Can Psychology Students and Researchers Distinguish between an AI-generated and a Human-authored Journal Abstract?

[redacted]

Faculty of Health and Medical Sciences, University of Adelaide

September 25, 2023

Word Count: 6,771

This thesis is submitted in partial fulfilment of the Honours degree of Bachelor of Psychology (Advanced) (Honours)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Limited research has explored the application of Large Language Models (LLMs), such as ChatGPT, in writing journal abstracts. To determine the practicability of LLMs in psychology research writing, I investigated the distinguishability of AI-generated versus human-authored abstracts through a mixed between-within participants experiment. Fifty-six participants with varying levels of researcher experience in psychology completed a survey in which they viewed five AI-generated and five human-authored psychology abstracts. Using Signal Detection Theory (SDT) to investigate whether participants were able to distinguish between AI-generated and human-authored abstracts, it was found that students and researchers performed poorly when distinguishing between the two types of abstracts, with a mean performance of 52% correct and a d-prime of 0.11. I then investigated if the level of researcher experience between Undergraduate Students, Ongoing Postgraduate Students, and Completed Postgraduate Researchers, influenced ability to distinguish between AI-generated and human-authored abstracts. A one-way ANOVA and post-hoc tests revealed that Undergraduates and Ongoing Postgraduates were almost equally as proficient and demonstrated significantly better distinguishing ability than Completed Postgraduates. Finally, I investigated whether participants evaluated the textual quality of ChatGPT abstracts as equivalent to human-authored abstracts and discovered through a paired *t*-test that AI-generated abstracts were evaluated as significantly better. The results present important implications regarding the use of LLMs within psychological research, as well as the need for students and researchers to build their AI literacy. Research protocol is available on the Open Science Framework (OSF).

Key words: Artificial Intelligence (AI), Large Language Model (LLM), Abstracts, ChatGPT, Signal Detection Theory (SDT)

DECLARATION

This thesis contains no material which has been accepted for the award of any other degree of diploma in any University, and, to the best of my knowledge, this thesis contains no material previously published except where due reference is made. I give permission for the digital version of this thesis to be made available on the web, via the University of Adelaide's digital thesis repository, the Library Search and through web search engines, unless permission has been granted by the School to restrict access for a period of time

CONTRIBUTOR'S TABLE

ROLE	ROLE DESCRIPTION	STUDENT	SUPERVISOR 1	SUPERVISOR 2
CONCEPTUALIZATION	Ideas; formulation or evolution of overarching research goals and aims.	X	X	X
METHODOLOGY	Development or design of methodology; creation of models.	X	X	X
PROJECT ADMINISTRATION	Management and coordination responsibility for the research activity planning and execution.	X	X	X
SUPERVISION	Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team.	X	X	X
RESOURCES	Provision of study materials, laboratory samples, instrumentation, computing resources, or other analysis tools.	X	X	X
SOFTWARE	Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code.	X		X

INVESTIGATION	Conducting research - specifically performing experiments, or data/evidence collection.	X		X
VALIDATION	Verification of the overall replication/reproducibility of results/experiments.	X	X	X
DATA CURATION	Management activities to annotate (produce metadata), scrub data and maintain research data) for initial use and later re-use.	X		X
FORMAL ANALYSIS	Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data.	X	X	X
VISUALIZATION	Visualization/data presentation of the results.	X	X	X
WRITING – ORIGINAL DRAFT	Specifically writing the initial draft.	X		
WRITING – REVIEW & EDITING	Critical review, commentary or revision of original draft	X	X	X

Can Psychology Students and Researchers Distinguish between an AI-generated and Human-authored Abstract?

Introduction

Large Language Models (LLM) represent a powerful application of Artificial Intelligence (AI) technology. These models are a form of Machine Learning (ML) that have undergone comprehensive '*training'* from exposure to vast amounts of textual data, using Natural Language Processing (NLP) (Trott et al., 2023). LLM training involves processing of human text to identify patterns within the data. LLMs are subsequently able to predict and generate human-like text by analysing and mimicking patterns between words and phrases (Korinek, 2023). NLP describes a method that allows computers to process and interpret natural human text or speech. In terms of functionality, LLMs can analyse the emotional tone of texts (Tang et al., 2023), translate complex text between different languages (Lai, 2023), and power sophisticated 'chatbots' (Wei, 2023). A 'chatbot' is a software application that can simulate human-like conversation with other parties (e.g., human users); its main purpose is to respond to queries, or 'prompts', inputted by users. As LLMs increase in popularity and accessibility, it becomes pivotal to address their application in research writing in order to make effective use of them.

Development of LLMs

Concerns about AI technologies, such as LLMs (e.g., fears that human workers could be replaced by technology; broader concerns about ethics and intellectual property; potential use for the creation and dissemination of false information) gained momentum in the early 2020s (Intahchomphoo & Tschirhart, 2022). It should be noted that LLMs are not new technologies per se*;* there are notable references to general underlying concepts from the 1950s (Macdonald, 1954). Early statistical forms of ML yielded much greater efficiencies than their predecessors, but their reliance on resource-intensive 'feature engineering' (i.e., the process of extracting features [characteristics, properties, attributes] from raw data) limited their utility and broader application (Maucec & Donaj, 2019). 'Deep Learning' approaches to ML, which gained

prominence throughout the 2010s, led to substantial advancements as they facilitate the computation of 'neural networks', which use generalizable 'word embedding' approaches to capture (and interpret) the semantic properties of words much more efficiently than the specific feature engineering required by earlier statistical approaches (Asudani et al., 2023).

The refinement of deep learning and neural NLP approaches paved the way for the development of modern LLMs. There have been several well-known examples of LLMs since the technology's emergence (e.g., Meta's LLaMa (Touvron, 2023) and Google's PaLM (Chowdhery, 2022), however, none have quite reached the popularity of Open AI's Generative Pre-trained Transformer (GPT) models (e.g., GPT-3.5 and GPT-4), which power ChatGPT (Liu et al., 2023). Unlike many LLM interfaces, which require steep learning curves and programs inaccessible to the average person, ChatGPT is a publicly available chatbot, offering a relatively simple user interface (Zhai, 2022). Released by *OpenAI* in November 2022, ChatGPT took only five days to reach one million users and as of July 2023, reports over one billion visits per month (Shewale, 2023). ChatGPT can generate fluent and contextually appropriate responses to a range of queries (prompts) that users may present (Hassani & Silva, 2023); its application in research writing has not been richly explored.

Implications of LLM Application in Research

ChatGPT's text generation abilities, coupled with its high accessibility, present several practical implications for academic research that require further investigation (Sullivan & Kelly, 2023). Users can provide prompts to the chatbot, such as a question, or statement, and ChatGPT is able to generate high-quality text tailored to users' requests. LLMs of this nature have thus far presented a multitude of benefits and concerns within academia. LLMs such as ChatGPT showcase ability in making research more efficient (Dones, 2022). For instance, LLMs can significantly reduce effort and time spent in synthesising systematic reviews, as they use modelling to analyse vast amounts of data and categorise research in relation to the criteria of the systematic review (Wang et al., 2022). Scientists can therefore utilise LLMs to generate

textual data. However, despite LLMs' ability to make the research process more efficient, human involvement is still crucial; humans are required to create effective prompts for LLMs to interact with (Schwartz et al., 2023). The technology of ChatGPT also enables non-native English speakers to produce more fluent and coherent research compositions in English, thus improving the overall quality of research reports published in non-English speaking countries (Kitamura, 2023). Despite the benefits, the rising popularity of ChatGPT poses a range of threats to academia. While AI reduces human effort, it harbours an environment for plagiarism, manipulation, and deception (Jakesch et al., 2023). Users can copy information generated by ChatGPT and make use of it without appropriate acknowledgement or citation (Hill-Yardin, 2023). More concerningly, ChatGPT may be prone to bias and inaccuracy. ChatGPT responds to prompts based on the dataset it is trained on; should the dataset contain biases, the chatbot may reinforce or magnify existing biases within the dataset and may present information that is factually inaccurate though sounding entirely believable; a phenomenon termed 'hallucination' (Deng & Lin, 2022). As acknowledged by its developers, ChatGPT's responses do tend to be well-written, but can be outdated and use non-existent evidence to support claims (Fitria, 2023). Due to the rising popularity of ChatGPT, the risk of research falsification becomes amplified, and so, its involvement within research writing becomes more critical to explore empirically.

Comparison of AI-generated and Human-authored Research

Few studies have compared AI-generated research writing to human-authored research writing. Haq et al. (2023) explores content analysis conducted by a panel of scientists to evaluate the completeness, credibility, and scientific content of research articles written by ChatGPT and human authors. The study found that ChatGPT could effectively compose articles that seemed believable, but deeper evaluation showed weakness in reliability and accuracy, as well as structure. Dowling and Lucey (2023) examined ChatGPT's text generation performance throughout the individual stages of the research process and found that the platform performed best in the initial stages of generating a meaningful research idea but performed poorly on the

literature review, stating that the model lacks the capability of adequately connecting multiple internally generated research ideas. Blinded human reviewers in Gao et al. (2023) struggled to discern between AI-generated and human-authored abstracts as the reviewers were only able to correctly identify AI-generated abstracts 68% of the time. The study also showed that ChatGPT was able to generate fabricated yet believable statistics, making it difficult to distinguish from human-authored abstracts.

Using Linguistic Cues to Distinguish between AI-generated and Human-authored Text

It is important to establish whether AI text generated by modern LLMs is distinguishable from human-authored text if it is to be implemented in research writing. Individuals can utilise *linguistic cues* to distinguish between AI and human-authored text. Defined by Bates and MacWhinney (2014), linguistic cues refer to linguistic information of text such as syntax, grammar, semantics, and writing style; these linguistic indicators can be used to reveal information about the authorship of a given text. Clark et al. (2021) reported that less experienced evaluators of AI-generated text focussed on the format of text such as the grammar and style to guide their evaluation. Experienced evaluators focussed more significantly on the content of text, and unlike the lesser experienced evaluators, did not underestimate the quality of AI-generated text. Regardless, both groups of evaluators struggled to distinguish between AI-generated and human-authored text. Casal and Kessler (2023) sampled a panel of linguists evaluating AI-generated and human-authored content; the linguists attributed vagueness and a lack of detail to AI-generated content, and richness of information to human-authored content. However, the panel was largely unsuccessful in distinguishing between the two types of content when content was anonymised. On the other hand, Uchendu et al. (2021) and Dowling and Lucey (2023) reported that LLMs such as ChatGPT struggle in constructing a natural flow in their constructed arguments whereas human authors showcase ability to organically transition between their points. Markovitz et al. (2023) reported that human reviewers attribute emotion to human-generated content while inauthenticity is attributed to AI-generated textual expression. It

is important to note that as LLMs become more sophisticated and further improve in self-learning (Gordijn & Have, 2023), distinguishing between AI and human text becomes more challenging. Despite LLMs being in development for several years, they have become more powerful now due to the rise of more quality datasets, improvements in processing power, as well as the incorporation of human feedback into the creation of these models (Dis et al., 2023). Inability to differentiate between AI-generated and human-authored text could have both positive and negative consequences. For one, it would suggest that AI text has addressed its textual flaws and evolved effectively to organically integrate within human-authored research. On the other hand, if the author of the text is unknown, the reliability and scientific integrity of the data may become compromised. These concerns highlight the need for further research, particularly regarding the implementation LLMs in research writing.

LLMs in Abstract Writing

Journal abstracts are a useful element of research; presented at the front of the paper, they are designed to concisely convey the main points of a study. An effective abstract is comprehensive and gives the reader a complete picture of the study. However, due to a lack of specialised training (Sanganyado, 2019), journals are filled with studies containing poorly structured abstracts; Lazarus et al. (2015) analysed 128 human-authored abstracts published within BioMed Central Medical journals between 2011 to 2013 and found that 107 (84%) contained some form of misleading information. LLMs have the potential to produce articulate and informative abstracts, however, prior studies do not delve deeply into the textual efficacy of AI-generated abstracts in comparison to human-authored abstracts, hence the implications remain unclear.

Lazarus et al. (2015) and Gao et al. (2023) both evaluate the efficacy of biology abstracts and the implications of LLMs within biology abstract-writing. However, no studies within the literature delve into psychology abstracts. Unlike biological research, which often cites concrete indisputable findings, psychological research tends to delve into topics of human

emotion, attitudes, and subjective experiences to fully comprehend human behaviour, making it a field that is inherently different (Chamberlain & Broderick, 2007). LLMs such as ChatGPT may struggle to mimic the specific objective methodologies mentioned in biology research, making it easier to distinguish from human-authored abstracts. As ChatGPT continues to develop in mimicking human language, it improves its ability to generate commentary on human topics; hence improving its potential to compose psychology abstracts that seem human in nature. This phenomenon may make AI-generated psychology abstracts harder to distinguish when compared to other fields of science such as biology.

The Present Study

In this thesis, I aim to determine the practicability of LLMs, such as ChatGPT, for the purposes of producing abstracts for psychological studies.

To investigate the application of LLMs in research, it is beneficial to explore whether individuals can distinguish between AI-generated and human-authored abstracts. The first research question therefore asks, 1) Can psychology students and researchers distinguish between AI-generated and human-authored journal abstracts?

Furthermore, the literature highlights that experienced individuals such as expert linguists utilise specific linguistic cues to distinguish between AI-generated and human-authored abstracts yet tend to be generally unsuccessful. However, researcher experience specifically has not been explored to this degree. As LLMs such as ChatGPT have undergone extensive development in recent years to effectively mimic human language, and application to psychological abstracts has not been explored, it is unclear how experienced and inexperienced researchers will differ in distinguishing between AI-generated and human-authored psychology abstracts. Hence the second research question asks, 2) Does the level of researcher experience influence ability to distinguish between AI-generated and human-authored journal abstracts?

Furthermore, the literature describes AI-generated texts as vague and inaccurate yet also as coherent and believable. As I aim to explore the practicability of LLMs within research writing, it becomes important to evaluate whether AI-generated text produced by ChatGPT is noticeably different in linguistic standard than human abstracts. However, there is a lack of text evaluation measures regarding LLM text; Lee et al. (2021) acknowledges this and recommends best practice in evaluating LLM-generated text. I created a new scale using the recommended guidelines, to explore how participants evaluate text features of given abstracts. Hence, the final research question asks: 3) Do psychology students and researchers evaluate textual features of AI abstracts differently to human abstracts?

Method

Participants

I recruited participants from Schools of Psychology by contacting heads of schools for each Australian Psychology Accreditation Council (APAC) accredited program in Australia, and asking them to distribute my study details and link to current students, staff, alumni, and titleholders. Undergraduate students were additionally recruited via the University of Adelaide first-year participant pool platform, SONA, and given course credit for completing the study. All recruited participants were studying psychology or had previously studied psychology. Participants were recruited in accordance with three levels of researcher experience, 1) 'Undergraduate Psychology' students, 2) 'Ongoing Postgraduate Psychology' students, and 3) 'Completed Postgraduate Psychology' researchers. The sample (N = 56) comprised 42 female, 13 male, and one non-binary identification. The mean age was 27.6 years (SD = 11.2) and the range was 18-55 years. Among the 56 participants, there were 27 'Undergraduates', 15 'Ongoing Postgraduates', and 14 'Completed Postgraduates'. All participants were recruited within a three-week timeframe. Upon completion, participants were invited to enter a prize draw for a $50 gift card (students receiving course credit were excluded from this draw).

Stimuli

To create the abstract stimuli, I reviewed abstracts from a collection of high impact psychology journals as classified by *SCImago Journal Rank (SJR)*. Upon my review, I collected 10 peer-reviewed empirical studies to be used in the final study apparatus with the aim of producing a sample that covered a range of fields in psychology such as Cognitive, Clinical, and Neuropsychology (See Appendix A for list of studies). The abstracts were then extracted from each study and compiled separately.

To create AI-generated versions of the abstracts, I developed a specific prompt for each abstract, using a standardised structure populated with study-specific details, which were then inputted to ChatGPT. I took inspiration from Gao (2023) in developing the prompt structure,

particularly, mentioning the title and journal in the prompt. I refined their approach by mentioning specific details of the study such as the Method and Results to ensure ChatGPT would not fabricate findings (See example prompt in Appendix B); this approach of using specific details of the study was in line with my experimental purpose of determining the real-life practicability of LLMs in research writing.

ChatGPT is designed to retain information and remembers previous prompts within the same chat session to inform its responses (Buholayka, 2023); because of this trait, I entered the prompt for each abstract in separate sessions so that ChatGPT would not learn from past prompts and refine its responses based on that information. Each prompt was delivered in isolated sessions to ensure fair representation of ChatGPT's language processing ability.

Measures

Forced Choice Confidence Scale

Participants rated whether an abstract was AI-generated or human-authored on a 12-point confidence scale. The confidence scale was split into two 6-point scales for each of AI or human abstracts, with '1' being not at all confident, and '6' being totally confident. The scale was presented on one line, with the left side of the scale (6 to 1), representing confidence in rating the presented text as AI-generated. The right side of the scale (1 to 6), represented confidence in rating the presented text as human-authored. By excluding a 'zero' option, participants were forced to choose between an abstract being AI-generated or human-authored, while the use of a confidence scale revealed participants' conviction or uncertainty in their choices (See Appendix C for scale layout).

Text Evaluation Score (TES)

Given the lack of consensus, I generated a novel measure drawing on previous literature around text evaluation and scale design. Lee et al. (2021) outlines recommended practice regarding human evaluation of LLM text and suggests that Likert scales tend to be the most effective in experimental literature. Participants were presented with five variables of text

features that Lee et al. (2021) deemed significant to evaluate: Fluency, Informativeness, Quality, Readability, and Clarity. Text evaluation score items were designed as 5-point Likert scales (See Appendix D for example item). The individual scores of each of the five variables were scored out of five and summed to attain an overall score of 25. I named this measure the Text Evaluation Score (TES); this score is attributed to each abstract and exhibits how participants evaluate the features of each abstract.

I conducted a McDonald's Omega to test the internal reliability and factor structure of the newly created TES; the RMSEA was 0.142 indicating poor fit, The CFI was 0.977 indicating very good fit. The TLI was 0.944 indicating good fit. The fits, though not perfect, were considered sufficient to treat the TES as a single factor. The TES measure consists of five variables, with individual reliability coefficients, g, of 0.74, 0.80, 0.81, 0.72 and 0.80, indicating high reliability (Charter, 2003).

Procedure

Once I had used ChatGPT to generate an abstract for each of the human-authored abstracts, there were a total of 10 human-authored abstracts, and 10 corresponding AI-generated versions of those abstracts. A collection of 10 abstracts were composed in four separate blocks and participants were randomly assigned to completing one of these four blocks. Abstracts were randomly allocated to each block, and it was ensured that each block had 10 abstracts with five AI-generated and five human-authored.

Ethics approval was granted by the University of Adelaide's School of Psychology Human Research Ethics Subcommittee (approval #23/60). The study was administered on Qualtrics as a mixed between-within participants survey. Upon providing consent, participants began the survey by providing demographic information (See Appendix E). Participants were then randomly presented ten abstracts one at a time. Underneath each abstract, participants rated, on a forced-choice confidence scale, whether they thought the presented abstract was AI-generated or human-authored. After this, they responded to the TES measures to evaluate the

textual features of the abstract. A multiple-choice comprehension question was presented as a follow-up to each abstract to measure whether participants retained salient information from each abstract. Once participants had completed their evaluation of the ten abstracts, they were asked to respond to the following questions: 1) “What factors did you use to decide whether an abstract was human or AI-generated?”; 2) “Do you think AI has a place in research writing in the future?”

The function of these open-ended questions was to further inform the research questions and determine the directions for future research. At the end of the survey, participants were invited to register for a $50 gift card prize draw. To ensure transparency about how this study was conducted, I included a preregistration in the Open Science Framework (OSF), outlining key details of the research process including the analyses (available at: https://osf.io/hkmpe/?view_only=22f505d10ad044908207e83f5b805b7a).

Analysis

Data analysis was completed in R *Version 2022.12.0+353* and data was cleaned to remove incomplete responses. The Signal Detection Theory (SDT) framework (Pastore & Scheirer, 1974) was used to quantify how participants discriminated between AI-generated and human-authored abstracts. I used SDT to evaluate participants’ ability to distinguish *signal* from *noise*. In this case, the *signal* is correctly distinguishing an abstract as AI-generated, and *noise* is incorrectly distinguishing an abstract as AI-generated. I used SDT to gain values for three key insights: accuracy, discriminability, and response bias. I quantified participants’ overall accuracy in correctly distinguishing between human and AI abstracts by evaluating the proportion of correct answers across the 10 abstracts.

I measured discriminability using d-prime, which explained how effectively participants were able to discriminate *signal* from *noise* by evaluating the standardised difference between the means of hits (evaluating an abstract as AI-generated when it is AI-generated) and false alarm (evaluating an abstract as AI when it is human-authored) distributions. A higher d-prime

value of around 2 indicates better discriminability; a value of around zero indicates chance-performance (Crossman & Lewis, 2006). As measured by C, response bias values explained the tendency of participants to favour one response over another. Response bias evaluates the proportion of hits and false alarms and determines whether participants are more inclined to one type of response. For instance, participants may be more inclined to evaluate given abstracts as 'AI' in comparison to evaluating 'Human'. A response bias value of around zero indicates little to no bias in responses (Cull, 2005). I also incorporated an ROC (Receiver Operating Characteristic) curve within my analysis which graphically represented AI abstracts hits against false alarms.

I aimed to determine significant differences in discriminability and response bias amongst the three groups of varying researcher experience (Undergraduates, Ongoing Postgraduates, Completed Postgraduates) by conducting a one-way ANOVA; d-prime and response bias values were calculated for each participant. I then conducted Tukey post-hoc tests to determine which groups differed significantly from one another. I conducted assumption tests such as the Shapiro Test, Bartlett Test, and histogram visualisations to conclude that conducting a parametric ANOVA and Tukey post-hoc tests would render effective results. A paired *t*-test was conducted to determine mean differences on the TES between the two abstract conditions (AI and Human). Assumption testing using QQ plot, Shapiro-Wilk Test, and histogram visualisations, indicated I could substantiate proceeding with a parametric *t*-test.

Outliers were found and kept for each conducted analysis as they provided valuable information on demographics; for instance, some individuals were exceptional in distinguishing between abstracts. Presence of outliers across all levels of the variable also shows that results were not skewed towards one level of variable.

Results

Signal Detection Theory (SDT) was used to examine participants' ability to distinguish between AI-generated and human-authored journal abstracts. Preliminary calculations showed that participants were able to distinguish correctly 346 out of 663 responses, leading to an overall 52% success rate in distinguishing between AI-generated and human-authored abstracts. However, these calculations present no clear distinction in accuracy across the two types of abstracts. To better understand performance, I conducted an SDT analysis (See Table 1 for confusion matrix).

Table 1

Confusion Matrix of AI vs Human Abstracts

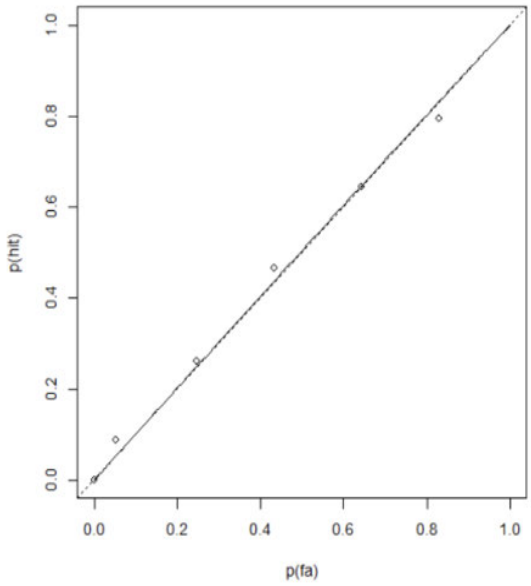
		Abstract category	
		AI	Human
Participant response	AI	157 (hits)	143 (false alarms)
	Human	174 (misses)	189 (correct rejection)

SDT analysis showed that the overall mean d-prime value was 0.11 (*SD* = 0.87*, min value* = -2.06*, max value* = 2.06), which is a relatively low value indicating poor ability to discriminate between signal and noise. The overall mean response bias value was 0.12 (*SD* = 0.40*, min value* = 0.39*, max value* = 2.55). As this is a positive value, it indicates participant bias

towards evaluating ‘AI’ to given abstracts. However, this value is relatively small, suggesting a very subtle bias. I also implemented an ROC curve (Figure 1) to graphically depict AI abstract hits against false alarms which shows participants’ ability to detect AI abstracts. A straight line indicates low discriminability and shows chance-level performance for distinguishing between AI-generated and human-authored abstracts, which is also reflected in the relatively low d-prime value.

Figure 1

First Principal Component ROC Curve of AI abstract hits vs false alarms



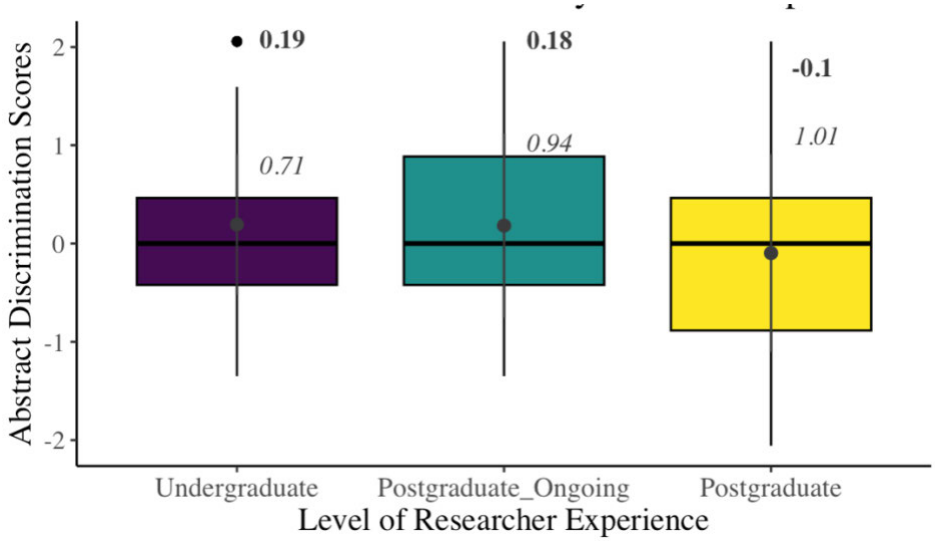
Note. A receiver-operator characteristic curve of detection of AI abstracts (hits) against false alarms for all participants including levels of confidence (plotted points along the line).

To determine whether level of researcher experience influenced being able to distinguish between AI-generated and human-authored abstracts, I conducted a one-way ANOVA for d-prime (dependant variable), to determine differences in mean discrimination scores between the levels of researcher experience (independent variable). Figure 2 demonstrates the scores on

discrimination across different levels of researcher experience.

Figure 2

Abstract Discrimination (d-prime) Scores by Level of Researcher Experience



After determining that the requirements of the assumption tests were met (See Appendix F), a parametric one-way ANOVA was conducted to examine the differences in d-prime values among the three levels of researcher experience. The one-way ANOVA revealed a significant main effect of level of researcher experience, $F(2, 557) = 5.88$, $p < .005$. The effect size, calculated as eta-squared (η^2), was 0.021, indicating a small effect size.

As the ANOVA revealed a significant effect, I proceeded to conduct a Tukey post-hoc test to determine which of the groups among the levels of researcher experience had a significant mean difference in d-prime values. Results, including the *p*-value and the effect sizes of the associated group relationships, are shown in Table 2.

Table 2

Post-hoc Tukey Test Results for d-prime

Comparison between Levels of Researcher Experience	*p*-value	Cohen's *d* Effect Size
Undergraduates - Ongoing Postgraduates	.99	0.01
Ongoing Postgraduates - Completed Postgraduates	< .05	0.33
Undergraduates - Completed Postgraduates	< .005	0.29

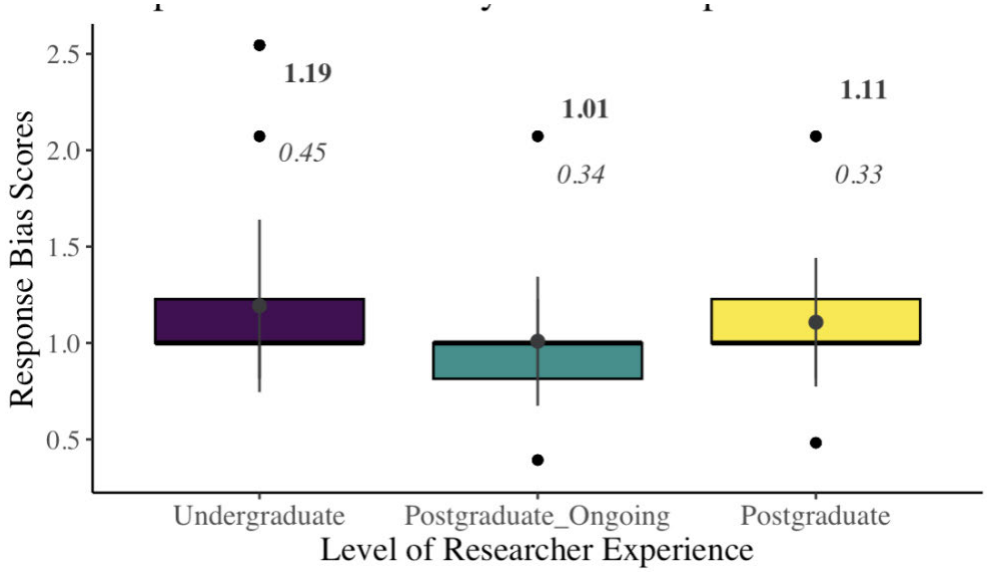
Notes. *p*-value below .05 indicates that there is significant mean difference between given groups

The Tukey post-hoc test results reveal that there is a significant mean difference in d-prime values between Ongoing Postgraduates (M = 0.18, SD = 0.94) and Completed Postgraduates (M = -0.10, SD = 1.01) as well as between Undergraduates (M = 0.19, SD = 0.71) and Completed Postgraduates; Undergraduates and Ongoing Postgraduates have a higher discrimination score than Completed Postgraduates. The mean difference in discrimination scores between Undergraduates and Ongoing Postgraduates was non-significant.

After the d-prime analysis, I aimed to determine if there was any response bias amongst the levels of researcher experience. Hence, I conducted a one-way ANOVA for the response bias values (dependent variable), amongst levels of researcher experience (independent variable). Figure 3 demonstrates the scores on response bias across the levels of researcher experience.

Figure 3

Response Bias (C) Scores by Level of Researcher Experience



After determining that the requirements of the assumption tests were met (See Appendix G), I conducted the parametric one-way ANOVA to examine response bias value differences amongst the groups. The one-way ANOVA revealed a main effect of level of researcher experience, $F(2, 557) = 10.63$, $p < .05$. The effect size, calculated as eta-squared (η^2), was 0.037, indicating a small to medium effect size.

Due to the significant result, I conducted a post-hoc Tukey test to determine if there were any significant response bias mean differences amongst the three groups. Results, including *p*-values and effect sizes of the associated group relationships, are shown in Table 3.

Table 3

Post-hoc Tukey Test Output for Response Bias

Comparison between Levels of Researcher Experience	*p*-value	Cohen's *d* Effect Size
Undergraduates - Ongoing Postgraduates	< 0.01	0.46
Ongoing Postgraduates - Completed Postgraduates	.08	0.29
Undergraduates - Completed Postgraduates	.10	0.21

Notes. p-value below .05 indicates that there is significant mean difference between given groups.

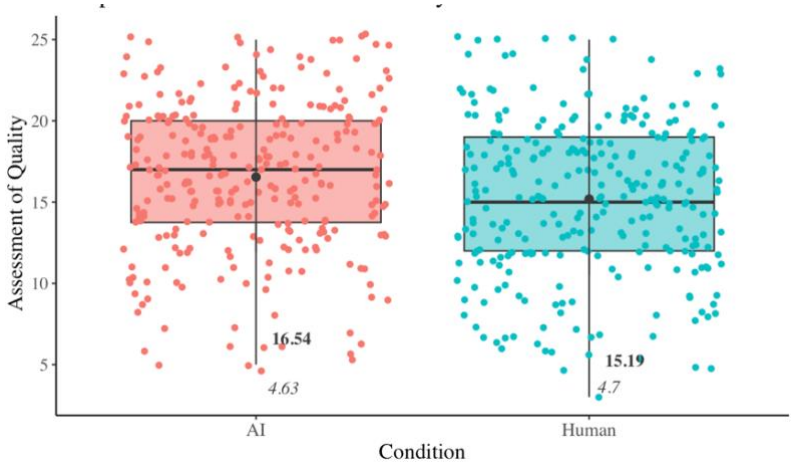
The Tukey post-hoc test results reveal that there is a significant mean difference in response bias values between Undergraduates (M = 1.19, SD = 0.45) and Ongoing Postgraduates (M = 1.01, SD = 0.34) with Undergraduates skewing towards 'AI' evaluation of abstracts. The mean difference in response bias scores between Ongoing Postgraduates and Completed Postgraduates (M = 1.11, SD = 0.33), as well as between Undergraduates and Completed Postgraduates were non-significant.

Finally, I aimed to discover whether psychology students and researchers evaluate text features of AI abstracts differently to human abstracts. After satisfactory output from the t-test assumption tests (See Appendix H), I conducted a paired-samples *t*-test to determine if there was a significant mean difference in the TES between AI (M = 16.5, SD = 4.63) and Human (M = 15.2, SD = 4.70) abstracts. The *t*-test revealed a significant mean difference between the two conditions, $t(279) = 3.65$, $p < .01$, with higher TES scores for AI abstracts. I obtained a small to

medium Cohen's *d* effect size of 0.29. Figure 4 visualises the difference in TES scores between AI-generated and human-authored Abstracts. Table 4 shows the mean TES scores of AI and Human abstracts given to each individual variable of the TES.

Figure 4

Text Evaluation Score (TES) by Abstract Condition



Note. TES of AI and human-authored abstracts (scored out of 25); AI abstracts scoring significantly higher than human abstracts.

Table 4

Mean Scores (alongside standard deviation scores) for each Individual Variable of the TES

Abstract	Fluency		Informative		Quality		Readability		Clarity	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
AI	3.41	1.06	3.25	1.07	3.28	1.02	3.34	1.09	3.27	1.11
Human	3.12	1.03	3.03	1.10	3.02	1.11	3.01	1.10	3.00	1.14

Note: Scores were out of five.

Discussion

I ran a mixed between-within participants experiment to discover if psychology students and researchers could distinguish between AI-generated and human-authored abstracts, and whether the participants evaluated a difference in textual quality between them.

Psychology students and researchers in this study were generally not very effective at distinguishing between AI-generated and human-authored psychology abstracts. Participants could only correctly distinguish between the abstracts 52% of the time. In comparison, Gao et al. (2023) reviewed Biology abstracts and showed an overall accuracy of 68%; the difference in results may be due to more detailed prompts being used in this study, leading to ChatGPT producing abstracts more difficult to distinguish. As shown in Figure 1, the ROC 'curve' is essentially a straight line; this indicates that participants responses to distinguishing between abstracts were equivalent to random guessing. The SDT analysis showed a d-prime value of 0.12, indicating very low discriminability between AI-generated and human-authored abstracts. Response bias was evaluated to be a value of 0.12; a low value like this indicates that there was no significant response bias in abstract distinguishability. Participants struggled to distinguish correctly between abstracts; SDT research explains that difficulty in distinguishability between *signal* and *noise* can render less response bias as participants make more neutral responses (Lerman, 2010).

I further utilised the SDT analysis approach to determine whether researcher experience played a role in ability to distinguish between AI-generated and human-authored abstracts. The lesser experienced Undergraduate and Ongoing Postgraduate participants scored equally well, and significantly better, than the more experienced Postgraduate researchers. In general, it appears that typically younger students and researchers were better able to distinguish between AI-generated and human-authored text than Postgraduate researchers who are generally older. Hermann (2023) outlines that older individuals generally perceive AI text as more natural and therefore more difficult to distinguish from human text, while younger individuals are more

exposed to AI systems and experience more digital technology. Undergraduate students tend to receive more exposure to AI systems education in contemporary educational settings through their university curriculum (Long & Magerko, 2020). However, Completed Postgraduate researchers comprised the smallest comparison group (n = 14) in the current study, limiting the generalisability of the results. I suggest these results be interpreted with caution; despite the analysis showing significant findings, the real-world impact of these results may be limited.

In my final analysis, considering whether the participants evaluate the text features of AI-generated abstracts differently to human-authored abstracts, I discovered that the participants proffered significantly higher text evaluations for AI-generated abstracts. AI-generated abstracts were evaluated as textually better than human-authored abstracts on each of the five variables of the TES (Table 5). This may in part be attributed to the efforts that LLM developers have made in honing the human language production abilities of applications such as ChatGPT (Dis et al., 2023). This is also the first study in the literature that attempted to minimise ChatGPT's weaknesses to test its real-world efficacy; that is, I gave ChatGPT extensive and rich prompts, so that it produces a thorough and sharp abstract, free of fabrication and vagueness. This likely played a factor in ensuring ChatGPT generated a high-quality composition and hence, received higher quality evaluations scores by participants. However, confirmation bias (Klayman, 1995) may also have contributed to these results. Confirmation bias refers to the tendency for participants to favour responses that confirm their beliefs. Within the survey, participants are first asked to review and classify the given abstract as 'AI' or 'Human', and only then proceed to evaluate the text features of the abstract. If their initial classification is incorrect, participants may rate the text features expecting they are rating an AI-generated abstract when in reality it is a human-authored abstract, and vice versa. Hence, it is possible that text evaluation scores could have been impacted by confirmation bias. Future research can counteract this by utilising counterbalancing (Zeelenberg & Pecher, 2015); in my experiment, I solely presented the

abstract classification measure to either ‘AI’ or ‘Human’ before the text evaluation measures, yet future studies can present their range of measures in a more randomised order.

Theoretical and Practical Implications

The results imply that providing detailed and accurate prompts to ChatGPT when tasked to generate psychology abstracts enables the chatbot to produce abstracts that are difficult to distinguish from human-authored abstracts. As participants could not identify varying features between AI-generated and human-authored texts, we could determine that integrating AI text within psychology research writing could go undetected; AI-generated text could be seen as equivalent to human-authored text. With this advancement, researchers could delegate textual tasks to LLMs such as ChatGPT, and themselves focus on more critical aspects of research. Additionally, researchers could improve the quality of their worded compositions; the results indicate that AI-generated text typically had better text evaluation scores than human-authored abstracts. If researchers incorporate AI-generated content within their research, they could see an improvement in their writing quality. It could allow researchers in non-English speaking countries to compose written research to a high linguistic calibre in English and therefore, publishable in international research journals.

On the other hand, the notion that AI-generated text is indistinguishable from human-authored text poses significant threats to the conventions of research writing. Determining the true author of research text becomes a more challenging task; if AI-generated text can mimic human-authored text in research writing, authorship of text becomes difficult to attribute, plagiarism and academic breaches may also go undetected (Perkins, 2023). While detailed prompts may enable users to ensure produced data is accurate and precise, integration of AI-generated text within research may invite discussions on how writing quality and authenticity is assessed within academia (Rudolph et al., 2023). The usage of AI prompts raises a new concern altogether; quality of generated text may vary depending on provided prompts. If AI-text integration within research becomes more prevalent, researchers may need to undergo training

on how to utilise prompts to generate the most effective texts (White et al., 2023). Concerns would arise when researchers do not have access to specialised training, software and equipment that would enable them to make effective use of LLMs such as ChatGPT. It should be acknowledged that ChatGPT may assist non-native English speakers in producing linguistically sound research text; however, certain countries may not have access to working with these AI systems. National policies and financial constraints may act as barriers, which create or magnify inequality across the global research community that would want to take advantage of this AI technology (Xames & Shefa, 2023).

As LLMs grow in popularity, it becomes paramount to incorporate AI literacy education within psychology teaching programs at all levels. AI literacy education refers to the ability to use, monitor and critically reflect on AI application (Long et al., 2021). There is a need for psychology students and researchers to build on their knowledge of current LLM systems. AI literacy is a scarcely studied topic within the literature; Laupichler et al. (2022) explores the value of AI literacy within students and researchers and how systematically incorporating AI application in university curriculums may prove to be useful. There needs to be encouragement to understand the ethical implications of using LLMs, such as ChatGPT, and the benefits and concerns they bring to research. Psychology researchers should engage in regular exposure to AI-generated content as well as collaborating with AI-related departments to build stronger understanding of the AI systems. Due to the topic's novelty, there is a lack of understanding of measures that evaluate AI literacy, and this makes it difficult to implement effective interventions that promote AI literacy (Ng et al., 2021). Nonetheless, there must be a stronger push to explore AI literacy in all facets of education.

Gil (2022) explores the possibility of a future in which researchers view AI systems, such as LLMs, as partners in scientific endeavour. Scientific research is becoming more complex, involving the collaboration of a multitude of scientists and several years to make meaningful discoveries (Aed et al., 2012). There is a genuine function for AI systems to be more actively

involved in the research process for which tedious and/or textual tasks can be delegated to. More research is required to create thoughtful AI systems that will act as powerful partners in collaboration with researchers. Gil (2017) brings attention to this concept and proposes principles that should be upheld to establish LLMs as research partners. Gil (2017) states that information given by LLMs should be governed by knowledge within the AI system; ChatGPT suffices in this regard when provided sufficiently detailed and meaningful prompts. Furthermore, the LLM should not just be limited to a particular task or domain and should be capable of responding to a vast range of queries; ChatGPT has been trained on 570GBs of data (Shen, 2023), responds to a multitude of varying queries, and is therefore likely adequate in this regard.

A matter in which ChatGPT fails to suffice in accordance with Gil (2017) is the network principle. At the time of writing, ChatGPT cannot connect to online resources to inform its responses and it is only trained on data available up to the year 2021 (Shewale, 2023), compromising its ability to act as a research partner. Another significant principle that is considered is ethics; ChatGPT must act in a responsible manner that adheres to research policy; more specifically, it must recognise limitations within its decision making. ChatGPT is trained on unmonitored online data and is therefore prone to making biased assertions. Nonetheless, developers of ChatGPT have made significant strides in combating bias by creating more refined approaches to information output (Harrer, 2023). There is potential for ChatGPT to become an effective research partner, yet as evident, there are also limitations associated with ChatGPT's intrinsic properties that act as barriers to such a future.

Limitations and Future Research

While the results pose significant implications, I acknowledge the limitations within the study. Reproducing results stands as a challenge due to the nature of AI models such as ChatGPT. These models may naturally exhibit variability in their outputs on different occasions despite receiving the same prompt on those occasions (Sordoni et al., 2023). Furthermore, the continuing evolution of ChatGPT modelling will impact how the chatbot interprets inputted

prompts and will therefore produce varying texts. Because of this, future researchers may find it challenging to replicate the output of ChatGPT. I attempted to minimise this issue of reproducibility in two ways: 1) I have publicly provided the prompt framework I used as well as their associated abstracts so that future researchers may use these to generate abstracts of their own; 2) I entered the prompt for each abstract in separate chat sessions to ensure that ChatGPT would not learn from its previous responses to generate text.

Participant fatigue may also have played a role. While the study includes results from 56 participants, 36 additional participants had opened the survey but failed to complete it. It is possible that many participants grew fatigued due to the repetitive nature of the task. Fatigue may subsequently compromise the quality of responses as participants may not have invested equal attention to the abstracts. Future researchers may benefit by actively increasing participant engagement to ensure completion of the survey.

The small sample size limits the generalisability of the results as it is less likely to effectively capture the diverse characteristics of the broader population. This consequently compromises the external validity of the study, limiting its application to real-world situations. The smaller sample size also elevates proneness to sampling bias, which serves to inaccurately represent the characteristics of the larger population (Lin, 2018). Most of the evaluated effect sizes were small to moderate, indicating that the difference between variables is noticeable but not particularly large. In particular, the comparison of the TES between AI-generated and human-authored abstracts resulted in a Cohen's *d* effect size of 0.29. Given the relatively modest sample size, I cannot assert the precision of the effect size; larger sample sizes contribute to providing a more precise estimate of the effect size (Anderson & Maxwell, 2017).

Nonetheless, the effect sizes may not be particularly large primarily due to the difficulty participants experienced in distinguishing between the abstracts. Participants could not identify textual differences between AI-generated and human-authored abstracts, therefore leading to more modest effect sizes. Due to this, the findings cannot be dismissed solely due to the effect

sizes; the observed difficulty in distinguishing between AI-generated and human-authored abstracts is a finding with valuable implications to research practice.

Due to a lack of measures that evaluate text features generated by LLMs, I designed a new measure named the TES that incorporates best practice as outlined by Lee et al. (2021). The TES has not been practically implemented outside of this study so should be used with caution in future studies. The TES lacks the extensive validation that established scales have so concerns regarding the measure's reliability and validity are justified. The validity of the TES may also be threatened by the rapidly ongoing improvement of LLMs.

Despite this, the TES underwent a McDonald's Omega analysis which provided satisfactory evidence that allowed it to be qualified as a single factor. The measure additionally showed high reliability amongst its variables giving sound substantiation to its use as an effective measure in evaluating the textual features of the abstracts. As there is a lack of measures of this type within the literature, the TES exhibits psychometric properties that may render it an effective tool to evaluate texts generated by LLMs.

The novelty of this area of research allowed me to conduct fundamental research regarding LLMs and how psychology students and researchers evaluate and distinguish between AI-generated and human-authored abstracts. I also implemented a new scale, the TES, to compare textual evaluations given to both AI-generated and human-authored abstracts, for which very little exploration has been done prior. Future studies should attempt to implement or revise the TES measure and evaluate its performance on similar tasks. As LLM technology develops, it becomes pivotal to reliably evaluate its linguistic potential. Furthermore, future studies attempting to replicate this study should employ a larger sample size to improve generalisability and better understand the practical significance of findings. As LLMs become more prevalent within research, it becomes important to 1) improve AI literacy amongst students and researchers and 2) evaluate the impact of AI literacy on their ability to distinguish between AI-generated and human-authored abstracts.

References

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological science*, *28*(11), 1547-1562.

Asudani, D. S., Nagwani, N. K., & Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: a review. *Artificial Intelligence Review*, 1-81.

Bates, E., & MacWhinney, B. (2014). Competition, variation, and language learning. In *Mechanisms of language acquisition* (pp. 157-193). Psychology Press.

Buholayka, M., Zouabi, R., & Tadinada, A. (2023). The Readiness of ChatGPT to Write Scientific Case Reports Independently: A Comparative Evaluation Between Human and Artificial Intelligence. *Cureus*, *15*(5).

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, *2*(3), 100068.

Chamberlain, L., & Broderick, A. J. (2007). The application of physiological observation methods to emotion research. *Qualitative market research: an international journal*, *10*(2), 199-216.

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of general psychology*, *130*(3), 290-304.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*.

Crossman, A. M., & Lewis, M. (2006). Adults' ability to detect children's lying. *Behavioral Sciences & the Law*, *24*(5), 703-715.

Cull, W. L., O'connor, K. G., Sharp, S., & Tang, S. F. S. (2005). Response rates and response bias for 50 surveys of pediatricians. *Health services research*, *40*(1), 213-226.

Deng, J., & Lin, Y. (2022). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems*, *2*(2), 81-83.

Dones V. (2022) Systematic Review Writing by Artificial Intelligence: Can Artificial Intelligence Replace Humans?. *Journal of Musculoskelet Disorders and Treatment* 8:112. doi.org/10.23937/2572-3243.1510112

Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters*, *53*, 103662.

Fitria, T. N. (2023). Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. In *ELT Forum: Journal of English Language Teaching* (Vol. 12, No. 1, pp. 44-58).

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.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. *Psychological review*, *109*(1), 75.

Gil, Y. (2017). Thoughtful artificial intelligence: Forging a new partnership for data science and scientific discovery. *Data Science*, *1*(1-2), 119-129.

Gil, Y. (2022). Will AI write scientific papers in the future?. *AI Magazine*, *42*(4), 3-15.

Gordijn, B., & Have, H. T. (2023). ChatGPT: evolution or revolution?. *Medicine, Health Care and Philosophy*, *26*(1), 1-2.

Harrer, S. (2023). Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, *90*.

Haq, Z., Naeem, H., Naeem, A., Iqbal, F., & Zaeem, D. (2023). Comparing human and artificial intelligence in writing for health journals: an exploratory study. *medRxiv*, 2023-02.

Hassani, H., & Silva, E. S. (2023). The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field. *Big data and cognitive computing*, *7*(2), 62.

Herrmann, B. (2023). The perception of artificial-intelligence (AI) based synthesized speech in younger and older adults. *International Journal of Speech Technology*, 1-21.

Hill-Yardin, E. L., Hutchinson, M. R., Laycock, R., & Spencer, S. J. (2023). A Chat (GPT) about the future of scientific publishing. *Brain Behav Immun*, *110*, 152-154.

Intahchomphoo, C., & Tschirhart, C. (2022). The evolution of data and freedom of expression and hate speech concerns with artificial intelligence. *Legal Information Management*, *22*(1), 45-48.

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, *120*(11), e2208839120.

Kitamura, F. C. (2023). ChatGPT is shaping the future of medical writing but still requires human judgment. *Radiology*, *307*(2), e230171.

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of learning and motivation*, *32*, 385-418.

Korinek, A. (2023). *Language models and cognitive automation for economic research* (No. w30957). National Bureau of Economic Research.

Lai, V. D., Van Nguyen, C., Ngo, N. T., Nguyen, T., Dernoncourt, F., Rossi, R. A., & Nguyen, T. H. (2023). Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2307.16039*.

Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence*, 100101.

Lazarus, C., Haneef, R., Ravaud, P., & Boutron, I. (2015). Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC medical research methodology*, *15*, 1-8.

Lerman, D. C., Tetreault, A., Hovanetz, A., Bellaci, E., Miller, J., Karp, H., ... & Toupard, A. (2010). Applying signal-detection theory to the study of observer accuracy and bias in behavioral assessment. *Journal of applied behavior analysis*, *43*(2), 195-213.

Lin, L. (2018). Bias caused by sampling error in meta-analysis with small sample sizes. *PloS one*, *13*(9), e0204056.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-16).

Macdonald, N. (1954). Language translation by machine-a report of the first successful trial. *Computers and automation*, *3*(2), 6-10.

Maučec, M. S., & Donaj, G. (2019). Machine translation and the evaluation of its quality. *Recent trends in computational intelligence*, *143*.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, 100041.

Pastore, R. E., & Scheirer, C. J. (1974). Signal detection theory: Considerations for general application. *Psychological Bulletin*, *81*(12), 945.

Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, *20*(2), 07.

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).

Sanganyado, E. (2019). How to write an honest but effective abstract for scientific papers. *Scientific African*, *6*, e00170.

Schwartz, S., Yaeli, A., & Shlomov, S. (2023). Enhancing Trust in LLM-Based AI Automation Agents: New Considerations and Future Challenges. *arXiv preprint arXiv:2308.05391*.

Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and other large language models are double-edged swords. *Radiology*, *307*(2), e230163.

Shewale, R. (2023, September). *"32 Detailed ChatGPT Statistics - Users, Revenue, and Trends"*. Demand Sage. https://www.demandsage.com/chatgpt-statistics/

Sordoni, A., Yuan, X., Côté, M. A., Pereira, M., Trischler, A., Xiao, Z., ... & Roux, N. L. (2023). Deep Language Networks: Joint Prompt Training of Stacked LLMs using Variational Inference. *arXiv preprint arXiv:2306.12509*.

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning.

Tang, R., Chuang, Y. N., & Hu, X. (2023). The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Uchendu, A., Ma, Z., Le, T., Zhang, R., & Lee, D. (2021). Turingbench: A benchmark environment for turing test in the age of neural text generation. *arXiv preprint arXiv:2109.13296*.

Van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature*, *614*(7947), 224-226.

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, *67*, 101151.

Wang, Q., Liao, J., Lapata, M., & Macleod, M. (2022). Risk of bias assessment in preclinical literature using natural language processing. *Research synthesis methods*, *13*(3), 368-380.

Wei, J., Kim, S., Jung, H., & Kim, Y. H. (2023). Leveraging large language models to power chatbots for collecting user self-reported data. *arXiv preprint arXiv:2301.05843.*

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Xames, M. D., & Shefa, J. (2023). ChatGPT for research and publication: Opportunities and challenges. *Available at SSRN 4381803*.

Zeelenberg, R., & Pecher, D. (2015). A method for simultaneously counterbalancing condition order and assignment of stimulus materials to conditions. *Behavior research methods*, *47*, 127-133.

Zhai, X. (2022). ChatGPT user experience: Implications for education. *Available at SSRN 4312418*.

Appendix A

List of studies whose abstracts were used in stimuli

1. Loughnan, R. J., Palmer, C. E., Thompson, W. K., Dale, A. M., Jernigan, T. L., & Chieh Fan, C. (2023). Intelligence Polygenic Score Is More Predictive of Crystallized Measures: Evidence From the Adolescent Brain Cognitive Development (ABCD) Study. *Psychological Science*, *34*(6), 714-725.
2. Gurven, M., Von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013). How universal is the Big Five? Testing the five-factor model of personality variation among forager–farmers in the Bolivian Amazon. *Journal of personality and social psychology*, *104*(2), 354.
3. Parham, T. A., & Helms, J. E. (1981). The influence of Black students' racial identity attitudes on preferences for counselor's race. *Journal of counseling psychology*, *28*(3), 250.
4. Akhtar, H., & Kovacs, K. (2023). Which tests should be administered first, ability or non-ability? The effect of test order on careless responding. *Personality and Individual Differences*, *207*, 112157.
5. Geurten, M., Catale, C., & Meulemans, T. (2015). When children's knowledge of memory improves children's performance in memory. *Applied Cognitive Psychology*, *29*(2), 244-252.
6. Moore, K. N., Lampinen, J. M., Nesmith, B. L., Bridges, A. J., & Gallo, D. A. (2022). The effect of feedback and recollection rejection instructions on the development of memory monitoring and accuracy. *Journal of Experimental Child Psychology*, *221*, 105434.
7. Wang, X., Cai, L., Qian, J., & Peng, J. (2014). Social support moderates stress effects on depression. *International journal of mental health systems*, *8*(1), 1-5.

8. Cemalcilar, Z., Canbeyli, R., & Sunar, D. (2003). Learned helplessness, therapy, and personality traits: An experimental study. *The Journal of social psychology*, *143*(1), 65-81.
9. Head, J., & Helton, W. S. (2014). Sustained attention failures are primarily due to sustained cognitive load not task monotony. *Acta psychologica*, *153*, 87-94.
10. Wang, Y., Zhang, J., & Lee, H. (2021). An online experiment during COVID-19: Testing the influences of autonomy support toward emotions and academic persistence. *Frontiers in Psychology*, *12*, 747209.

Appendix B

Example prompt used to generate AI abstract

Study source**:** Wang, X., Cai, L., Qian, J., & Peng, J. (2014). Social support moderates stress effects on depression. *International journal of mental health systems*, *8*(1), 1-5.

Prompt:

Write a scientific abstract for the article titled [Social support moderates stress effects on depression] in the style of the journal, [International Journal of Mental Health Systems]. The abstract should summarise the research aims/questions, main findings, methods, and implications of the study in a clear and concise manner that flows. Consider the target audience of the journal and use appropriate terminology and language. The abstract should be no more than 250 words in length. The length should be just one paragraph.

Adhere to the following information in generating the abstract:

Aim/Question: [one sentence]

Method: Hierarchical regression analysis to determine association between perceived stress, perceived social support, and depression

Participants: 632 undergraduate students

Results: Hierarchical regression analysis showed that social support moderated the association between stress and depression; undergraduate students with high stress reported higher scores in depression than those with low stress with low social support level.

Conclusion: [one sentence]

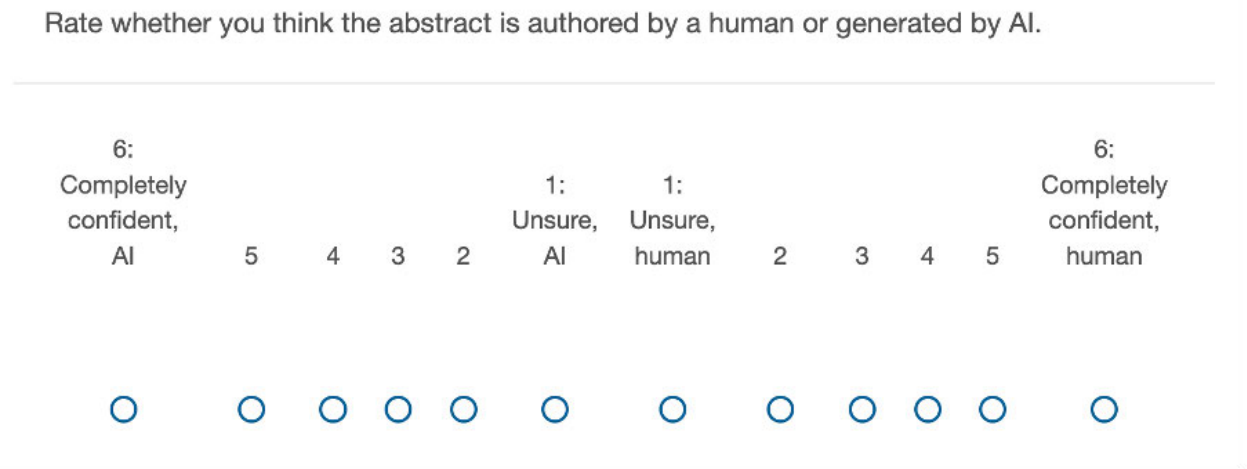
Note that Method, Participant, and Results section of prompts is different for each study.

Appendix C

Confidence scale layout

Figure C1

Forced-choice confidence scale layout

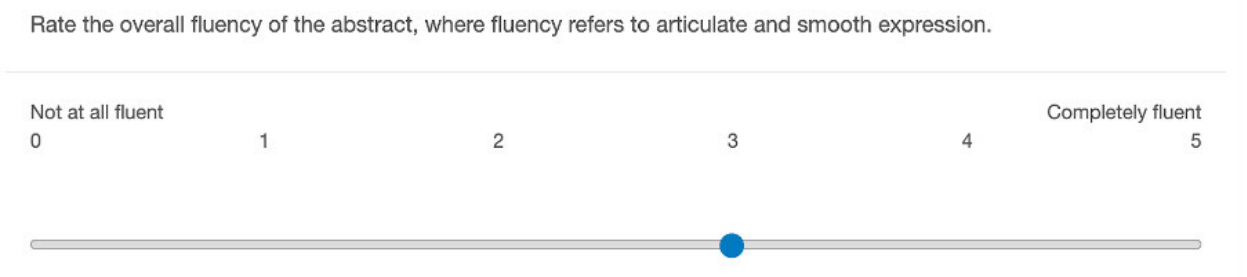


Appendix D

Visual depiction of example item, Fluency, in the TES

Figure D1

Visual depiction of the item, Fluency



Appendix E

Demographic information requested by survey

Figure E1

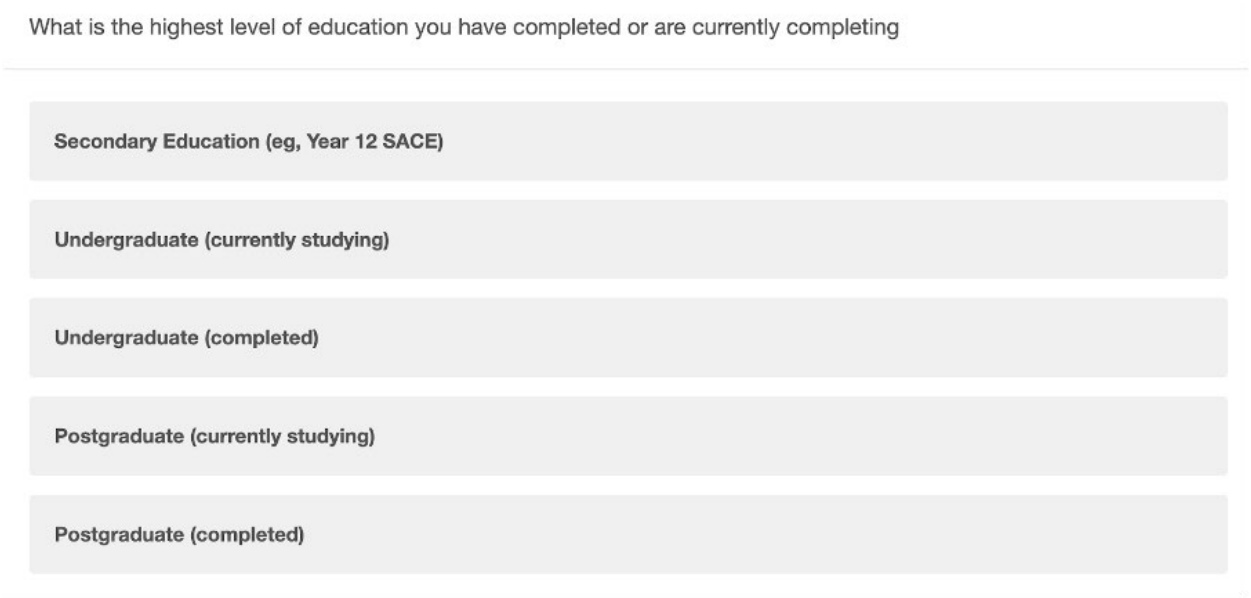
Gender demographic

Figure E2

Age demographic

Figure E3

Level of researcher experience demographic



Appendix F

One-Way ANOVA Assumption tests for d-prime

Figure F1

Assumption test 1: Shapiro-Wilk test

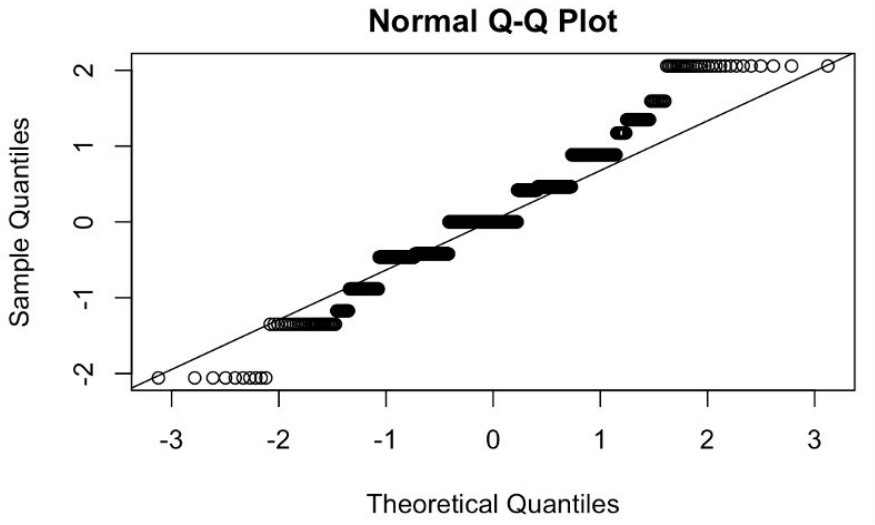
```
        Shapiro-Wilk normality test

data:  ai_abstracts_SDT_data$dprime
W = 0.96381, p-value = 1.633e-10
```

Note. Output indicates non-normality

Figure F2

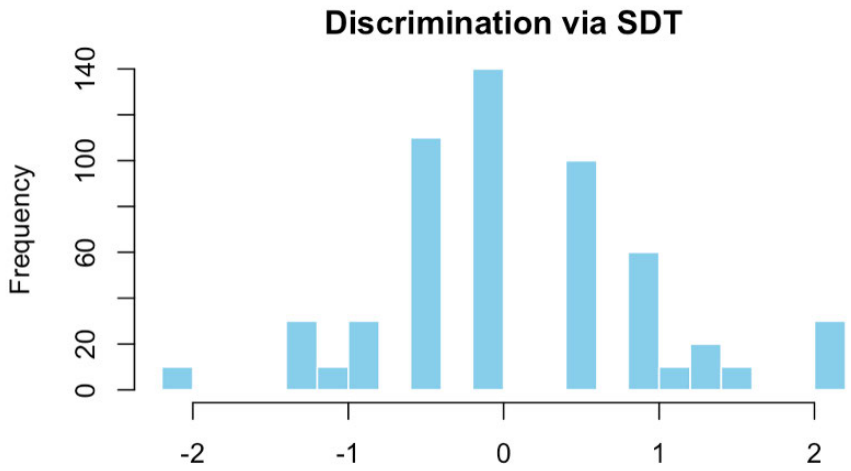
Assumption test 2: QQ Plot



Note. Output indicates normality

Figure F3

Assumption test 3: histogram



Note. Output indicates normality

Figure F4

Assumption test 4: Bartlett Test

```
        Bartlett test of homogeneity of variances

data:  dprime by Demo_Education
Bartlett's K-squared = 26.546, df = 2, p-value = 1.72e-06
```

Note. Output indicates assumption of variance violated

Appendix G

One-Way ANOVA Assumption tests for response bias

Figure G1

Assumption test 1: Shapiro-Wilk Test

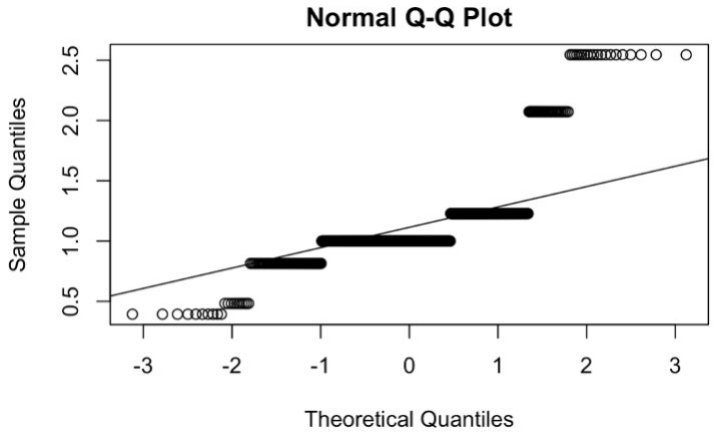
```
        Shapiro-Wilk normality test

data:  ai_abstracts_SDT_data$beta
W = 0.68362, p-value < 2.2e-16
```

Note. Output indicates non-normality

Figure G2

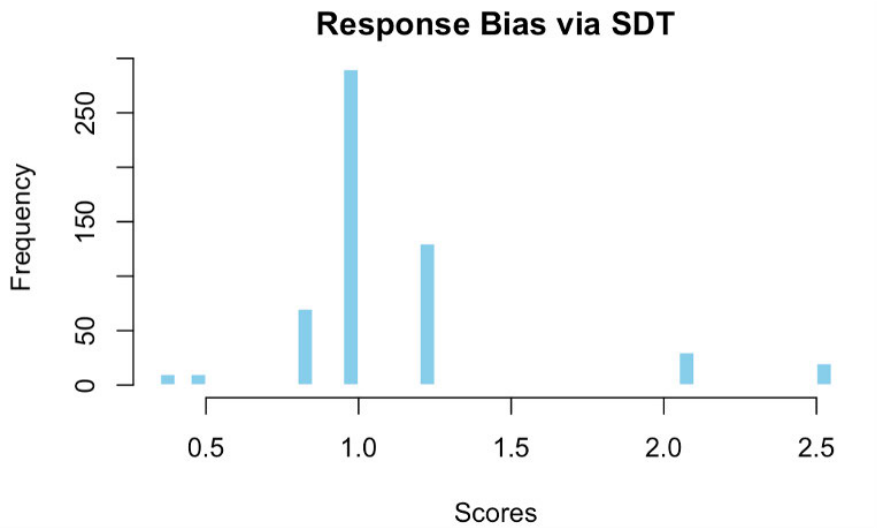
Assumption test 2: QQ Plot



Note. Output indicates non-normality

Figure G3

Assumption test 3: histogram



Note. Output indicates non-normality

Figure G4

Assumption test 4: Bartlett Test

```
        Bartlett test of homogeneity of variances

data:  dprime by Demo_Education
Bartlett's K-squared = 26.546, df = 2, p-value = 1.72e-06
```

Note. Output indicates assumption of variance satisfied

Appendix H

Paired t-test assumptions

Figure H1

Assumption test 1: Shapiro-Wilk Test

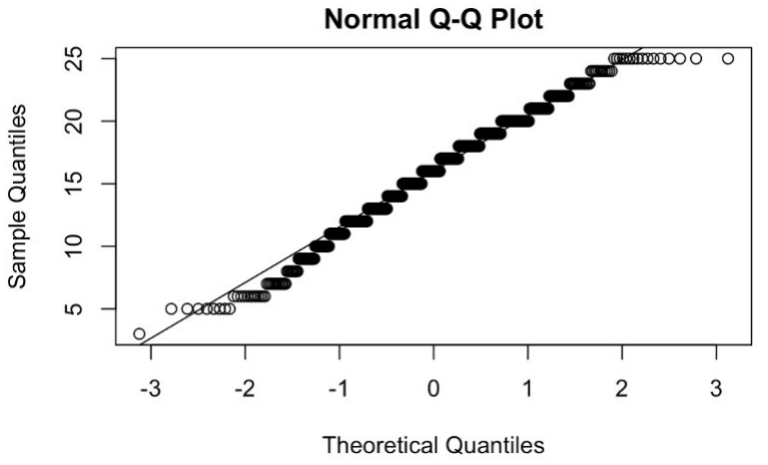
```
        Shapiro-Wilk normality test

data:  ai_abstracts_cleaned_data_complete$Sum_of_Quality_Variables
W = 0.98361, p-value = 6.175e-06
```

Note. Output indicates non-normality

Figure H2

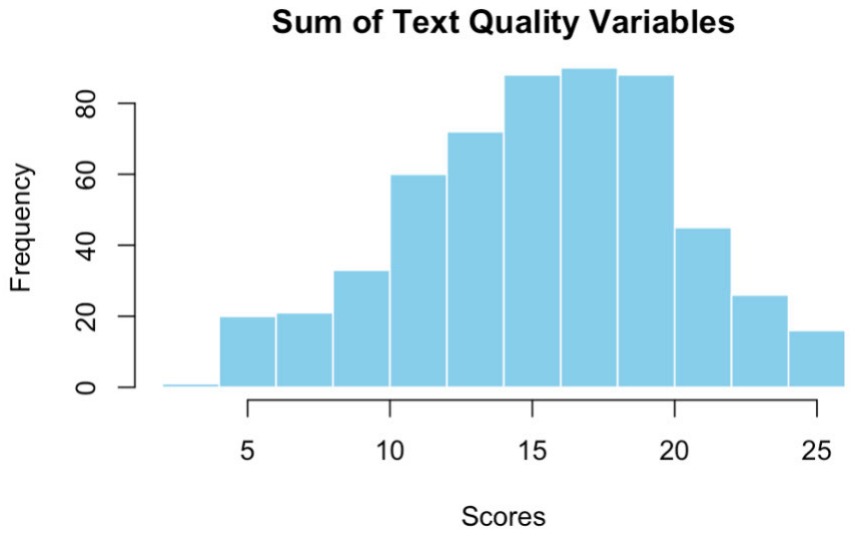
Assumption test 2: QQ Plot



Note. Output indicates normality

Figure H3

Assumption test 3: histogram



Note. Output indicated normality