Research Methods in Applied Linguistics 2 (2023) 100068

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing

J. Elliott Casal [a,*], Matt Kessler [b]

[a] Department of English (Institute for Intelligent Systems Affiliate), The University of Memphis, Memphis, TN, USA
[b] Department of World Languages, University of South Florida, Tampa, FL, USA

ARTICLE INFO

Keywords:
Research Abstracts
Academic Publishing
Artificial Intelligence
ChatGPT
Research Ethics

ABSTRACT

There has been considerable intrigue surrounding the use of Large Language Model powered AI chatbots such as ChatGPT in research, educational contexts, and beyond. However, most studies have explored such tools' general capabilities and applications for language teaching purposes. The current study advances this discussion to examine issues pertaining to human judgements, accuracy, and research ethics. Specifically, we investigate: 1) the extent to which linguists/reviewers from top journals can distinguish AI- from human-generated writing, 2) what the basis of reviewers' decisions are, and 3) the extent to which editors of top Applied Linguistics journals believe AI tools are ethical for research purposes. In the study, reviewers ($N = 72$) completed a judgement task involving AI- and human-generated research abstracts, and several reviewers participated in follow-up interviews to explain their rationales. Similarly, editors ($N = 27$) completed a survey and interviews to discuss their beliefs. Findings suggest that despite employing multiple rationales to judge texts, reviewers were largely unsuccessful in identifying AI versus human writing, with an overall positive identification rate of only 38.9%. Additionally, many editors believed there are ethical uses of AI tools for facilitating research processes, yet some disagreed. Future research directions are discussed involving AI tools and academic publishing.

1. Introduction

The field of Applied Linguistics is maturing in a time of important ethical discussions in educational and scholarly domains. Applied linguists are celebrating the methodological strides made in recent decades (Gass et al., 2021; Plonsky, 2014) while calling for further commitments to elevate methodological rigor and critically interrogate issues of research ethics (e.g., De Costa et al., 2021; Kubonyiova, 2008). At the same time, like other fields, applied linguists are grappling with the emergence of freely available Large Language Model (LLM) powered AI chatbots (e.g., OpenAI's ChatGPT, Microsoft's Bing, Google's Bard), which have the power to quickly generate text that is designed to resemble human-produced language. Scholarly and practical concerns abound, with scholars across disciplines tuning in to language-related ethical concerns and scrambling to find pedagogical applications (e.g., Kasneci et al., 2023; Kohnke et al., 2023; Yang et al., 2022), attempting to understand the potential uses in misinformation campaigns (e.g., Kreps et al., 2022), and investigating the extent to which AI-produced texts differ from human-produced texts (e.g., Ma et al., 2023). The later issue is the basis for the primary focus of the current study. Notably, previous studies have aimed to profile the linguistic differences between

* Corresponding author at: Patterson Hall, The University of Memphis, Memphis, TN, USA.
E-mail address: jecasal@memphis.edu (J.E. Casal).

https://doi.org/10.1016/j.rmal.2023.100068
Received 3 June 2023; Received in revised form 18 July 2023; Accepted 18 July 2023
Available online 7 August 2023

AI- and human-produced texts, and to some extent, the more practical concern of whether humans can tell the difference. In the current study, we are interested not only in whether humans – in this case linguists who review for top journals in the field – can reliably distinguish AI- from human-produced texts, but what the basis for the distinction is. Further, and returning to the issue of research ethics, we are also concerned with how editors of these journals view the ethics of using LLM-backed technologies in the production of scholarly research and writing.

In this study, we focus on research abstracts as the target genre, a decision that is both practical and theoretical. First, LLM-backed technologies such as ChatGPT (GPT4 Plus) have a limited amount of word tokens they can produce, and it is widely known that longer texts are more prone to hallucination (i.e., the tendency to invent content). At the same time, considerable scholarship on research article abstracts and conference abstracts has found that they are rhetorically predictable (e.g., Halleck & Connor, 2006; Samar et al., 2014; Yoon & Casal, 2020) and, perhaps due to their constrained length requirements, characterized by highly formulaic language (e.g., Casal & Yoon, 2023; Omidian et al., 2018). Thus, we have reason to believe that abstracts are a genre that may be particularly vulnerable to AI-generation. Abstracts are also a relevant and critical genre because they play a primary role in editors' initial decision-making processes, reviewers' choices to accept reviews, and in the promotion of one's work. After all, reviewers often receive only the abstract and title before making a decision to accept/decline a review, and the abstract is also part of only a small amount of information available on the free side of scholarly paywalls. For these reasons and more, in the current study, we investigate (a) the extent to which experienced scholars can tell the difference between AI- and human-produced abstracts, (b) what the basis for scholars' decisions are, and (c) the extent to which editors of journals in Applied Linguistics believe the use of AI is ethical for research purposes.

2. Literature review

2.1. AI Chatbots in recent scholarship

For detailed discussions of the history of Large Language Models (LLMs) that power modern generative chatbots or the particular workings of specific tools, readers are directed towards the considerable wealth of scholarship from other disciplines on the development, application, and drawbacks of such technologies over time (e.g., Bender et al., 2021), as well as to the websites of current versions of such tools. Nevertheless, in this section, we provide an overview of the topic with an eye towards the first publications on ChatGPT (and similar tools) in Applied Linguistics and fields with related interests.

The potential benefits, ills, and perceived inevitabilities resulting from the existence of freely available AI chatbots powered by LLMs have become pervasive talking points in the public eye and across disciplinary domains, particularly with the 2022 launch of ChatGPT. Despite much recent intrigue, the technology had been gaining enough steam for Bender et al.'s (2021) high profile article to address the "development and deployment of ever larger language models" (p. 10) and to consider the consequences of such growth. Generally speaking, AI chatbots are generative tools which are designed based on advances in natural language processing and deep learning models to respond to human inquiries. This is accomplished through algorithms that leverage highly parameterized and massive linguistic datasets. GPT 3, for example, had 175 billion parameters on an undisclosed number of texts from across the internet, which can be confirmed by asking ChatGPT.[1]

Our widespread anecdotal experience of AI Chatbots is that public interest largely targets what the capacities of such technologies are and how this may impact a variety of current practices. Much of what extant scholarship there is aligns with this, as early commentaries in scholarly venues often emphasize the general ubiquity of such novel tools (e.g., Kurian et al., 2023 in Dentistry), or the balance of their potential affordances and dangers (e.g., Kasneci et al., 2023 in Education; Kohnke et al., 2023 in Language Teaching; Shen et al., 2023 in Medicine). A trend in these studies, which is emblematic of their publication in the early stages of these discussions, is that AI Chatbots have immense potentials including general productivity, richness in interactivity, and myriad educational affordances. Conversely, discussions highlight fears that they may also carry dire risks including human professional replacement, intellectual offloading, and misinformation. An exception to these discussions of broad affordances and general concerns is Ma et al. (2023), who include a small-scale investigation of how two graduate students in Computer Science differentiated between AI- and human-produced abstracts and wiki item descriptions as part of a broader analysis of formal differences in AI- and human-produced writing. They found that their two human participants identified AI writing by a lack of concreteness in research motivation and methods, but the scope of the analysis was notably small-scale.

Although scholarly publications on AI Chatbots such as ChatGPT in Applied Linguistics and beyond are scarce (but undoubtedly more are on the way), it is important to consider how such tools will impact our productivity as researchers and educators, what affordances and challenges they will offer language learners and language users, and how they may invite some of us to reconsider or recalibrate theoretical concepts and pedagogical practices in the field. And yet, it seems essential that before entering into such discussions, there must be ethical discussions surrounding how we, as a field, want to proceed in these matters.

2.2. Research ethics in applied linguistics and with the use of AI

As discussed, scholarship investigating the affordances of AI chatbots and similar tools in Applied Linguistics is scant, particularly when it comes to understanding their capacity to support various facets of the research process. Simultaneously, there is also a pressing

[1] Although GPT 4 was available when the current article was written, OpenAI had yet to reveal the total number of parameters.

need to explore the use of these tools in conjunction with research ethics. That is, despite much excitement and intrigue surrounding the extent to which AI chatbots *can* be used, there has been little discussion of ethics and the extent to which these tools *should* be used for research purposes.

Within Applied Linguistics in particular, there has been a surge of interest in research ethics during the past decade. In a synthesis and timeline by Yaw et al. (2023), the authors noted that such interest stems from the fact that there is a growing realization that "the entire research cycle—from conceptualization to design and data collection, to analysis, writing, and dissemination—is laden with decisions that can be viewed through the lens of research ethics" (p. 1). This has prompted many scholars to attempt to better understand general research practices across the field at-large, in addition to improving their own practices, respectively (e.g., De Costa et al., 2021; Isbell et al., 2022; Sterling & Gass, 2017). Within the scholarship on research ethics, multiple studies such as Isbell et al. (2022) and Larsson et al. (2023) have broadly been interested in what have been deemed *Questionable Research Practices* (QRPs). These QRPs consist of numerous activities that researchers may engage in, which range from clearly defined, unacceptable instances of misconduct (e.g., data fabrication) to relatively less clear, ethical gray areas (e.g., reporting effect sizes for significant results only). Much recent attention has been paid to the topic of QRPs, and rightfully so. However, comparatively speaking, fewer scholars have formally investigated ethical issues involved with research that leverages digital tools and technologies. As some prior studies in this area have noted (e.g., Kessler et al., in press; Marino et al., in press; Spilioti & Tagg, 2017), when it comes to technology, research ethics are particularly important. This is because both digital tools and online environments are increasingly presenting novel ethical challenges, such as determining what constitutes public vs. private data, whether acquiring informed consent is necessary (or even possible) in certain situations, and more. Relatedly, research ethics with technology are crucial because determining what constitutes ethical behavior often falls on the researchers themselves since university and institutional review boards may lag behind new developments in technology.

In terms of ethical uses of technology – particularly with AI – most of the (published) dialogues to date have occurred among researchers who work directly with AI's development in areas such as Computer Science and Machine Learning (e.g., Hagendorff, 2020; Jobin et al., 2019; Mittelstadt, 2019; Siau & Wang, 2020). In a study by Hagendorff (2020), the researcher synthesized 22 different sets of guidelines produced on the ethical use of AI tools, highlighting similarities among them. Across guidelines that were produced within governmental, industrial, or academic contexts, Hagendorff identified recurring ethical issues that were discussed, including *accountability* (i.e., ascertaining the parties responsible when something goes wrong), *privacy* (i.e., protecting users' data), and *fairness* (i.e., minimizing bias and promoting inclusion). However, as Hagendorff pointed out, "it is noteworthy that almost all guidelines suggest that technical solutions exist for many of the problems described…[yet few] contain genuinely technical explanations at all" (p. 104). In sum, Hagendorff notes that although potential ethical issues are widely acknowledged by AI designers and users, few (if any) concrete solutions are ever provided.

In returning to the field of Applied Linguistics, as noted, prior studies have tended to focus on exploring AI's general affordances and capabilities rather than engaging in discussions of ethics. One notable exception is Kohnke et al. (2023). In their technological review of ChatGPT for language learning/teaching purposes, the authors provided their views regarding three potential ethical issues. These issues included general concerns of (1) students' cheating on assessments (e.g., AI producing students' papers), (2) AI producing inaccurate or false information, and (3) AI potentially perpetuating cultural biases (i.e., since many systems, including earlier versions of ChatGPT, rely primarily on English). Apart from Kohnke et al.'s work, empirical research is lacking that investigates the potential ethical issues with AI's use, particularly when it comes to AI's adoption in the research process. Studies are needed that attempt to understand the extent to which gatekeepers of scientific knowledge (such as experienced researchers and journal editors) believe that AI's use is (un)ethical, including what appropriate uses AI may have for research and publishing purposes.

2.3. The current study

With these concerns in mind, the current study targets the ability of reviewers of top journals in Applied Linguistics to distinguish between AI- and human-produced research article abstracts, what they report to base their decisions on, and what editors of top journals consider to be ethical and unethical use of AI chatbots in publication practices. The research questions (RQs) guiding the current study are:

RQ1: To what extent can reviewers tell the difference between written abstracts produced by ChatGPT/AI versus humans?
RQ2: What is the basis of reviewers' decisions (i.e., their rationales or criteria) for determining whether writing is produced by ChatGPT/AI or by humans?
RQ3: What are journal editors' views regarding the acceptable use of ChatGPT and similar AI/LLM tools for academic publishing purposes?

3. Method

To address the three RQs, we collected survey and semi-structured interview data. For RQs 1-2, a pool of frequent article reviewers was compiled of the editorial board members for the top 30 journals in Applied Linguistics using SCImago Journal Rank (SJR). Participants from this category are henceforth referred to as *reviewers*. For RQ3, a pool of editors (including co-editors and assistant editors), was compiled from the top 40 journals in Applied Linguistics using SCImago Journal Rank (SJR). Participants in this category are henceforth referred to as *editors*. Participants who were listed more than once in either list were only contacted once for each pool. Semi-structured interviews were conducted as follow-ups to the surveys for RQs 2-3.

3.1. Participants

For RQ1 (i.e., reviewers' capacity to differentiate between ChatGPT/AI- versus human-produced abstracts), 87 reviewer participants began the survey (described in detail in the next section), but only 72 completed it and are included in this analysis. The 72 reviewers self-reported an average age of 46.3 years ($SD = 11.7$, range 28-75 years). In terms of gender, they identified as female (39), male (31), or other (2). Regarding their professional profile, the majority indicated they currently held various tenure-line professor titles, such as assistant professor (12), associate professor (17), full professor (6), or professor with no rank specified (22). In addition, some held emeritus status (5), lecturer or senior lecturer positions (5), and the remaining indicated a variety of research-oriented positions. The reviewers represent a wide-range of research areas, such as L2 writing (15), second language acquisition, instructed second language acquisition, or TESOL (11), psycholinguistics (7), L2 pronunciation/phonology (7), corpus linguistics (6), CALL (3), L2 vocabulary (3), and a number of other areas listed by one or two reviewers (e.g., conversation analysis, discourse analysis, education, genre studies, identity, literacy studies, pragmatics, teacher education, sociocultural theory, and sociolinguistics). When asked about the number of reviews they had completed for journals within the last 12 months, participants indicated they had completed between 1-3 reviews (19 people), 4-6 reviews (20), 7-9 reviews (11), or 10 or more reviews (22).

Of the 72 survey participants, 28 volunteered to participate in follow-up interviews for RQ2 (i.e., discussing the basis of reviewers' decisions). From this potential pool of 28 reviewers, 10 were contacted because they reflected the diversity in our dataset. Specifically, efforts were made to contact reviewers who broadly represented the field of Applied Linguistics, both in terms of biographical attributes (e.g., different ages, genders) and in terms of research specializations. Of the 10 potential reviewers who were contacted, seven ultimately participated. These seven reviewers (see Table 1) were roughly balanced in gender (4 male, 3 female), and they had completed a reasonable-to-high number of reviews over the previous 12 months. The interview participants were also roughly representative of the sample's ability to distinguish between AI- and human-produced texts, with most correctly identifying two of the four texts they were presented (to be discussed later).

For RQ3 (i.e., editors' views regarding the ethical use of ChatGPT/AI tools), 35 journal editors began the survey, but only 27 completed it and had their responses included in the analysis. The 27 editors self-reported an average age of 49.6 years ($SD = 9.6$, range 36-80 years), with a greater number of males (17) than females (9) and one participant choosing not to disclose. The editors were not asked to be as specific in identifying their subfields as the reviewer population, since the lower number of potential participants might render editor participants identifiable. The average number of years of editorial experience was 7.2 years ($SD = 7.4$ years) with a broad range (1-40 years). Of the 27 editor participants, seven indicated a willingness to participate in follow-up semi-structured interviews. Although all seven were contacted, five ultimately participated. Editor participant information is kept general in what follows, as to not clue in identity given the small pool and general notoriety of individuals.

3.2. Instruments

3.2.1. Reviewer and editor surveys

Two surveys were designed and used for this study with Qualtrics (Qualtrics, Provo, UT). For RQ1, after reading the instructions, participants from the reviewer pool completed an 11-question survey which took approximately 15 minutes to complete. First, participants were asked a series of general questions pertaining to demographic and professional information. Next, participants were given a brief overview of an AI- and human-produced research abstract identification task and asked to indicate their confidence in distinguishing AI- from human-produced texts of this genre. They were then shown four abstracts (described below) from a pool of eight possible texts and were asked to indicate whether the author was AI, human, or if they were not sure using a 7-point Likert scale (*Definitely AI, Likely AI, Possibly AI, Not Sure, Possibly Human, Likely Human, Definitely Human*). Finally, participants were asked if they would be willing to participate in a semi-structured interview to discuss their responses in greater depth (by providing their email address).

Eight research abstracts were used in this study, including four human-generated abstracts and four AI-generated abstracts that corresponded to the human abstracts. For the human abstracts, four articles were selected from two journals on the basis that they were published in 2021 or 2022 in indexed, peer-reviewed journals of Applied Linguistics with impact factors of approximately 1.0. Recent articles were selected (i.e., 2021-2022) to reduce the likelihood that they had been widely read.[2] Peer-reviewed journals with impact factors of 1.0 were used to ensure a base level of quality with the abstracts (and corresponding articles) yet were also unlikely to be widely read/cited as to influence the results or reduce our reviewer pool. Two qualitative and two quantitative abstracts were selected, but details of the articles will not be shared in order to protect the authors from any undue consequences of reviewers' analyses. Reviewers were told that the articles would not be discussed in depth individually in this article to ensure that their comments regarding writing practices could not be interpreted specifically as feedback or evaluation of their peers.

Since the four human-produced abstracts were published along with four research articles, authors' names were removed. For this reason, reviewers were asked to skip judging an abstract in the event they had encountered the abstract/paper before (i.e., simply by hitting the 'next' arrow key and not entering a judgement on that abstract). No reviewers skipped any abstracts. The AI-produced abstracts were generated using ChatGPT (GPT4 Plus, OpenAI, 2023) on March 29, 2023 based on each of the four human-produced research articles. Due to the overall text processing limit for individual queries, the articles were first broken down into part-genres

[2] Additionally, since ChatGPT was not introduced until November 2022, this made it highly unlikely that the published journal abstracts had been produced by AI.

Table 1
Semi-structured interview participants (i.e., reviewers) for RQ2.

Participant	Age	Area	Reviews completed in past 12 months	Confidence in identifying AI vs. human writing	Accuracy
Reviewer 1	30	Corpus	10+	Somewhat Unconfident	50%
Reviewer 2	73	L2 Writing	4-6	Somewhat Unconfident	50%
Reviewer 3	37	Teacher Education	1-3	Somewhat Unconfident	50%
Reviewer 4	40	L2 Pronunciation	10+	Somewhat Confident	0%
Reviewer 5	42	Instructed-SLA	7-9	Fairly Confident	50%
Reviewer 6	56	SLA	10+	Somewhat Unconfident	50%
Reviewer 7	32	L2 Writing	4-6	Somewhat Confident	50%

Note: Confidence column reflects the reported confidence of the participant using the scale: *Completely Unconfident, Fairly Unconfident, Somewhat Unconfident, Neutral, Somewhat Confident, Fairly Confident, and Completely Confident* (discussed further in Table 2); Accuracy is the percent correctly identified of the four abstracts considered, with Likert ratings not factored in.

(i.e., Introduction/Literature Review, Methods, Results, Discussion), and ChatGPT was asked to summarize each of the part-genres. Then ChatGPT was prompted to create a research abstract following the journal's guidelines (e.g., length and content; not reproduced for author privacy) based on the part-genre summaries. No edits were made to the human- or AI-produced abstracts.

The editor survey for RQ3 was an 8-question instrument that took editors approximately 4 minutes to complete. Participants were asked for information regarding demographics and editorial profiles. Editors were then asked to indicate if they felt, as a journal editor (but not as a representative of their journal), that the use of AI tools such as ChatGPT was ethical in research activity related to a journal article submission a) under no circumstances, b) under specific circumstances, or c) in all circumstances. They were also asked to select which activities they considered to be ethical in research activity and writing for publication in academic journals (i.e., writing the abstract for a study; writing the majority or all of the main manuscript; writing specific parts of the main document; writing a public-facing summary of the article; analyzing data for the study; writing computer code of some kind for the study; editing text; as well as 'other' uses they could specify). Finally, editors were asked to comment generally on issues of ethics and the use of ChatGPT and similar tools in research activity and publication. Editors were also invited to provide information for a follow-up semi-structured interview.

3.2.2. Reviewer and editor semi-structured interviews

For participants of both reviewer and editor pools who indicated a willingness to participate in semi-structured interviews, those interviews were conducted within one week of completing the survey, or shortly afterwards. The goal of the reviewer interviews was to gain insights into the basis for distinguishing between human- and AI-produced abstracts, and the goal of the editor interviews was to better understand aspects of their responses.

For the reviewer interviews, after consent and procedures were discussed, participants were asked to discuss their prior experiences with ChatGPT or similar tools (e.g., Google's Bard). They were then shown each of the abstracts they had previously viewed in the survey and their previous human/AI author decision on the Likert-scale. They were given time to review the abstract and asked to discuss what features of the abstract informed their decision, including any concrete elements and/or vague holistic impressions. Participants were then asked more generally about the basis for distinguishing between AI- and human-produced texts, after which they were shown whether each of the four texts was produced by a human or AI. Participants were given time for a final reflection.

For the editor interviews, participants were asked about their previous experience with ChatGPT and similar tools. They were also asked probing questions about their survey responses, including their rationales for why the items they had identified in the survey were (un)ethical, principles they would use to guide decisions of ethicality, and a general reflection on how they foresee (or currently experience) changes in their editorial role brought about from ChatGPT and AI-based technologies.

The first author conducted all interviews using synchronous video communication tools. Each interview was audio recorded and lasted approximately 30 minutes for reviewers and 15 minutes for editors.

3.3. Analysis

For RQ1, reviewers' responses from the survey were aggregated and presented numerically using simple descriptive statistics, as this allowed for reflections on accuracy of author identification and other patterns. For RQs 2-3, the second author transcribed all audio interviews using verbatim transcription in order to enable qualitative thematic analysis (see Polio & Friedman, 2017), which was guided by a two-cycle coding model (e.g., Miles et al., 2018). In the first cycle of coding, the second author created descriptive and in-vivo codes to reflect the topics of individual sentences discussed by reviewers (in RQ2) or editors (in RQ3). In the second round of coding, similar codes were then grouped together with similar themes. To ensure the reliability of the coding, inter-coder reliability was obtained. Using the thematic codes developed by the second author, the first author then double-coded 54.5% of the interview data, achieving high reliability (86.1%). Any/all discrepancies were resolved via discussions. In the results section for RQ2, we showcase all codes that were developed, including illustrative examples for each code. For RQ3 involving editors' comments, we present several codes and interview excerpts in a supporting role to the survey data as a means of interrogating and accounting for the observed patterns.

4. Results

4.1. RQ1: Reviewers' ability to identify ChatGPT/AI versus human writing

The first aim was to identify the extent to which reviewers from top journals in Applied Linguistics were able to distinguish between AI-produced research abstracts and human-produced research abstracts. Before asking participants to identify the origin of the abstracts, they were asked to rate their confidence in successfully completing the task on a 7-point Likert scale. As can be seen in Table 2, 77.7% of all respondents either adopted a neutral position or the lowest degree of confidence in either direction (neutral or somewhat confident/unconfident), although slightly more participants leaned towards confidence in their ability to distinguish ($n = 29$) than inability to distinguish ($n = 25$). Overall, this paints a broad landscape of uncertainty among even trained linguists, which resonates with broader cultural apprehension.

Next, reviewers completed the identification task, in which they were shown a random sample of four total abstracts, one at a time. Reviewers' identification accuracy and confidence levels for this task are presented in Fig. 1 (human-produced abstracts) and Fig. 2 (AI-generated abstracts). In each figure, the X-axis (called 'Text Label') represents the individual text that was shown to participants (i.e., human abstract #1, #2, etc.). The colored bars spanning the Y-axis (called 'Cumulative Percent') show the percentage of participants who assigned specific ratings (i.e., human or AI), with a black bar seperating those who correctly or incorrectly identified the origin of the text, regardless of confidence level, with 'not sure' responses counted as incorrect identification. Human- and AI-produced texts are seperated into two figures to avoid inversion of color schemes and to reduce the complexity of interpretations.

As can be seen, overall, the reviewers were not particularly effective, with an average total positive identification rate of only 38.9%. Notably, the reviewers were more effective at identifying human authors of research abstracts (44.1%) than AI authors of abstracts (33.7%). The raters only reached 50% accuracy when rating one abstract, which was Human Abstract 3. Thus, it is readily observable that experienced reviewers of Applied Linguistics research article manuscripts are not particularly effective at identifying AI-produced texts under these circumstances. Further, they are less likely to correctly identify AI-produced texts as AI than they are to identify human-produced texts as human.

It is also noteworthy that there is a pronounced difference in reviewer identification abilities and accuracy on a text-to-text basis. Resonating with the overall low rates of correct AI/human author identification, none of the 72 reviewers properly identified all four abstracts, 18.1% properly identified three of the four, 34.7% properly identified two of the four, 34.7% properly identified only one of the four, and 12.5% did not identify any properly. On a text-to-text basis, the most difficult text was AI-2, which only 22.6% of all respondants correctly identified. As mentioned, the most consitently identified was Human-3, which 55.6% of all respondants correctly identified, with the rest following in a narrower band.

4.2. RQ2: Basis of reviewers' decisions for determining ChatGPT/AI vs. human writing

After participating in the identification task for RQ1, several reviewers ($n = 7$) engaged in follow-up interviews, in which they discussed the basis of their decisions. We identified a total of 10 different rationales that reviewers gave for determining whether a writing sample was produced by ChatGPT/AI or humans. Due to space limitations, we do not explicitly discuss all 10 rationales in this section. Instead, we highlight the top four recurring criteria that reviewers relied upon. However, all 10 rationales are shown in Table 3 with example quotes to illustrate these themes. Notably, none of the reviewers we interviewed relied exclusively on only one rationale. Instead, when discussing their decision-making processes, each reviewer relied on a range of criteria to sort the abstracts they read, which ranged from as few as three criteria (Reviewer 3) to as many as five criteria (Reviewers 1 and 2).

The top four rationales provided by reviewers pertained to a text's (1) *continuity and coherence,* (2) *specificity or vagueness of details,* (3) *familiarity and voice,* and (4) *writing quality at the sentence-level.* The first of these criteria, *continuity and coherence*, was the most frequently used, with five of seven reviewers mentioning it at least once. This rationale was used to describe the general ease or difficulty of reading an abstract. Reviewers attributed any writing flow or continuity issues to AI-produced writing, whereas they considered fluid, connected, and coherent compositions to humans. For instance, reviewers often made direct reference to textual coherence "I found that there was a certain level of coherence here" (Reviewer 5) in their evaluations. Human texts were identified as "easier to read" (Reviewer 2) due to connections in writing and ideas, while cohesion issues such as "the connections between the sentences are not always natural" (Reviewer 6) motivated AI classifications. However, this did not appear to be a reliable criterion, as reviewers who based their decisions on *continuity and coherence* correctly identified an author's identity only 22.2% of the time (2 of 9 instances).

Table 2
Reviewers' ($N = 72$) confidence in their ability to distinguish ChatGPT/AI from human writing.

Response item	Frequency	Percent (%)
1) Completely confident I cannot distinguish	1	1.4%
1) Fairly confident I cannot distinguish	9	12.5%
1) Somewhat confident I cannot distinguish	15	20.8%
1) Neutral	18	25.0%
1) Somewhat confident I can distinguish	23	31.9%
1) Fairly confident I can distinguish	6	8.3%
1) Completely confident I can distinguish	0	0.0%

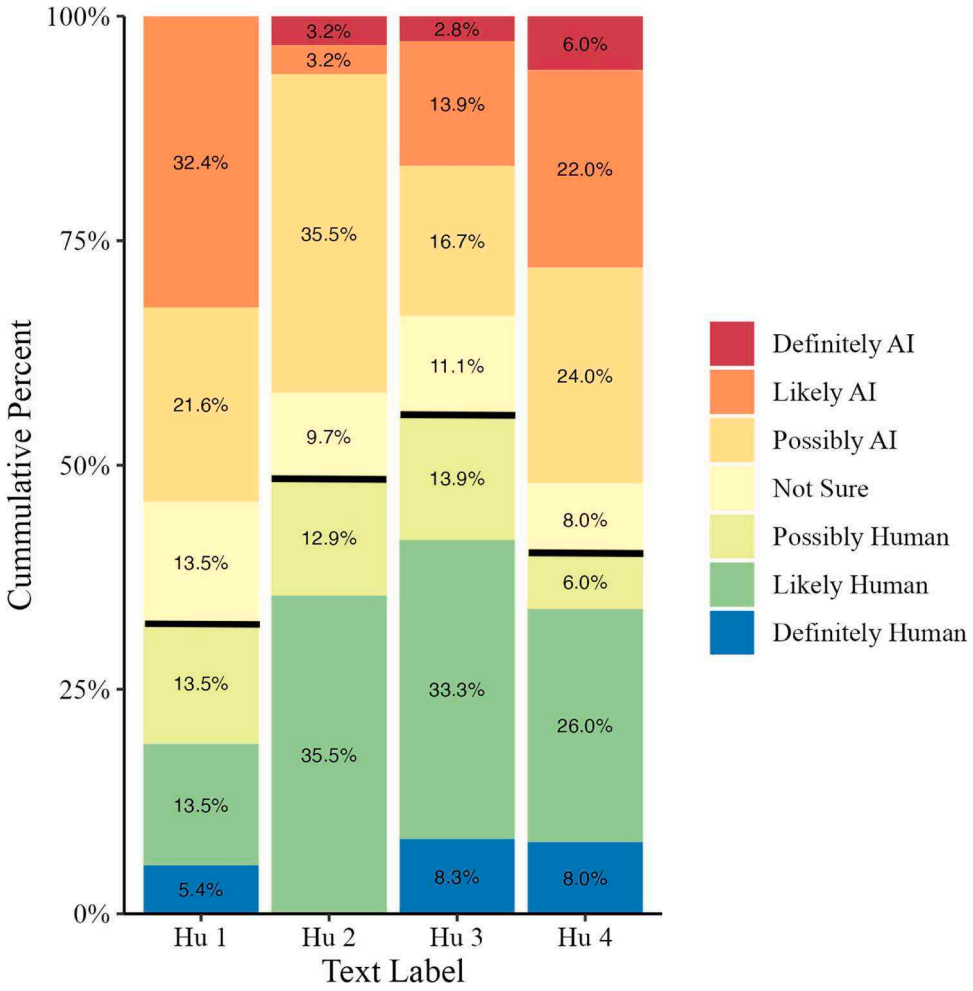

Fig. 1. Reviewers' author identifications and confidence levels on human-produced texts by percent.

The second frequent rationale reviewers provided was *specificity or vagueness of details,* with three of seven reviewers discussing it at least once. With this rationale, when reviewers found an abstract to contain considerable detail or specific information about different aspects of the study such as the methods or findings, they attributed the text to being human-generated. Conversely, if the writing sample was vague or thin on details regarding the study, reviewers assumed that the text was AI-produced. For example, after reading an abstract, Reviewer 7 commented: "It talks about this specific university [in the abstract]. I thought that's another specific fact AI might not have mentioned." *Specificity or vagueness of details* also did not seem to be reliable in making determinations. With this rationale, reviewers correctly identified an abstract's origins 28.6% of the time (2 of 7).

The third criterion reviewers adopted was *familiarity and voice.* Three of seven reviewers discussed it. In all instances, reviewers discussed this theme in terms of a writing sample feeling generally familiar or human-like, oftentimes referring to the tone or voice within the text. For instance, after reading an abstract and judging it as 'likely human' (which it was not), Reviewer 5 stated "It seems like this author's style is coming through… It felt more unique… I would call this what is commonly referred to as voice, which I would describe as some unpredictability or some uniqueness." Similar to the previous two rationales, *familiarity and voice* did not appear to be a reliable indicator, as reviewers judged abstracts correctly only 20% of the time (1 of 5).

The fourth and final rationale highlighted here is *writing quality at the sentence-level.* Four of seven reviewers discussed it at least once. This consisted of reviewers commenting that, when the composition of particular phrases or sentences appeared generally well constructed, they attributed this to humans. Conversely, if individual phrases or sentences were not well written and judged to be of poor quality, reviewers attributed it to AI. An example is shown in Table 3, in which Reviewer 4 commented positively on the "writing quality at the sentence level," thereby believing the sample to be human. However, in another instance, Reviewer 6 felt the writing quality was poor and thus attributed the abstract to AI, stating: "The last sentence in the first paragraph…That's something weird to say, like, the way they present the methodology," which they later connected to phrasing. When using this criterion, reviewers tended to be somewhat more successful, correctly identifying the origins of the abstract 60% of the time (3 of 5).

As discussed earlier, there were other criteria reviewers leveraged when attempting to identify ChatGPT/AI- versus human-

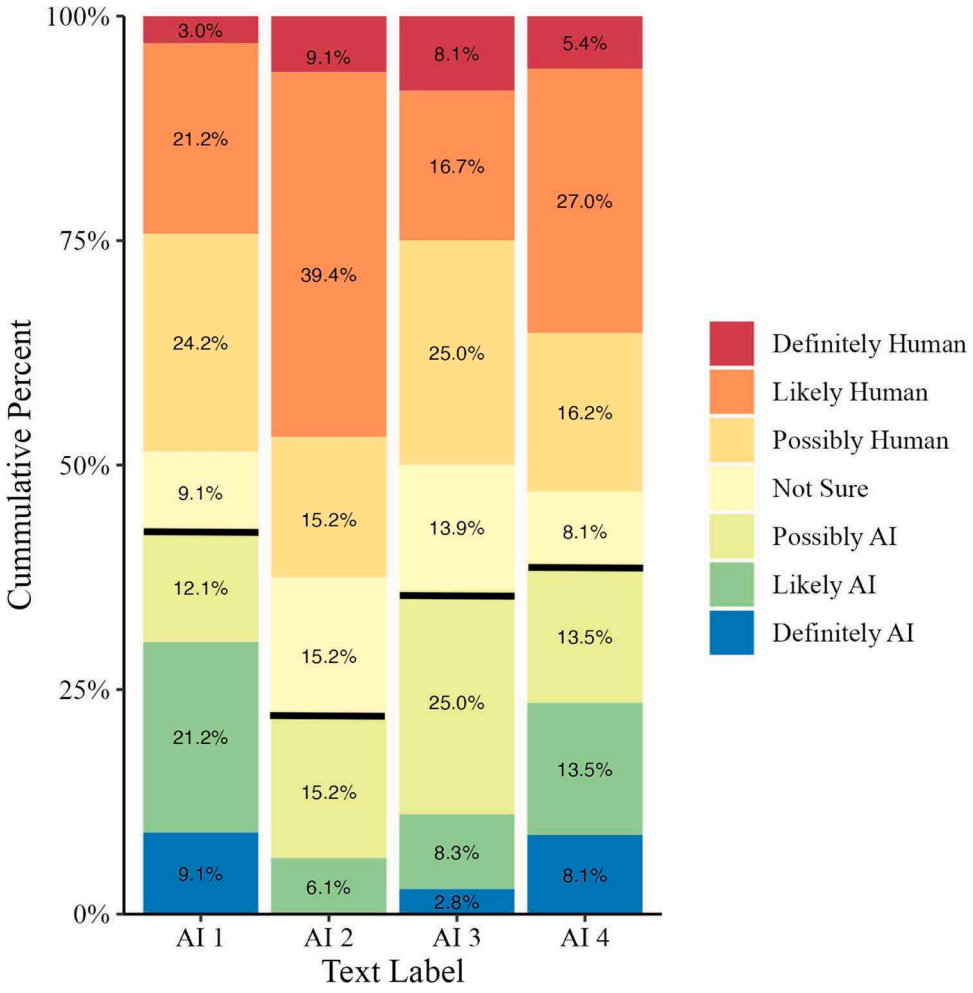

Fig. 2. Reviewers' author identifications and confidence levels on ai-produced texts by percent.

produced writing. However, these rationales tended to occur less frequently in the dataset. They also tended to be more idiosyncratic in nature, typically being discussed by only one reviewer (e.g., Reviewer 4's rationale of *methods [il]logical*) or by two reviewers (e.g., Reviewer 3 and 7's discussion of *abbreviations or contractions*). Due to the idiosyncratic nature of these items, we do not discuss them in-depth here and instead refer readers to Table 3. Interestingly, most of these points informed AI identification if negative and human identification if positive.

4.3. RQ3: Journal editors' beliefs about the ethicality of using ChatGPT/AI tools for publishing

The final RQ examined the extent to which editors of Applied Linguistics journals believe the use of ChatGPT and AI tools are ethical for academic research and publishing purposes. The first survey question asked editors about their general approval/disproval of researchers' use of such tools when submitting work to their journals. Editors' responses to this question are displayed in Table 4. As shown, a large majority believe that AI tools are acceptable under specific circumstances, while a smaller (yet not insignificant) sample stated that such tools were not acceptable under any circumstance. Notably, one editor believed that the use of AI tools was always okay for authors, regardless of the situation.

The next survey question asked editors to select and provide specific examples of instances in which ChatGPT/AI tools might be appropriate for researchers to use. Table 5 lists editors' responses. As shown, journal editors identified nine different ethical uses of such tools, not including one additional option, which enabled editors to select "under no circumstance is it acceptable" (25.9%). Of the 27 editors, over 50% felt that using ChatGPT/AI tools was acceptable for two purposes, which included *editing text* or *writing computer code* (e.g., R scripts) for one's study. Following these two uses, approximately 40% believed it was also acceptable to use ChatGPT/AI tools for *writing a summary for public consumption* or for *writing an abstract for a study*. Additional uses were also identified (e.g., *analyzing [some] data for the study*), yet fewer editors agreed on the acceptability of such uses.

Following the survey, five editors participated in follow-up interviews to elaborate on what they felt was (un)ethical, why they felt

Table 3
Rationales provided by reviewers for determining ChatGPT/AI vs. human-produced writing.

Code	Description	Example quote[1]	Total mentions by reviewers
Continuity and coherence	• the general ease or difficulty of reading the abstract (e.g., the flow, the connections between concepts, overuse of transitions)	• *There's a whole bunch of concepts going on that are not connected.* (R4)	10
Specificity or vagueness of details	• the information provided is very specific (e.g., in terms of the details of the methods, results); or, that the information provided is vague, too general, or non-specific	• *It was a little bit too general. If it were a real person who had actually conducted this study themselves, I felt that they would have given more specific information.* (R7)	8
Familiarity and voice	• the tone or voice of the abstract seems familiar, human-like, or personal in nature	• *This one reads like a human, and that's my bias. It's friendlier.* (R2)	5
Writing quality at sentence-level	• a particular phrase or sequence of words seems well produced or constructed	• *Sentence wise, the writing quality at the sentence level is good.* (R4)	5
Methods (il)logical	• the methods used in the abstract either (a) makes sense, or (b) doesn't make sense	• *There are too many concepts that are impossible to be investigated in one study.* (R4)	4
Templatic/ formulaic in nature	• the abstract feels overly formulaic, like a template, and without any creativity	• *The overall structure or frame that it's within is very general or templatic to me.* (R1)	4
Abbreviations or contractions	• the use of abbreviations (e.g., L2, TBLT) or contractions (e.g., can't, won't) in the writing sample	• *I don't see ChatGPT doing like… when we have [the phrase] "critical pedagogy" and then in parentheses "CP"… So for me, that's an indicator as soon as I open the text.* (R3)	3
Unexpected rhetorical move	• information is provided that is not typically given or seen in the abstract genre	• *In the second paragraph, he said "However, the study's limitations…" I rarely see that in an abstract.* (R6)	3
No idea	• s/he has no evidence to make a decision	• *I don't see any evidence either way.* (R3)	2
Specialist/technical language	• author uses specific word sequences or phrases that only a subject-matter expert would use	• *The phrase "online teacher inquiry model" seems really specific, like only somebody who does online teacher inquiry models would even know to put those words together.* (R1)	1

[1] In this 'Example quotes' column, 'R' stands for 'reviewer.'

Table 4
Editors' ($N = 27$) general approval of authors' use of ChatGPT/AI tools for publishing.

Response item	Frequency	Percent (%)
Under specific circumstances	19	70.4%
Under no circumstance (i.e., it is never okay)	7	25.9%
Under all circumstances (i.e., it is always okay)	1	3.7%

Table 5
Editors' ($N = 27$) responses regarding ethical uses of ChatGPT/AI tools for publishing.

Response item	Frequency	Percent (%)
Editing text	16	59.3%
Writing computer code for the study	14	51.9%
Writing a summary for public consumption/open access purposes (e.g., OASIS)	11	40.7%
Writing the abstract for a study	10	37%
Under no circumstance is it acceptable	6	22.2%
Writing parts of the main document	4	14.8%
Analyzing (some) data for the study	3	11.1%
Writing the entire main document	1	3.7%
Writing the conclusion section	1	3.7%
Data visualization	1	3.7%

that way, and also, what some of their uncertainties and fears were with ChatGPT/AI tools. One of the first comments made by multiple editors pertained to the realization that such tools were unlikely to be stopped now or in the future. As Editor 2 simply stated: "With AI…it's inevitable." Echoing this sentiment, Editor 1 stated: "I don't think we can stop technological advancement. We never can… So instead of avoiding it, using it for benefits is a better direction." Thus, among the editors that participated in the interviews, there was a general sense of agreement that it was better to proactively figure out the positives and negatives of AI tools, rather than trying to avoid their use entirely.

As one example of this, as mentioned, over 50% of the editors who were surveyed agreed that AI tools were acceptable for either *editing text* (discussed in the next section) or for *writing computer code for the study* (e.g., R scripts). In terms of the latter, some of the editors we interviewed commented on this. Editor 5, for instance, noted that he views AI tools as very similar to researchers' current

use of other statistical tools and software (e.g., SPSS), stating: "In the case of R or writing script, we all use tools, especially for statistical analysis." Thus, some of the editors believed that AI tools have the capacity to help researchers process and analyze data they have already collected in ways that resonate with current tool usage in the field.

Apart from such comments, there were three primary themes that were discussed by the editors we interviewed. In what follows, we discuss these three themes, which included: (1) *using AI as an editing tool*, (2) *using AI for summarizing completed work*, and also some of the editors' fears about (3) *perpetuating biases and data fabrication.*

4.3.1. Using AI as an editing tool

One emergent theme from the interviews pertained to editors discussing AI as a new tool that could generally assist humans in editing their work. Three editors discussed this theme. For instance, Editors 3 and 4 commented on how, to some extent, humans have already been using AI tools for academic writing purposes:

> [With] editing… You know, people use Grammarly. It's not any different. You don't say 'I have used Grammarly' [when you go to publish], right? (Editor 3)

> I just see it as a more advanced version of Microsoft Word's autocorrect or autocomplete. We've actually been using AI tools for a number of years likely without realizing that we're using them. (Editor 4)

As these editors noted, when writing a manuscript, authors typically rely on AI tools within Microsoft Word, Grammarly, and those embedded in similar software to help them write faster and more accurately. Of course, we acknowledge (and many of the editors did, too), that ChatGPT is different from such tools in important ways. In another interview, Editor 2 agreed, extending the point by discussing how people's opinions of AI might evolve in the future:

> Our understanding of ethics and our standards may change, as…different technologies, change. Like, for example, with calculators. It may be that when calculators were first invented, that people thought, 'No. That's cheating. You need to do your own hand calculations.' But I think now we feel like, 'No. It's a tool to help us make our work faster and more accurate.' (Editor 2)

Thus, multiple editors framed ChatGPT use through previous tool use, emphasizing the general usefulness in assisting researchers with editing or revising their writing.

4.3.2. Using AI for summarizing completed work

The second theme that emerged from the editors was how researchers might ethically leverage AI tools for research purposes. Four editors mentioned that they believed using ChatGPT/AI tools to do different types of summary writing were acceptable. This included using AI tools to write abstracts and to write summaries intended for public consumption/open access (e.g., OASIS). As some editors explained:

> Abstracts are basically stylized passages that fit very well-defined structures, and these tools are absolutely fantastic at knowing those structures and able to summarize that based on the text that that people tend to write. Because it's just summarizing your own text, I don't see any issue… Similarly, executive summaries for like OASIS or LinkedIn posts, Twitter, things like that. It just saves so much time to feed what you've done into the software and have it spit that out… I mean, you've already produced the study. The abstracts is one of the last things. (Editor 4)

> Those are the areas of the writing process…[where] the tools are doing something with something you've already done as an author, which feels to me more okay than actually producing something from scratch. (Editor 5)

In these passages, Editors 4 and 5 both point to summarizing as an acceptable use of AI tools. This is primarily because they believe it saves time, but more importantly, because authors often write abstracts or open-access summaries once the research has already been completed. Thus, rather than being used to produce new knowledge, editors believed AI tools were ethical for aiding researchers in the final stages of writing.

4.3.3. Fears: perpetuating biases and data fabrication

Although many editors agreed that ChatGPT/AI tools could be used in ethical ways that involved editing or summary writing, they also voiced fears about possible misuses. For example, Editor 5 feared how large publishing houses might try to use AI tools to replace human labor, along with the potential of AI tools to perpetuate human biases:

> From the reviewer and editorial standpoint, I worry about…reviewers and journals making decisions based on AI models that are built on unjust human practices… If models are based on biased human practices, they will be consistent, but they will be consistently biased. (Editor 5)

Specifically, Editor 5 feared that publishing houses might try to use AI tools to make recommendations to editors about which manuscripts to desk reject or to send out for review based on the actions of previous (or current) editors. However, as Editor 5 noted, such AI tools might use data that is based on a select few individuals' preferences, which may be biased. Editor 5 feared that such uses of AI might perpetuate inequities in publishing by unfairly discriminating against specific research topics, methods, or authors from certain regions, while also not prioritizing innovation.

Apart from perpetuating biases, Editors 1 and 4 were particularly worried about researchers who might misuse AI for data fabrication. As Editor 4 explained:

> What I'm terrified of from an editorial point-of-view is someone who has gone to ChatGPT and said, 'Give me 20 people's survey data randomized in an Excel file.' And then they've come along and said, 'This is my survey data.' So where people are now using ChatGPT for data fabrication as well as using AI image generation to make things like charts that aren't based on real data. I think data fabrication is going to be [a] massive [issue] for editors. (Editor 4)

Clearly, although many of the editors we interviewed agreed that AI tools could be ethically used, they also feared that such tools would be used in unethical ways either unintentionally (e.g., by publishing houses) or intentionally (e.g., by authors fabricating data). In terms of the general use of AI in academic publishing, Editor 4 closed by saying:

> It's gonna help good researchers, but who are bad writers do very well. But then there's bad researchers out there who now have the tools to help them get noticed and to do things. (Editor 4)

5. Discussion and conclusions

The findings of the current study suggest that experienced reviewers' and linguists' have limited capacity to distinguish ChatGPT/AI from human-produced abstracts (RQ1). In fact, none of our 72 participants were able to correctly identify all four abstracts they viewed. Of the reviewers, only 18.1% correctly identified 3/4 abstracts, 34.7% identified 2/4, 34.7% identified 1/4, and 12.5% misidentified all four. This resulted in an overall positive identification rate of only 38.9%. Notably, the reviewers in our study were better at identifying human writers than AI writers. This was reflected in the fact that reviewers correctly identified human texts 44.1% of the time versus ChatGPT/AI texts 33.7% of the time.

Regarding reviewers' decision-making processes and rationales for distinguishing between human- and AI-generated texts (RQ2), multiple and diverse rationales and principles emerged for navigating the task (between three and five justifications, depending on the reviewer). Some of these rationales were employed by multiple reviewers, while some justifications were more idiosyncratic in nature. Although there is limited scholarship in this area to date, we do note that one of the rationales adopted by three of seven reviewers in our study pertained to a *specificity or vagueness of details,* which resonates with Ma et al.'s (2023) findings. Similar to the two participants in Ma et al., our participants attributed considerable detail/information to human writers, while attributing vagueness and a lack of details to AI. Despite this similarity, we note that our participants were generally unsuccessful when adopting this criterion (correct only 28.6% of the time). Importantly, the use of positive criteria for describing human writing and negative criteria for describing ChatGPT/AI writing underscores that reviewers viewed human-produced writing as superior.

Humans' general inability to correctly identify human or AI writing may have negative consequences, including the potential for humans (e.g., teachers, professors) to accidentally accuse people of 'plagiarism,' when in fact, the writer was actually human – (consider the numerous recent stories appearing in the news). This particularly may be the case with shorter, more formulaic genres such as research abstracts, which as noted earlier, have been found to be both rhetorically predictable (e.g., Halleck & Connor, 2006; Samar et al., 2014; Yoon & Casal, 2020) and characterized by highly formulaic language (e.g., Casal & Yoon, 2023; Omidian et al., 2018). This is further complicated by reviewers' greater ability to identify human authors (44.1%) than AI authors (33.7%).

Yet, regarding the ethicality of ChatGPT/AI's use for research and academic publishing purposes (RQ3), many journal editors feel that abstract writing constitutes an acceptable usage of AI, along with other tasks such as summary writing, writing computer code, and editing text, among others. Thus, in terms of implications, it seems that there may be many existing applications for ChatGPT and similar AI tools for research activity and publishing. Such tools may be particularly useful for (a) serving as a facilitative tool to help researchers analyze or process their data, and also (b) aiding in the final writing/editing stages of the research process. Thus, we encourage researchers to experiment with such uses, both informally (on their own) and formally (in a research capacity), but we also highlight the importance of disclosing AI use as is customary for any methodological procedure or tool. Of course, additional research is also needed to further investigate those uses described in the current study, since our study does not report on ChatGPT/AI's capacity to accurately perform these tasks.

However, it is also important to note here that numerous journal editors disagreed (25.9%), stating that AI's use is *never* acceptable for academic research and publishing in any way. Unfortunately, none of these editors volunteered to participate in follow-up interviews, so we were unable to learn more about their particular opinions. That being said, the fact that so many editors believe that AI's use should never be permitted leads to many potential issues and questions. For instance, this raises questions as to how – as a community and discipline – we might move forward when a sizable portion of influential gatekeepers are strongly against AI's use. That is, when it comes to research ethics and understanding the nature of QRPs (i.e., *Questionable Research Practices,* see Isbell et al. (2022), and Larsson et al. (2023), in the future, will academic journals vary in what they consider to be (un)acceptable or instances of misconduct with AI? The findings of this study suggest that such discussions need to take place, particularly among the editor community and academic publishing houses.

Given the novelty of this research space and these questions within Applied Linguistics, we close this study by turning to a brief discussion of future research directions. In particular, more studies are needed that investigate humans' capacity to differentiate between human- and AI-produced writing, particularly regarding how humans reach or justify their decisions. Since we leveraged a relatively short, formulaic genre for the current study, we encourage future studies to adopt similar methods and to explore other genres beyond abstracts, such as research article part-genres (e.g., introductions, literature reviews), which are typically longer and thus may be more prone to hallucination. Importantly, such studies have the capacity to help our field learn more about the emergent systematicity of human judgements, the (in)consistent application of emergent principles, and the accuracy and of such judgements. Relatedly, as mentioned, studies are also needed that investigate those summary-oriented genres that were identified by the editors in

our study as being acceptable for AI-generation (e.g., creating OASIS, executive, or LinkedIn summaries). That is, to what extent can ChatGPT/AI tools accurately perform such tasks, and to what extent do AI-produced texts mirror the typical rhetorical moves and lexicogrammatical features employed by human writers?

Finally, as is evidenced by the current study (and by broader, informal discussions across academic and professional contexts), more research is needed that investigates the ethicality of AI's use for research and publishing purposes. Our findings point to many similarities among editors' openness to AI as a facilitative tool. However, our findings also point to potential fears and conflicts, all of which must be explored in greater depth.

Declaration of Competing Interest

None.

References

Bender, E., Gebru, T., McMillan-Major, A., & Shmitchel, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623).

Casal, J. E., & Yoon, J (2023). Frame-based formulaic features in L2 writing pedagogy: Variants, functions, and student perceptions in academic writing. *English for Specific Purposes, 71*, 102–114. https://doi.org/10.1016/j.esp.2023.03.004

De Costa, P., Lee, J., Rawal, H., & Li, W. (2021). Ethics in applied linguistics. In J. McKinley, & H. Rose (Eds.), *The Routledge handbook of research methods in applied linguistics*. Routledge.

De Costa, P. I., Sterling, S., Lee, J., Li, W., & Rawal, H. (2021). Research tasks on ethics in applied linguistics. *Language Teaching, 54*(1), 58–70. https://doi.org/10.1017/S0261444820000257

Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching, 54*(2), 245–258. https://doi.org/10.1017/S0261444819000430

Hagendorff, T. (2020). The ethics of AI: An evaluation of guidelines. *Minds and Machines, 30*, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Halleck, G. B., & Connor, U. M. (2006). Rhetorical moves in TESOL conference proposals. *Journal of English for Academic Purposes, 5*, 70–86. https://doi.org/10.1016/j.jeap.2005.08.001

Isbell, D., Brown, D., Chan, M., Derrick, D., Ghanem, R., Gutiérrez Arvizu, M. N., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *Modern Language Journal, 106*(1), 172–195. https://doi.org/10.1111/modl.12760

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kasneci, E., Sesler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kessler, M., Marino, F., & Liska, D. (In press). Netnographic research ethics in applied linguistics: A systematic review of data collection and reporting practices. *Research Methods in Applied Linguistics*.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, 1–13. https://doi.org/10.1177/00336882231162868

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science, 9*(1), 104–117. https://doi.org/10.1017/XPS.2020.37

Kubonyiova, K. (2008). Rethinking ethics in contemporary applied linguistics: The tension between macroethical and microethical perspectives in situated research. *The Modern Language Journal, 4*, 503–516. https://doi.org/10.1111/j.1540-4781.2008.00784.x

Kurian, N., Cherian, J., Sudharson, N., Varghese, K., & Wadhwa, S. (2023). AI is now everywhere. *British Dental Journal, 234*(72). https://doi.org/10.1038/s41415-023-5461-1

Larsson, T., Plonsky, L., Sterling, S., Kytö, M., Yaw, K., & Wood, M. (2023). On the frequency, prevalence, and perceived severity of questionable research practices. *Research Methods in Applied Linguistics, 2*(3), Article 100064. https://doi.org/10.1016/j.rmal.2023.100064

Ma, Y., Liu, J., & Yi, F. (2023). AI vs. human - Differentiation analysis of scientific content generation. *arXiv Computation and Language*. https://doi.org/10.48550/arXiv.2301.10416

Marino, F., Liska, D., & Kessler, M. (In press). Ethical considerations for research involving computer-assisted language learning, social media, and online environments. In P. I. De Costa, A. Rabie-Ahmed, & C. Cinaglia (Eds.), *Ethical issues in applied linguistics scholarship*. John Benjamins.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2018). *Qualitative data analysis: A methods sourcebook* (4th ed.). Sage.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*, 501–507. https://doi.org/10.1038/s42256-019-0114-4

Omidian, T., Shahriari, H., & Siyanova-Chanturia, A. (2018). A cross-disciplinary investigation of multi-word expressions in the moves of research article abstracts. *Journal of English for Academic Purposes, 36*, 1–14. https://doi.org/10.1016/j.jeap.2018.08.00

OpenAI. (2023). *ChatGPT4 Plus [Computer software]*. Retrieved from https://www.openai.com.

Plonsky, L. (2014). Study quality in quantitative L2 research (1990-2010): A methodological synthesis and call for reform. *The Modern Language Journal, 98*, 450–470. https://doi.org/10.1111/j.1540-4781.2014.12058.x

Polio, C., & Friedman, D. (2017). *Understanding, Evaluating, and Conducting Second Language Writing Research*. Routledge. https://doi.org/10.4324/9781315747293.

Samar, R. G., Talebzadeh, H., Kiany, & G. R., & Akbari, R. (2014). Moves and steps to sell a paper: A cross- cultural genre analysis of applied linguistics conference abstracts. *Text & Talk, 34*(6), 759–785. https://doi.org/10.1515/text-2014-0023

Shen, Y., Heacock, L., Elias, J., Hentel, K., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and other large language models are double-edged swords. *Radiology, 307*(2). https://doi.org/10.1148/radiol.230163

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management, 31*(2), 1–14. https://doi.org/10.4018/JDM.2020040105

Spilioti, T., & Tagg, C. (2017). Ethics of online research methods in applied linguistics. *Applied Linguistics Review, 8*(2–3). https://doi.org/10.1515/applirev-2016-1033

Sterling, S., & Gass, S. (2017). Exploring the boundaries of research ethics: Perceptions of ethics and ethical behaviors in applied linguistics research. *System, 70*, 50–62. https://doi.org/10.1016/j.system.2017.08.010

Yang, D., Zhou, Y., Zhang, Z., Li, T. J. J., & LC, R. (2022). AI as an active writer: Interaction strategies with generated text in human-AI collaborative fiction writing. In *, 10. Joint proceedings of the ACM IUI workshops*.

Yaw, K., Plonsky, L., Larsson, T., Sterling, S., & Kytö, M. (2023). Research ethics in applied linguistics. *Language Teaching*, 1–17. https://doi.org/10.1177/00336882231162868

Yoon, J., & Casal, J. E. (2020). Rhetorical structure, sequence, and variation: A step- driven move analysis of applied linguistics conference abstracts. *International Journal of Applied Linguistics*, 1–17. https://doi.org/10.1111/ijal.12300