Please read the writing samples below and try to determine whether they were written by a human or ChatGPT. There are 8 samples in total. All human written samples were taken from a corpora of CEFR A2 level Japanese students. The students wrote using MS word and were allowed to use spell checkers but no dictionaries or other references. They were asked for written responses to the following two prompts: "It is important for college students to have a part-time job" and "Smoking should be completely banned at all the restaurants in the country". Please try to determine whether each response was written by ChatGPT or by L1 Japanese learners of English. You are welcome to use any online tools including zerogpt, ChatGPT, google translate, etc. to try to help identify which short essays are human and which are AI generated.

Please note that the total time given to complete the survey is 33 minutes. This is made to simulate conditions for part time instructors who make up the majority of university instructors of English in Japan (A survey of 8 part time instructors of English in Japanese universities found that the average part time instructor had a mean of 316.6 students plus additional work). So please only click the link to the survey when you have 33 minutes of free time to focus on the survey.

Finally if you have any questions or feedback on the survey please let me know in the comments or send me a message. Any discussion about AI and essay writing is most welcome. Also if you happen to get a high score please tell me so that I can learn your ways!

Thank you in advance for your help. It is greatly appreciated.

All responses were taken anonymously in order to respect the privacy of participants but the author welcomed emails, messages, and comments about the project. The author chose not to publish results or the correct answers to the survey until data collection was completed on February 11th, 2024. In total, 47 participants who claimed to have experience teaching in Japanese higher education responded to the survey. Because of the online nature of the survey, personal verification of each participant was not possible. However, during the data collection process many participants contacted and identified themselves to the author.

Results

Results of Teaching Experience Survey

All responses to questions related to teaching experience can be viewed in Appendix B. To summarize, the 47 participants stated that they had an average of 13.3 years of experience teaching in Japanese higher education. 40% of participants identified their current position as part time lecturers, 30% of participants identified as full time lecturers, 19% of participants identified as professor / assistant professor / associate professor, 4% of participants identified as researchers, and 6% of participants identified as other positions such as retired lecturer or visiting faculty member. In terms of native language, 89% of participants stated English as their native language or one of their native languages. Participants' other native languages included Japanese, Spanish, Setswana, etc.

Results of Identification Items

The eight identification items were presented to 47 EFL university teachers in Japan in order to investigate the first research question - *Can teachers with EFL experience in Japanese higher education identify which of eight essays were generated by ChatGPT and which were written by CEFR A2 level Japanese university students?* Full responses to the essay identification items can be found in Appendix C. In total the 47 participants produced a global identification score of 54.25% of essays identified correctly. In raw numbers, 204 out of 376 essays presented to all participants were correctly identified as being generated by ChatGPT or Human. In particular, questions 5 and 6 which were both essays generated by ChatGPT were not identified correctly by the majority of teachers. On average, question 5 was identified correctly by 34% of participants and question 6 was identified correctly by 30% of participants. These results indicate that EFL university teachers in Japan don't have the ability to accurately distinguish between ChatGPT and human written essays when ChatGPT is prompted to mimic grammar mistakes of EFL university students.

Differences in Teaching Experience

In order to satisfy Research question 2, data was collected to discern whether teachers with different teaching experience and positions were more or less effective at identifying between ChatGPT and Human essays. Teachers were grouped into three teaching experience levels based on the years of experience served in Japanese institutions of higher education which participants disclosed in the survey. The three experience levels were those with 0-9 years of experience, those with 10-19 years of experience, and those with 20+ years of experience. It was found that teachers with 0-9 years of higher education experience scored highest on the identification items at an average of 60.42%. The difference between teachers with 10-19 years and 20+ years of experience was negligible. Teachers with 10-19 years of experience identified exactly 50% of items correctly. Teachers with 20+ years of experience identified 51.25% of items correctly.

Table 1. Teacher Experience in EFL

EFL experience in higher education	Number of teachers	Average of items identified correctly
0-9 years	18	60.42%
10-19 years	19	50%
20+ years	10	51.25%

Differences in Teaching Position

Teachers were also grouped into five different teaching position categories based on the results of the survey. The five teaching categories were part-time lecturer, full time lecturer, professor / assistant professor / associate professor, researcher, and other. Those who reported being part-time lecturers and full time lecturers scored similarly. Part-time lecturers scored on average 54.61% of identification items correctly. Full time lecturers identified the essays correctly 55.36% of the time. Surprisingly, the group consisting of professors, assistant professors, and associate professors identified correctly between ChatGPT and human essays on just 40.28% of

items. Because of a lack of respondents, not enough statistical data was made available to calculate reliable averages for the researcher and other categories.

Table 2. Current Work Position at Japanese Institutions

Position	Number of teachers	Average of items identified correctly
Part-time lecturer	19	54.61%
Full-time lecturer	14	55.36%
Professor, Asst Prof, etc.	9	40.28%
Researcher	2	Not enough data
Other	3	Not enough data

Comments by High Scoring Teachers

During and after the data collection numerous teachers contacted the author by email, private message, and Facebook comments. Many teachers seemed eager to know their score on the survey and divulged details such as time and date stamps, years of experience, and teaching position in order for the author to help identify them. Thanks to this important feedback the author was able to verify the identity of five of the six highest scorers on the essay identification items. High scorers were defined by the author as teachers who were able to identify at least seven out of eight essays correctly. Just two participants were able to identify all items correctly while four participants identified seven out of eight items correctly. It should be noted that the six high scorers reported av average of just 5.83 years of experience teaching EFL or ESL in institutions of higher education. The author then proceeded to ask the high scorers how they were able to identify the essays.

Teacher A, who achieved a perfect score, stated that they were very good at pattern recognition. They also read parts of the essays out loud. It helped that they had experience from teaching at the Kindergarten level all the way up to university as that experience helped them to recognize typical *keyword patterns* which stood out to them very quickly. Therefore, Teacher A claimed to be able to make accurate judgments within 15 seconds of reading each essay.

Teacher B, who also achieved a perfect score, also used pattern recognition to identify the essays correctly. They claimed to analyze the essays carefully and found that essays generated by ChatGPT which mimicked low level university students tended to use the phrase *I not agree* quite often. Also the use of advanced connectors and synonyms caused Teacher B to correctly suspect that certain essays were ChatGPT generated. The human written essays tended to have more personal experiences and contextualized examples.

Teacher C, who identified seven out of eight items correctly, found that human written essays were likely to use set English phrases that are explicitly taught in the Japanese school system. Students are often taught these set phrases at more formative levels of education. For example prepositions were used in a particular way rather often. Teacher C also found the ChatGPT generated essays to be too contrived. Teacher C was also able to discern between identification items fairly quickly, claiming to spend about ten minutes on the survey in total.

Teacher D, who identified seven out of eight items correctly, said that they read each essay and thought carefully about the structure. They claimed that essays which were too orderly and linear were likely to be written by ChatGPT. They believed Japanese EFL student writing to be more random, tending to *zigzag* between different ideas.

Teacher E, who identified seven out of eight items correctly, said that they knew the prompts had directed ChatGPT to make grammar mistakes. Teacher E analyzed the mistakes made in the essays. They found mistakes atypical of Japanese students such as misusing verbs ending in -ing. Teacher E expressed that in their opinion ChatGPT was trying too hard to add mistakes.

Discussion

As of writing, this seems to be the first paper seeking to quantify how accurately EFL teachers in Japanese higher education can distinguish between ChatGPT generated essays and human written student essays. The results showed that when ChatGPT was prompted to make grammar mistakes similar to an EFL university student, teachers could identify between ChatGPT and human written essays with 54.25% accuracy. This is slightly worse than an AI identification survey conducted in the US which demonstrated that identification between ChatGPT essays and human essays could be performed with 62% accuracy by high school students and with 70% accuracy by high school teachers (Waltzer et al., 2023). This result is further supported by Gao et al. (2023) who found that humans raters were not able to reliably identify the differences between human written and ChatGPT generated essays. However, in contrast to Gao et al. (2023) the AI detectors ZeroGPT and GPT-2 Output Detector were also found to be ineffective in detecting ChatGPT generated essays when ChatGPT was prompted to make grammar mistakes. The results of the current study also finds support from Weber-Wulff et al. (2023) who found discrepancies and inconsistencies in the detection ability of 14 different AI detectors. The AI detectors often failed to detect AI generated content as well as flagged human written text as false positives at varying rates (Weber-Wulff et al., 2023). It should be concluded then that both humans (Gao et al., 2023; Waltzer et al., 2023) and AI detectors (Weber-Wulff et al., 2023) are unable to reliably detect AI generated content.

The 54.25% detection rate found in this study was slightly better than random chance. It shows that theoretically, EFL or ESL students who are determined and knowledgeable on how to use ChatGPT prompts can use it in English or even in their own native languages to attempt acts of academic dishonesty. Cotton et al. (2024) demonstrates clearly that even academic papers worthy of submission to academic journals can be mostly fabricated by ChatGPT with minimal edits. The amazing capability and ease of use demonstrated by ChatGPT can make using it for writing essays extremely tempting for students and academics alike, and already there have been several cases of plagiarism via ChatGPT detected at universities in the United States (Cotton et al., 2024).

Rates of academic dishonesty in US institutions of higher learning were found to be similar in 1963 and 1996 (McCabe & Trevino, 1996). Despite lower rates of academic dishonesty being reported in Japanese universities, there are a considerable number of students engaging in academic dishonesty according to a recent study (Carlson,

2020). The advent of ChatGPT has made writing precise essays easier than ever (Cotton et al., 2024). In fact, it has already been found that between 30-72.5% of students in Hungary, Germany, the US, and Vietnam already engage in usage of AI chatbots for academic purposes (Forman et al., 2023; von Garrel and Mayer, 2023; intelligent.com, 2023; Ngo, 2023). Although many of these students may use AI chatbots ethically, high rates of chatbot usage coinciding with high rates of academic dishonesty in university settings (Carlson, 2020; McCabe & Trevino, 1996) should warrant thorough research and investigation into how prevalent academic dishonesty is with the aid of AI chatbots.

In the author's opinion, it is only a matter of time until EFL and ESL students discover that with the correct prompts they can create short and convincing essays with typical grammar mistakes using their L1s in a matter of seconds. As has been demonstrated in the present study, teachers at this point don't have the detection tools and training necessary to accurately and reliably detect ChatGPT generated essays which contain grammatical errors. This is especially the case with many part time instructors being employed at multiple universities, with the theoretical burden of having to read short essays from hundreds of students in a short amount of time (OECD, 2014; Strong, 2023).

Conclusion and Further Research

This study surveyed 47 teachers with experience teaching nonnative speakers of English in higher education institutions. The survey used the ICNALE written corpus to randomly select four human written CEFR A2 level Japanese students' essays and ChatGPT to generate four AI chatbot generated essays which mimicked common grammar mistakes. The teachers were asked to identify whether they believed each essay was human written or ChatGPT generated. On average the teachers were able to identify 54.25% of items correctly. It must be concluded that teachers were unable to accurately distinguish between human written and ChatGPT generated essays.

Further research should be conducted on how to move forward in this new age of technology with different task types. In particular, additional research is needed to assist teachers in the development of writing tasks which can avoid AI chatbot usage or support in its detection. Alternatively, educators should ponder on how to utilize the vast knowledge and resources AI chatbots possess in productive ways which can aid in rather than hinder the learning of L2s.

Recommendations

In the opinion of the author, the first and most important recommendation is for teachers to move away from the short essay assignment format. Research shows that a large number of students in several different countries already engage in the use of ChatGPT or other AI chatbots for academic purposes (Forman et al., 2023; von Garrel and Mayer, 2023; intelligent.com, 2023; Ngo, 2023). Although the short essay is one of the most common assignments given to EFL and ESL students (Cheng & Wang, 2007), the current study demonstrates that if an EFL or ESL student decides to engage in academic dishonesty with a good understanding of how to write prompts, they could potentially deceive teachers and AI detection tools fairly easily. Additionally, the majority of students

in Japanese EFL classes have engaged in academic dishonesty at least one time including 31.25% of students who self-reported that they have submitted assignments directly copying and pasting from translation applications (Carlson, 2020). Therefore it seems quite likely that a high number of students who are prone to acts of academic dishonesty may start to or have already utilized ChatGPT in ways that may be considered unethical.

Teachers should instead focus on different types of productive writing and speaking tasks that are not easily replicated by AI chatbots (Rudolph et al., 2023). One way to think about writing assignments is as a *process* rather than a *product* (Brown & Lee, 2015). Brown and Lee recommend creating multiple drafts over the course of several class periods, monitoring the draft process, providing individualized feedback, emphasizing revision, and engaging in student-teacher conferencing. Of course, with ever improving technological abilities and the availability of translation apps and ChatGPT on smartphones, even this method is not without flaw. Perhaps the most infallible recommended writing task is the 10 minute free writing exercise on a blank piece of paper (Brown & Lee, 2015). Although somewhat archaic pen and paper style closed book exams are another way that students can be assessed with minimal fear of academic dishonesty (Rudolph et al., 2023). However, questions about relevance and the usefulness of cramming information in the age of ChatGPT lead Rudolph et al. (2023) to recommend assessments such as testing student skills in realistic situations, asking students to write about topics of genuine interest to them, and having students create presentations, videos, performances, and other relevant projects.

In terms of AI chatbot detection, there is still so much research that needs to be done. However, it seems impossible to catch up with the myriad of rapidly developing technological innovations and tools being created. Although many companies have started to offer AI detection services, they are not considered reliable ways to recognize evidence of academic dishonesty and can result in false positives and false negatives (Weber-Wulff, 2023). As shown in the methods portion of this paper, the author was able to fool the AI detectors ZeroGPT and GPT-2 Output Detector rather easily. However, for ChatGPT generated essays that don't contain grammatical errors but are prompted to mimic university students, the GPT-2 Output Detector may be a useful tool (Gao et al., 2023).

Another possible solution offered by Famaye et al. (2024) is to adopt different grading criteria and assessment methods to students who choose to use AI tools to aid with their submitted coursework. Famaye et al. (2024) underlines the importance of fairness in assessing students who choose to use AI tools and students who forgo the use of AI tools, and suggests that teachers aim to adopt policies that cater to the needs of both groups. However, as can be seen in the current study, teachers may have difficulty identifying human written and AI generated coursework if students attempt to present AI generated work as written without the aid of AI tools.

Besides paper and pencil free writing and simply moving away from short essay writing assignments, some recommendations that can be given for detection of AI chatbots is to follow the advice of the high scorers in this survey. This advice can be summarized as using pattern recognition, finding keywords, set phrases, and grammar that are typically taught explicitly in the national education system, looking for personal experiences and examples within the essay, determining whether the structure was organized too well, and finding repetitive grammar mistakes that are not common for EFL students of that country. Other detection methods include looking for

idiosyncratic language and vague language as signs of student written content and the consistent use of transitional words (firstly, additionally, etc.) and the word *overall* as signs of AI generated language (Waltzer et al., 2023).

.

Limitations

A significant portion of respondents to the survey (19.15%) failed to select *ChatGPT* for any of the essays. Six of the nine teachers who did not select *ChatGPT* for any items selected the option *Human* for all eight items. The other three teachers selected *Don't know* or did not select an answer on one item. It should be noted that all nine of these respondents received a score of exactly 50% as they all selected *Human* on the four human written essays. Possible explanations for why these teachers failed to select ChatGPT even a single time include a lack of experience or understanding of the capabilities of AI chatbots, or simply believing that all of the responses were human written. A number of teachers sent messages and comments to the author explaining that they did not anticipate ChatGPT to make so many human-like errors.

However, it should be mentioned that several teachers expressed that the instructions should have stated clearly that the ChatGPT prompts were designed to mimic the mistakes of EFL learners. These teachers said that they would have answered differently had this prompting methodology been stated. All in all, these teachers received scores indicating that they identified 50% of the items correctly, which is just 4.25% below average. Had the prompting methodology been stated clearly from the beginning of the survey, it may have resulted in a slight boost or even a slight decrease in correctly identified items.

Finally, it cannot be overstated how quickly AI chatbot technology seems to be developing. ChatGPT's proficiency is already much greater than any of its precursors (Waltzer et al., 2023). By the time this paper is published, it could already be terribly out of date. This arms race between technological innovation, discovery of potential problems related to the technology's use, and development of coping strategies will always take valuable time and energy to read about and implement for teachers. Time and energy are two things that a lot of teachers lack, and part-time teachers in particular may already feel marginalized by their insecure job situations (Nagatomo, 2014). Therefore, the author believes it is important to be aware of the general issues and to try to think of ways to avoid AI detection concerns altogether by choosing different task types. Particularly, moving away from short writing assignments like the short essay is highly recommended even though this type of assignment is utilized by the vast majority of EFL and ESL teachers (Cheng & Wang, 2007). However, we should also recognize that EFL and ESL teachers are finite human beings and will always face challenges and limitations in the face of the unlimited capacity of technological innovation.

References

Arai, N. (2017) English Program in the Faculty of International Studies, Takushoku University: English Education for 'Empowerment'. *Next Generation Studies, 1*, 21-42

Berry, D. M. (2023). The Limits of Computation: Joseph Weizenbaum and the ELIZA Chatbot. *Weizenbaum Journal of the Digital Society, 3*(3). 1-24. https://doi.org/10.34669/WI.WJDS/3.3.2

Brown, H. D., & Lee, H. (2015). *Teaching by principles*. (4th ed.) Pearson Japan

Carlson, G. D. (2021). Are You Honest? A Study of Self-Reported Academic Misconduct Among Japanese University English Language Learners. *大手前大学 IIE ジャーナル*, *7*, 7-19.

Cheng, L., & Wang, X. (2007). Grading, feedback, and reporting in ESL/EFL classrooms. *Language Assessment Quarterly*, *4*(1), 85-107. https://doi.org/10.1080/15434300701348409

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Famaye, T., Bailey, C. S., Adisa, I., & Irgens, G. A. (2024). What makes ChatGPT dangerous is also what makes it special": High-school student perspectives on the integration or ban of artificial intelligence in educational contexts. *International Journal of Technology in Education (IJTE)*, *7*(2), 174-199. https://doi.org/10.46328/ijte.651

Forman, N., Udvaros, J., & Avornicului, M. S. (2023). ChatGPT: A new study tool shaping the future for high school students. *International Journal of Advanced Natural Sciences and Engineering Researches*, *7*(4), 95-102. https://doi.org/10.59287/ijanser.2023.7.4.562

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ digital medicine*, *6*(1), 75. https://doi.org/10.1038/s41746-023-00819-6

von Garrel, J., & Mayer, J. (2023). Artificial Intelligence in studies—use of ChatGPT and AI-based tools among students in Germany. *Humanities and social sciences communications*, *10*(1), 1-9. https://doi.org/10.1057/s41599-023-02304-7

Hu, C., (2023, February 3) ChatGPT sets Record for fastest-growing user base – analyst note. *Reuters.* Retrieved from https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the World*, *1*, 91-118.

Intelligent.com (2023, January 23) Nearly 1 in 3 Students have used ChatGPT on Written Assignments. Retrieved from https://www.intelligent.com/nearly-1-in-3-college-students-have-used-chatgpt-on-written-assignments/

McCabe, D. L., & Trevino, L. K. (1996). What we know about cheating in college longitudinal trends and recent developments. *Change: The Magazine of Higher Learning*, *28*(1), 28-33.

Mortali, D. (2023). Class Size in Foreign Language Classrooms. *名古屋外国語大学論集*, *12*, 185-192.

Nagatomo, D. H. (2014). In the ivory tower and out of the loop: Racialized and gendered identities of university EFL teachers in Japan. In *Advances and current trends in language teacher identity research* (pp. 102-115). Routledge.

Ngo, T. T. A. (2023). The perception by university students of the use of ChatGPT in education. *International Journal of Emerging Technologies in Learning (Online)*, *18*(17), 4-19. https://doi.org/10.3991/ijet.v18i17.39019

OECD (2014), "How many students are in each classroom?", in *Education at a Glance 2014: Highlights*, OECD

Publishing, Paris. https://doi.org/10.1787/eag_highlights-2014-24-en

OpenAI. (2024). *ChatGPT* (Feb 11 version) [Large Langue Model] Retrieved from: https://chat.openai.com

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1). 342-363. https://doi.org/10.37074/jalt.2023.6.1.9

Strong, G. (2023). Supporting part-time ELT faculty in a Japanese university. *ELT Journal*, *77*(2), 241-244. https://doi.org/10.1093/elt/ccac008

Waltzer, T., Cox, R. L., & Heyman, G. D. (2023). Testing the Ability of Teachers and Students to Differentiate between Essays Generated by ChatGPT and High School Students. *Human Behavior and Emerging Technologies, 2023*. 1-9. https://doi.org/10.1155/2023/1923981

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Sigut, P. & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), 1-39. https://doi.org/10.1007/s40979-023-00146-z

Zemčík, M. T. (2019). A brief history of chatbots. *DEStech Transactions on Computer Science and Engineering*, *10*. 14-18. https://doi.org/10.12783/dtcse/aicae2019/31439

Author Information

Christopher Saarna

https://orcid.org/0009-0006-8791-2580

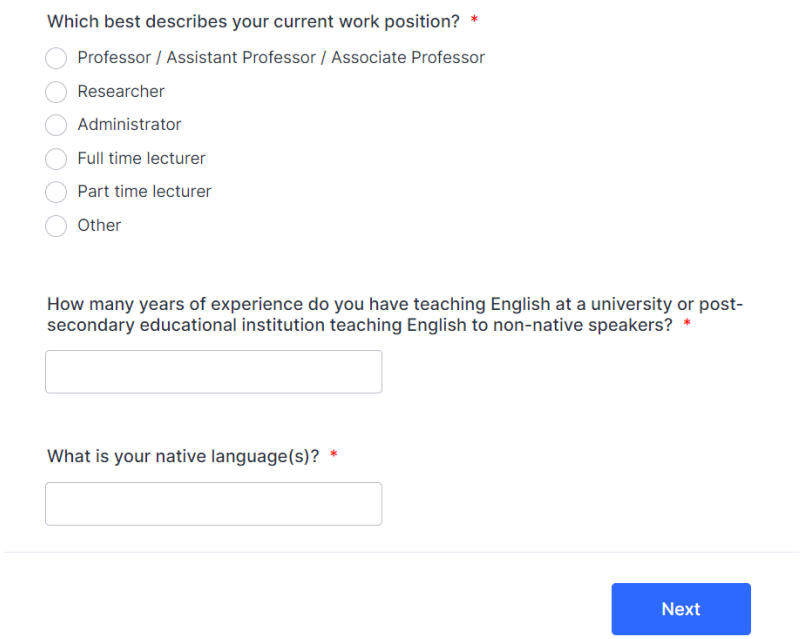
Asia University

5-8 Sakai, Musashino City, Tokyo 180-0022

Japan

Contact e-mail: *saarna_christopher@asia-u.ac.jp*

Appendix A. Survey



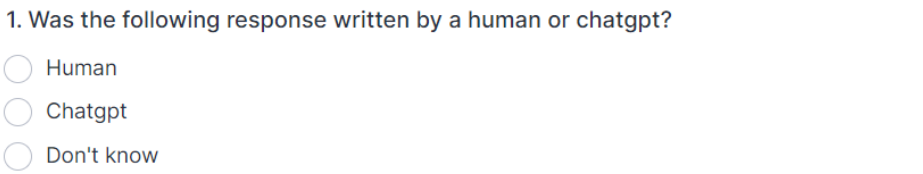
Which best describes your current work position? *

- Professor / Assistant Professor / Associate Professor
- Researcher
- Administrator
- Full time lecturer
- Part time lecturer
- Other

How many years of experience do you have teaching English at a university or post-secondary educational institution teaching English to non-native speakers? *

What is your native language(s)? *

Next



1. Was the following response written by a human or chatgpt?

- Human
- Chatgpt
- Don't know

Having part-time job not important for university students because many reasons. First, college is already very busy time with many classes and homeworks. If student work, they have less time for study and may get low grades. Also, college is for learning and getting degree, not for work. Student should focus on education and not distract with job. Second, part-time job can make student tired and stressed. Many students have to work late hours and then they feel tired in classes. This is not good for learning and can affect student's health. Also, if student work too much, they may not have time for social life and making friends. It is important to have balance in life and not just focus on work. Third, some students have to travel long distance to college. If they have part-time job, they need to spend more time for travel to work and back. This can waste valuable time that could be used for studying or relaxing. Also, it can be difficult to find job that is close to college. In conclusion, part-time job not necessary for college students. They should focus on education and have time for rest and social life. Work can affect grades and health, and students should prioritize their studies during this important time in their life.

2. Was the following response written by a human or chatgpt?

- ○ Human
- ○ Chatgpt
- ○ Don't know

I agree with college students have a part-time job. Because I think that is a way to stand on oneself. First I think people must earn money that to pay for one's hobby, cloth and meal (lunch or dinner) when they become college student or graduated from high school. I think people who graduated from high school is enough adult to be independent of their parents and to earn one's money. Those people will (or must) learn to earn money in public by oneself from part-time job. In that case we don't put a strain for our parents and we can pay money as we like. Second we can work in society. We had less chance to work in public. We don't know how to work in society and to communicate with people who haves same job (but they don't same age). We can learn what about working to earn money in society and we have responsible for our job. And we can learn how to communicate with boss, colleague, business connection and customer. Thus we can learn these things before we become member of society. For the reason of two things, I think part-time job is useful to learn work in public for college students. And we can make time class, study, club and so on, if we have part-time (not full-time) job. That is why I agree with college students having a part-time job.

3. Was the following response written by a human or chatgpt?

- ○ Human
- ○ Chatgpt
- ○ Don't know

I agree with the statement. Of course, college students must study hard. However, I think a part-time job has some advantages. I will show you advantages. First of all, I think it is not good that college students completely depend on parents for money that is used by playing, because almost college students are more than 20 years old. They are adult, so they should prepare for independence. So they should try to play by only using money that they earn by a part-time job. There is advantage that college students can independent. Second, a part-time job makes you notice the importance of money. When I did not have a part-time job, I usually waste money on buying unimportant things. However, I have a part-time job now, I cannot waste money. When I go shopping, I think carefully the goods which I want to buy are really necessary. And I get to save money for future. There is advantage that not only you notice the importance of money but also you can find new use for money. Finally, you can meet many people at workplace. For example the boss, senior, company, customers and so on. When you talk the boss, you must be polite. When you talk to customers, you must answer with polite words. There is advantage that you always be able to be polite unconsciously. On the other hand, when you talk to company, you may find new of life. You may be able to find good friends who you cannot meet at college. There is advantage that you can broaden your horizons. So I agree with the statement that it is important for college students to have a part-time job.

4. Was the following response written by a human or chatgpt?

- ○ Human
- ○ Chatgpt
- ○ Don't know

I think smoking should be completely banned at all restaurants in Japan. There are three reasons why I agree with the ban on smoking. First, my family and I don't smoke, so I smell the little flavor of cigarettes. And I don't like it very much. I think many people also don't like it. Moreover, because of the smell of cigarettes, we can't taste the meal at a restaurant and we can't eat delicious meal. So it is important to ban smoking at all restaurants. Secondly, there are smoking seats and nonsmoking seats at some restaurants. When nonsmoking seats are full and I don't want to wait, I may go to smoking seats. If smoking is completely banned at all the restaurants, I may sit soon at nonsmoking seats. Finally, I think that if the place where we can smoke decrease, people who stop smoking increase. And we can live more comfortable than today. For example, we may not smell the flavor of cigarettes at a station or in the city. In addition to that, illnesses such as lung cancer decrease. Man say cigarettes do harm for health. Smoking has a bad effect not only on smoker but also on non-smoker.

5. Was the following response written by a human or chatgpt?

- ○ Human
- ○ Chatgpt
- ○ Don't know

In my thinking, I not agree that smoking should totally ban at all the restaurants in the country because it removes personal freedom from people. Although I knowing that smoking can be bad for health, it important to thinking about the rights of individuals to make them own choices. Firstly, people going to restaurants for relaxing and enjoy themself. If smoking is ban, it maybe making some customers feeling like them freedom is limited. Everyone having different thinkings about smoking, and it not fair to force everyone follow same rules. People should be able to deciding for themself whether they want smoke or not. Secondly, some restaurants having special places for smoking, and this can be a balance. By having specific areas for smokers, nonsmokers still can enjoying them meals in a smoke-free place. This way, both groups can be together without hurt each other rights. Furthermore, banning smoking in all restaurants maybe have bad effects on the economy. Many restaurants make money from customers who smoke, and a total ban maybe leading to decrease in business. This maybe causing job losses for people working in the restaurant industry. In end, while I understanding the health problems about smoking, a complete ban in all restaurants maybe not be the best solving. It important to finding a balance that respecting the rights of both smokers and nonsmokers while still allowing people make them own choices.

6. Was the following response written by a human or chatgpt?

- Human
- Chatgpt
- Don't know

Smoking not good in restaurants, should stop everywhere. Smoking bad for health and bother people. First, smoking not good for health. If people smoke in restaurants, make others sick too. Secondhand smoke not nice. It hurt lungs and make breathing hard. We don't want people sick because of smoking in restaurants. Also, smoking make eating not fun. The smell bad, stay on food. Non-smokers no like it. If we stop smoking in restaurants, everyone can enjoy food better. Some people say it's their choice to smoke. But smoking in restaurants not just about them. It about others too. We need think about everyone's health. Banning smoking in restaurants good idea. In conclusion, no smoking in restaurants important. It help keep people healthy and make eating better for everyone. Let's stop smoking in all restaurants!

Back Next

7. Was the following response written by a human or chatgpt?

- Human
- Chatgpt
- Don't know

Recently, smoking has been banned in public places. This trend is spreading the entire world. For instance, in Canada, Australia, New Zealand, Norway, Finland, Singapore and the EU smoking is prohibited in public place and fine for smoking is very much. Yet in Japan there are a few places where smoking is banned. And fine is small. So compared with the law of smoking in developed countries that of Japanese is so slack that the government in Japan has to enforce the strict rule. Public places should be comfortable and relaxing for everyone. But smoking is a very bad smell for some people who don't smoke. Restaurant is public place and all of people have to be able to relaxing there. So smoking should be banned at all the restaurants. And the cigarette smoke make dishes awful. Some restaurants have divided the spaces into two areas, smoking area and nonsmoking areas. Yet this system is not enough to completely provide smoke from streaming into nonsmoking area. Besides lately, plenty of hotels have decided to forbid smoking there. After this, most of place, of course streets, smoking are banned. People have better smoke in smoking room not to hurt people's feeling.

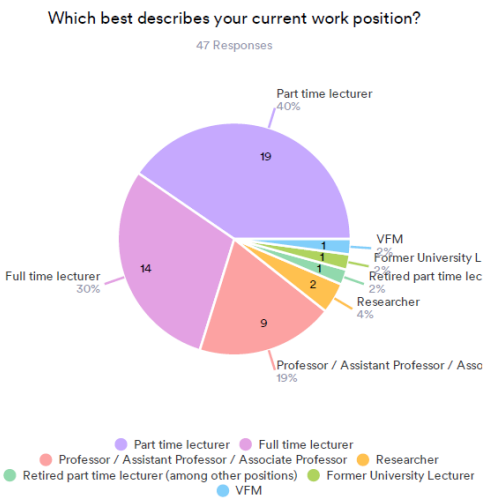
8. Was the following response written by a human or chatgpt?

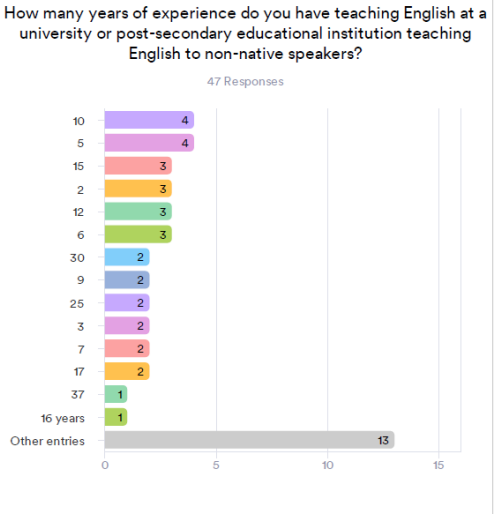
- ○ Human
- ○ Chatgpt
- ○ Don't know

Having part-time job is very important for college students because it helps in many ways. First, it gives us money which we need for many things like books, food, and rent. Without money, it is difficult to survive in college. Second, part-time job teach us about real world and how to work with people. We learn skills like communication and teamwork which are very important in future job. Also, having part-time job can make us more responsible. We need to manage time between classes and work. This teach us to be organized and make schedule. It is not easy, but it is good for future. Moreover, part-time job can help us in our career. When we apply for job after college, employers look for experience. If we have part-time job, we have experience and it can make us stand out. It shows that we can work hard and we are serious about our future. In conclusion, having part-time job is very beneficial for college students. It gives us money, teach us important skills, make us responsible, and help in our career. So, it is important to find balance between work and study. It is not easy, but it is worth it.

Back Submit

Appendix B. Work Experience, Position, and Native Language Results





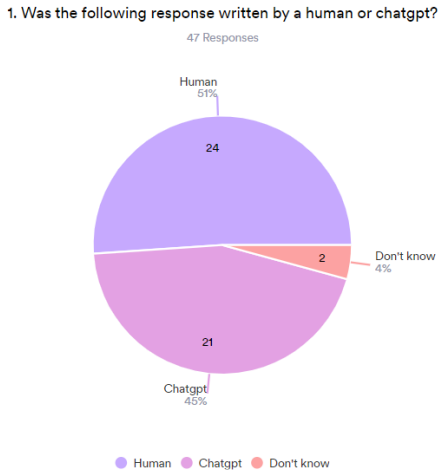
What is your native language(s)?

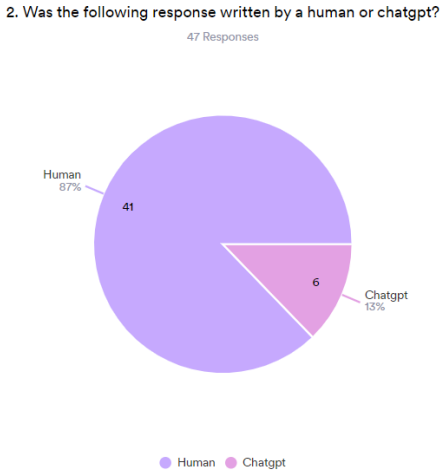
47 Responses

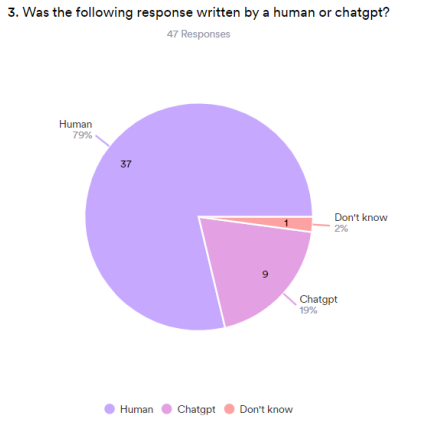
Data	Responses
English	40
Japanese	2
Spanish	2
Setswana	1
English & Hindi	1
Filipino, English	1

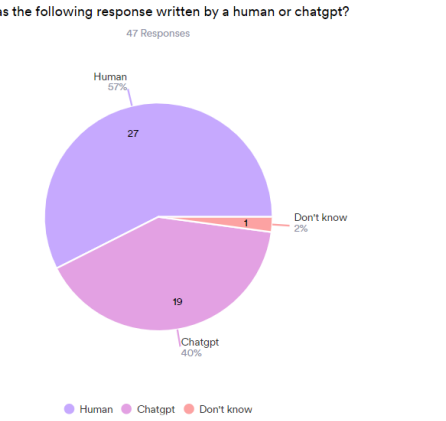
Appendix C. Identification Item Results

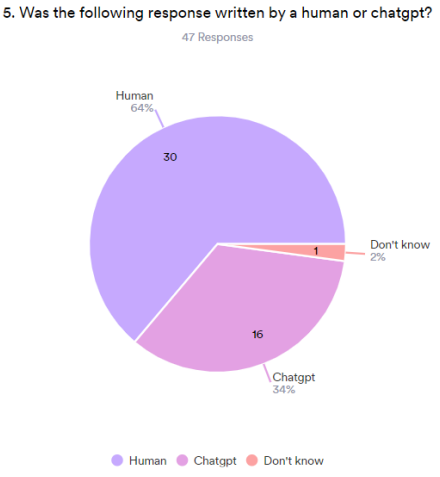
Answer key – 1. ChatGPT 2. Human 3. Human 4. Human 5. ChatGPT 6. ChatGPT 7. Human 8. ChatGPT

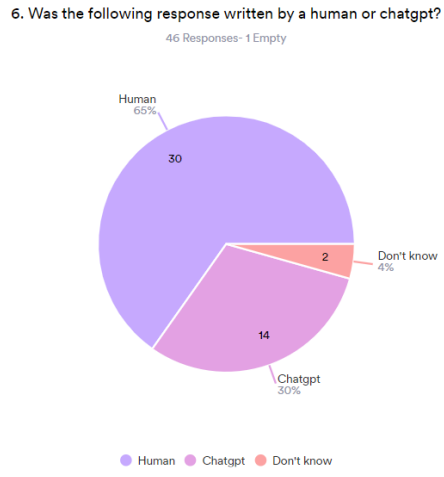


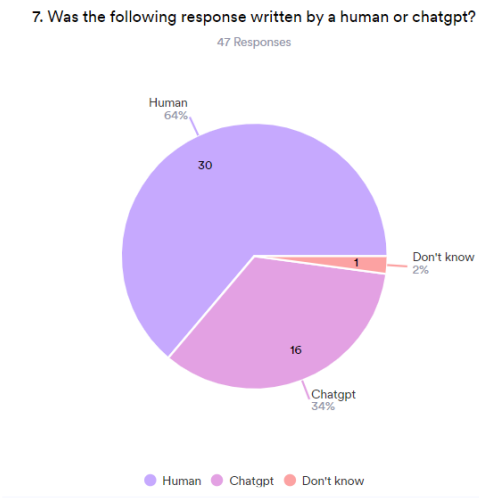


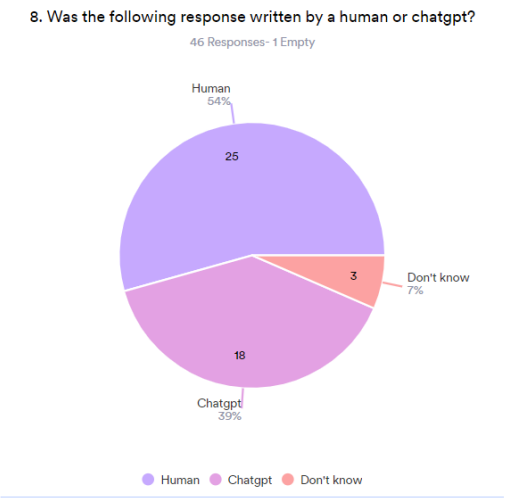









7. Was the following response written by a human or chatgpt?
47 Responses
Human
64%
30
1
Don't know
2%
16
Chatgpt
34%
Human
Chatgpt
Don't know


8. Was the following response written by a human or chatgpt?
46 Responses- 1 Empty
Human
54%
25
3
Don't know
7%
18
Chatgpt
39%
Human
Chatgpt
Don't know