A Comparative Analysis of Chatgtp-4/4.5 and Human-written Summaries in Linguistic Research

Aleksa Stošić[1[0009-0005-4601-0401]], Marija Milojković[1[0000-0003-4787-1017]], and Aleksandra Branković[1[0009-0007-7810-5581]]

[1] Belgrade Metropolitan University, Tadeuša Košćuška 63, Belgrade 11158, Serbia
aleksa.stosic@metropolitan.ac.rs
marija.milojkovic@metropolitan.ac.rs
aleksandra.brankovic@metropolitan.ac.rs

Abstract. This study evaluates the potential of ChatGPT-4 and ChatGPT-4.5 as research assistants in applied linguistics (AL) by examining their ability to generate annotated bibliographies of research articles. Five AL papers on technology in English pronunciation and speaking instruction were summarized by both models and by human researchers, producing 25 summaries. Fourteen expert raters assessed the summaries for quality and judged their authorship. Results show that both models produced factually accurate and structurally faithful summaries. However, both models lacked critical selectiveness, could only provide generalized statements on relevance, and relied on surface-level markers to assess credibility. Quantitative analysis indicated that ChatGPT summaries were rated as comparable in quality to human-authored ones, though inter-rater agreement was low and a bias against texts perceived as AI-generated was observed. Qualitative findings revealed that experts distinguished AI from human summaries based on information density, word choice, stylistic naturalness, and evaluative engagement. Overall, ChatGPT proved advantageous in accuracy, structural consistency, and efficiency, but its weaknesses in evaluative depth and authenticity suggest that, while it can accelerate the early stages of literature review, it cannot substitute for the nuanced judgment and interpretive reasoning required in applied linguistics.

Keywords: ChatGPT, applied linguistics, research assistant, research summarization, annotated bibliography.

1 Introduction

The rapid advancement of generative artificial intelligence (AI), especially through large language models (LLMs) such as ChatGPT, has had a wide impact on the academic community and the process of conducting research. Since its public release in November 2022, ChatGPT has been employed for tasks such as generating abstracts and summarizing research articles [1], while some authors have even used it as a co-author [2]. ChatGPT has undergone several iterations, and its current versions at the time of writing, ChatGPT-4 and its successor ChatGPT-4.5, have much enhanced

Research Paper
DOI: https://doi.org/10.46793/eLearning2025.205S
Part of ISBN: 978-86-89755-37-4

capabilities compared to previous iterations such as ChatGPT-3 [3][4]. These developments suggest an expanding potential for using ChatGPT as a research assistant.

So far, ChatGPT has been used as a research assistant in diverse fields, including medicine for generating research article abstracts [5], international business and management for summarizing from citations, retrieving citations from those summaries, and linking them back to the original abstracts [6]. In the field of linguistics, Bae's [7] demonstration study illustrated how ChatGPT-3.5 and ChatGPT-4 could assist experimental linguistics by suggesting sources for literature reviews, supporting experimental design, and facilitating statistical analyses. The author concluded that the tools could save researchers time and streamline data preparation, but constant verification was necessary as the tools tended to exhibit shallow reasoning and produce hallucinations[1] and outdated information. Furthermore, Uchida [8] assessed ChatGPT-3.5 for core corpus-linguistic tasks such as frequency analysis, collocation identification, and genre classification, finding partial alignment with established corpus data and highlighting significant limitations, particularly in genre analysis, suggesting that the tool should be used as an auxiliary rather than primary tool for rigorous research. However, there is a notable gap in the literature as to how ChatGPT can be used as an assistant for summarizing research articles. Its use could greatly assist researchers during the time-consuming stage of literature preview, which is foundational for subsequent stages of the research process. Reluctance to use ChatGPT for such purposes likely stems from its tendency to generate inaccurate information and hallucinate, which puts its ability to reliably summarize information into question [1] [9]. Ethical concerns have also been raised, with several studies demonstrating that ChatGPT can produce plagiarized information when used as a research assistant [10] [11] [12]. Nonetheless, if the accuracy of ChatGPT's latest models continues to improve, ChatGPT could substantially benefit researchers by reducing the time spent on reviewing literature, and allow for more time to be invested in conducting practical linguistic research. While field-specific applications vary, this study focuses on applied linguistics (AL), with the expectation that insights into ChatGPT's summarization capacity may yield benefits across disciplines.

Therefore, this study seeks to evaluate the performance of ChatGPT-4 and ChatGPT-4.5 in summarizing AL research papers by addressing the following research questions:

1. To what extent do ChatGPT-4 and ChatGPT-4.5 produce accurate summaries that capture key findings while adhering to the required structural conventions?
2. In what respects do ChatGPT-generated summaries differ from human-authored summaries?

By evaluating the performance of these models, this study aims to provide insight into the advantages and limitations of using ChatGPT for summarizing linguistic research papers, while also offering potential insights for the broader development of artificial intelligence. To the best of our knowledge, as of August 2025, no peer-reviewed

[1] AI generated content that is fluent and confident but factually incorrect, fabricated, or unsupported by its training data.

study has systematically examined ChatGPT's ability to summarize AL research articles. This study addresses that gap by analyzing summaries of AL papers generated by ChatGPT-4 and ChatGPT-4.5 and comparing them with those produced by human authors, drawing on expert evaluations to determine relative quality and reliability.

2 Previous Research

2.1 Technology Acceptance Models and the Use of ChatGPT as a Research Assistant

Previous research on the acceptance of ChatGPT has examined the factors that influence its adoption as a research assistant, but concerns about its accuracy and reliability for summarization tasks have also been raised. These studies primarily rely on the factors outlined in the Technology Acceptance Model (TAM) [13] and the Unified Theory of Acceptance and Use of Technology (UTAUT) [14]. These theories are closely connected as TAM focuses on perceived usefulness and ease of use, while UTAUT refines these factors and expands them with performance expectancy, effort expectancy, social influence, and facilitating conditions. Empirical studies have shown that there is a high acceptance rate of LLMs among academics, and that its adoption is primarily shaped by usefulness, ease of use, and researcher competence, alongside factors such as AI's perceived intelligence, enjoyment, social influence, and institutional practices [15]. Balaskas et al. [16] reported similar findings, adding that age and prior AI experience also play a moderate role. Nevertheless, concerns about accuracy persist. Salleh [6] documented frequent errors of omission (excluding relevant information), and commission (including irrelevant information) when ChatGPT is used as a research assistant. They noted that while ChatGPT can generate relevant summaries from citations, it often fails to understand the papers' theoretical significance. Moreover, it often mistakenly attributed the work to other authors, and often exhibited the Matthew Effect by showing a bias towards more prominent authors. These findings suggest that ChatGPT can help in summarization, but remains unreliable for citation and theoretical integration. Similarly, Rahman et al. [17] found that, even though ChatGPT-3.5 could generate acceptable abstracts if given precise prompts, it struggled to evaluate the studies critically and connect their significance when used for literature review. Therefore, in addition to examining whether ChatGPT-4 and ChatGPT-4.5 can accurately summarize AL research papers, this study also assesses their ability to critically evaluate research papers as an essential component of the literature review process.

2.2 Evaluation Criteria for ChatGPT-generated Summaries

When evaluating the effectiveness of ChatGPT in generating summaries of research papers to support AL research, the accuracy and faithfulness of the information in the generated output constitute a sine qua non condition and therefore represent the primary criterion in the present study.

Previous research has frequently assessed ChatGPT-generated summaries using ROUGE (Recall-Oriented Understudy for Gisting Evaluation), a set of automatic evaluation metrics used to measure lexical overlap between generated and reference summaries to determine how much of the essential content is captured [18]. Goyal et al. [19] reported that ChatGPT-3 exhibited lower ROUGE scores than traditional summarization methods when summarizing news articles and other similar content. In contrast, Yang et al. [20] found ChatGPT-3's performance on summarizing news, dialog transcripts, and Reddit posts to be on pair with traditional methods. Zhang et al. [21] investigated the performance of the more advanced model ChatGPT-3.5 in extractive summarization of diverse text types, including news articles, scientific articles, and government reports. Their findings also showed that ChatGPT-3.5 performed worse in terms of ROUGE scores when compared to traditional methods, but yielded better results in terms of faithfulness when provided with step-by-step summarization instructions. Such procedural prompting reduced the occurrence of hallucinations and errors, which earlier studies had identified as recurrent problems in ChatGPT-3 and ChatGPT-3.5 [19] [21] [22].

Although ROUGE remains widely used, Fabbri et al. [23] demonstrated that it was insufficient for evaluating academic or scientific summarization, as it failed to capture deeper qualities such as factual accuracy and coherence, and called for more refined metrics and the inclusion of human judgment. Responding to this limitation, Hake et al. [5] evaluated ChatGPT-3.5's summarization ability in medical research by asking experts to assess summaries or research articles generated from abstracts. Their findings confirmed that the tool could be of assistance since study participants claimed that summaries were high in quality, had high accuracy, and low bias, and they were often able to classify if the articles were relevant for different medical specialties. Nonetheless, the authors stressed that full-text evaluation remained necessary before making final research selection. To date, no comparable human-judgment studies have been conducted in applied linguistics, and the present study seeks to address this gap by drawing on the expertise of professional linguists to evaluate the effectiveness of ChatGPT-4 and ChatGPT-4.5 for AL research summarization.

Apart from accuracy and faithfulness, the quality of a summary also depends on its adherence to the conventions of academic writing. As a subgenre of academic discourse, summaries are expected to follow a recognizable structure and exhibit characteristic linguistic features [24]. These features include conciseness, high information density, explicit cohesion achieved through signaling devices, and the use of evaluative language to highlight significance or limitations. In the context of applied linguistics research, this further requires the consistent use of appropriate disciplinary terminology, since precision in technical language is essential for accurately conveying research findings and theoretical constructs. Accordingly, this study investigates whether ChatGPT-generated summarizes demonstrate these linguistic qualities.

As the structure of a summary may vary depending on particular research goals, the quality of ChatGPT-generated summaries should be judged by whether the tool can adhere to the required format. In this study, evaluation is based on a modified annotated bibliography format, as outlined by the Purdue Online Writing Lab of the College of Liberal Arts of Purdue University, though other formats could also be used when prompting ChatGPT to summarize research papers. The term annotated bibliography refers to a list of research sources where each entry includes the full citation and a short paragraph that describes the source's content and evaluates its relevance and quality, while also providing a statement why the source is relevant for the researcher's own work. It is thus a more robust format than a summary, which only condenses the main ideas of a single article without providing an evaluation. Therefore, a further criterion for assessing the effectiveness of ChatGPT-4 and ChatGPT-4.5 in AL research summarization is their ability to follow the prescribed structure specified in the prompt. The annotated bibliography format was selected because it not only requires presentation of the topic, methodology, and key findings of the research article, but also demands a critical evaluation of the article's credibility and its relevance to the researcher's topic and goals. As Rahman et al. [17] and Salleh [6] found, ChatGPT-3.5 lacks the ability to critically evaluate information for literature review, so employing the annotated bibliography format provides a means of testing whether the current models, ChatGPT-4 and ChatGPT-4.5, demonstrate improved capacity for critical evaluation alongside summarization.

3 Methodology

Given the insights from previous research and the evaluation criteria identified, the present study adopts a mixed-method approach to assess the summarization performance of ChatGPT-4 and ChatGPT-4.5 in applied linguistics and determine whether the two models have improved capabilities and become more suitable to be used as a research assistant.

3.1 Topic Selection and Corpus Selection

A research topic in the field of applied linguistics, the integration of technology in English language pronunciation and speaking instruction, was chosen. This topic was selected for practical reasons, as the researchers involved in this study possess professional expertise in this area and are therefore well positioned to evaluate the quality of the generated summaries. It is assumed that findings concerning the factual accuracy, faithfulness, structural conformity, and academic language features of ChatGPT-generated summaries in this domain are likely generalizable to other areas of applied linguistics.

Five research papers published between 2022 and 2024 were selected to form the study corpus [25] [26] [27] [28] [29]. These papers were selected on the basis that the subject matter of these papers closely aligns with our own interests in AL research, and because the survey participants would also be familiar with these topics. They were drawn from peer-reviewed journals to ensure scholarly credibility and rigor. All were experimental studies, as this methodological uniformity helped minimize variability in the source material and enabled a fairer comparison between ChatGPT-generated and human-authored summaries. This corpus size was deemed sufficient to obtain meaningful findings, while acknowledging that larger corpora might yield additional insights beyond the scope of the present study.

3.2 Human-written Summaries

The researchers in this study (three junior lecturers in Applied English and Linguistics at Belgrade Metropolitan University who have produced peer-reviewed publications in the field of applied linguistics) independently produced summaries of the five articles using the modified Purdue OWL annotated-bibliography template. The annotation procedure comprised the following steps:

- State what the topic of the paper is and what it is about.
- Summarize the methodology and most important results of the paper (if applicable).
- Evaluate the credibility of the cited works.
- State how the findings or methodology of the paper are applicable to our own research paper.

3.3 ChatGPT-generated Summaries

Each of the five articles was uploaded separately to ChatGPT-4 and ChatGPT-4.5. In line with Zhang et al. [21], both models received step-by-step instructions specifying the annotated-bibliography format and required elements to minimize hallucinations and improve faithfulness. The same prompt was used for both models to ensure comparability. The exact prompt text is provided below.

- You are working on a research paper that discusses the integration of technology and AI into the ESL classroom. You want to summarize the following research paper and write the summary in the style of an annotated bibliography. You need to write the summary using the following template. Do not omit any information.
- Provide a citation for this paper following the APA7 style of citation.
- State what the paper is about and what topic it discusses.
- State what the methodology of the paper is and how the research was conducted.
- State what the most important results of the paper were.
- Provide an evaluation of the paper's credibility and explain why it is or is not a reliable source.
- State how the paper is relevant for your own paper and research.

Additional contextual information about the research topic was intentionally withheld in order to test how the two models would generate critical links to previous literature and to determine whether their performance showed any improvements.

In total, 25 summaries were produced: 15 written by the researchers (five each) and 10 generated by ChatGPT (five by ChatGPT-4 and five by ChatGPT-4.5).

3.4 Summary Evaluation Survey Questions and Participants

Fourteen university professors teaching English linguistics at Belgrade Metropolitan University and the Faculty of Philosophy of the University of Niš participated in this study as expert raters. All participants teach applied linguistics courses, and were selected through convenience sampling based on their relevant teaching and research expertise. Each expert rated the overall quality of all 25 summaries on a five-point Likert scale (very poor, poor, acceptable, good, very good) and answered whether they thought each summary was AI-generated. To avoid priming effects, no explicit criteria for judging the quality of the summaries were provided, as raters were expected to rely on their professional judgment. An optional open-ended question followed each summary, allowing participants to elaborate on their ratings and their decisions regarding AI authorship.

4 Data Analysis and Results

4.1 Summarization Quality Analysis of ChatGPT-generated Summaries

The first stage of analysis focused on evaluating the summaries produced by ChatGPT-4 and ChatGPT-4.5 in terms of accuracy, faithfulness to the original texts, and adherence to the required structural format. This analysis was carried out manually by all three researchers, who systematically examined each summary against the source article and the prescribed annotated bibliography template.

ChatGPT-4

Upon first examination of the five summaries, it was noticed that the summary of Article 5 contained several obvious inaccuracies and hallucinations. The information in the summary appeared to have been merged with the content of a different study. In order to eliminate any possible mistakes made by the researchers when prompting ChatGPT, this output was discarded and regenerated in a new chat with temporary chat mode enabled. The prompt was not changed nor were the model settings altered in any way. The regenerated summary of Article 5 no longer contained hallucinations, suggesting the error resulted from context window confusion between the documents rather than an inherent model failure.

Closer inspection of the final set of summaries revealed that they did not contain any information that was not included in the articles themselves. In each summary, ChatGPT had successfully identified the topic of the paper, described the methodology accurately, and presented the results in a way that did not omit any important information. However, there was one discrepancy in how it identified the topic in its summary of Article 4.

Summary 4: *This paper explores how technology-enhanced learning (TEL) contributes to the improvement of English pronunciation and overall language proficiency, emphasizing the role of digital tools such as speech recognition software, mobile applications, and interactive learning platforms. It investigates how blending tradi-*

tional instruction with modern technology can help learners overcome pronunciation challenges, receive real-time feedback, and enhance self-directed language practice.

The research focus of the original article was on the use of digital tools to enhance pronunciation, fluency, and comprehension skills, while the results from interviews indicated that the advantages of these systems lay in their ability to provide instant feedback and provide guidance for further improvement in real time. The summary, on the other hand, presented these elements as research aims rather than reported outcomes. We regard this imprecision as a minor error that is unlikely to influence a researcher's decision to consult the full paper. Overall, the factual accuracy and faithfulness of the analyzed ChatGPT-4 summaries were found to be high, and the outputs can be considered reliable in assisting researchers in deciding whether to engage with the original texts further. These results are consistent with Zhang et al. [21], who reported that hallucinations and inaccuracies were greatly reduced when models were provided with structured, step-by-step instructions.

Regarding the evaluation of the articles' credibility, several patterns were observed. Four of the summaries (all except the fifth) cited publication in a peer-reviewed journal as evidence of credibility, while all five referred to the inclusion of detailed methodology or statistical analysis. Additional elements also appeared: the summary of Article 1 mentioned the study's limitations and the authors' affiliations, Summary 2 highlighted ethical considerations, and Summaries 2 and 3 pointed to the use of theoretical frameworks. While all of these factors are valid indicators of credibility, human-authored annotated bibliographies might not typically state them all explicitly. The findings indicate that ChatGPT-4 lacks the critical selectiveness of human researchers, however, it might produce different output if provided with more specific criteria focusing on article credibility in the prompt.

The final part of each summary contained one or two sentences that outline the relevance of the article. However, these sentences were found to be overly general and provided no specifics as to how these articles fit into the researcher's own work. This further illustrates ChatGPT-4's limited capacity for critical evaluation and corroborates earlier findings reported in the literature [6] [17].

In regard to whether ChatGPT-4 was able to follow the requested annotated bibliography structure as defined in the prompt, the results indicate that the model performed successfully. All five summaries followed the prescribed format, as illustrated in the table below.

Table 1. Structure of ChatGPT-4 Generated Summaries

	Total word count	% of words describing the topic, methodology, and findings	% of words providing an evaluation of credibility and own research usefulness
Article 1	299	65%	35%
Article 2	304	65%	35%
Article 3	299	79%	21%
Article 4	302	73%	27%
Article 5	219	66%	34%

ChatGPT-4.5

The same analytical procedure was applied to the summaries generated by ChatGPT-4.5. As with ChatGPT-4, the summaries did not include any information absent from the original articles. Each summary stated the topic faithfully, and both the methodology and results were described accurately, though the ChatGPT-4 summaries tended to provide more detail in these sections. The output of ChatGPT-4.5 were also comparatively shorter than those produced by ChatGPT-4.

With respect to the evaluation of the articles' credibility, patterns similar to those in ChatGPT-4 were observed. All five summaries referred to methodology and data analyses as evidence of credibility. Summaries 1, 3, and 5 additionally mentioned publication in a peer-reviewed journal, with Summary 1 explicitly naming the journal. Summary 5 further included information about the author and the article's DOI.

Summary 5: *Authored by a scholar with a solid publication record in language education, it appears in a peer-reviewed academic journal with a clear DOI.*

Additionally, Summaries 2 and 4 included sample size, Summaries 3 and 4 referenced theoretical frameworks, and Summaries 2 and 5 noted limitations as further evidence of credibility.

Comparing the two models, it can be concluded that both ChatGPT-4 and ChatGPT-4.5 consistently referred to aspects of methodology and data analyses as primary indicators of credibility, which aligns with accepted standards of academic reliability. Both models also frequently mentioned peer review status as evidence, even though this need not be stated explicitly when articles have already been selected from peer-reviewed journals. The same applies to references to author information, which is typically not noted in annotated bibliographies. Other forms of evidence offered by the models appeared somewhat random, though still reasonable and acceptable. These findings suggest that ChatGPT-4.5 does not demonstrate greater critical selectiveness than ChatGPT-4.

Finally, the last section of the ChatGPT-4.5 summaries addressed the articles' relevance to the personal study. As with ChatGPT-4, these sections consisted of one or two sentences making general links between the findings and the study at hand, showing no notable improvement in this respect over the earlier model.

As with ChatGPT-4, all summaries generated by ChatGPT-4.5 adhered to the requested structure, as shown in Table 2, indicating that both models are consistent in following structural instructions when clearly specified in the prompt.

Table 2. Structure of ChatGPT-4.5 Generated Summaries

	Total word count	% of words describing the topic, methodology, and findings	% of words providing an evaluation of credibility and own research usefulness
Article 1	218	61%	39%
Article 2	230	53%	47%
Article 3	233	70%	30%
Article 4	258	71%	29%
Article 5	220	60%	40%

The distribution of content across topic, methodology, findings, and evaluation was relatively similar in both models, with ChatGPT-4 devoting 65–79% of words to description and ChatGPT-4.5 allocating 53–71%. Given the questionable information occasionally used as credibility evidence and the generally vague statements regarding article relevance, ChatGPT-4 appears to be marginally more useful than ChatGPT-4.5 for generating annotated bibliographies.

Taken together, these observations directly address the first research question of this study, indicating that both models can produce accurate and structurally consistent summaries, with ChatGPT-4 showing marginally greater usefulness than ChatGPT-4.5.

4.2 Survey Results

Quantitative Analysis Results

Table 3 presents the quantitative results of the survey. All 14 participants provided ratings and AI-authorship judgments for each of the 25 summaries.

Table 3. Survey Results

Author	Summary	Rating			Perceived AI			
		Mean	SD	Count of *yes* answers	Count of *no* answers	Count of both *yes* and *no* answers	Mean	SD
Human 1	1	3.57	0.94	2	12	1	0.18	0.37
	2	3.79	0.70	2	12	0	0.14	0.36
	3	3.79	0.58	2	11	1	0.18	0.37
	4	4.07	0.62	2	11	1	0.18	0.37
	5	4.21	0.36	4	8	2	0.36	0.46
Human 2	1	4.21	0.89	9	5	0	0.64	0.50
	2	4.00	0.68	6	8	0	0.43	0.51
	3	4.07	0.62	10	3	1	0.75	0.43
	4	3.93	0.73	9	5	0	0.64	0.50
	5	4.14	0.53	7	5	2	0.57	0.47
Human 3	1	3.29	0.73	1	12	1	0.11	0.29
	2	3.93	0.92	0	14	0	0.00	0.00
	3	2.86	0.77	4	9	1	0.33	0.46
	4	3.86	0.95	3	11	0	0.21	0.43
	5	3.86	0.77	1	13	0	0.07	0.27
ChatGPT-4	1	4.00	1.04	11	2	1	0.82	0.37
	2	4.36	0.84	10	4	0	0.71	0.47
	3	4.00	1.04	9	4	1	0.68	0.46
	4	3.93	1.21	8	4	2	0.64	0.46
	5	3.86	0.95	12	2	0	0.86	0.36
ChatGPT-	1	3.86	0.77	8	6	0	0.57	0.51

4.5	2	4.14	0.86	8	6	0	0.57	0.51
	3	3.93	0.92	11	2	1	0.82	0.37
	4	4.21	0.89	6	7	1	0.46	0.50
	5	3.71	0.64	9	5	0	0.64	0.50

Since all the participants were asked to evaluate the same summaries, the degree of agreement or consistency among them was first calculated using the Interclass Correlation Coefficient (ICC). The result was ICC(2.k)=0.39 indicating poor-to-fair consistency among the 14 experts. This suggests that a considerable portion of the variability in ratings was attributable to differences among raters rather than true differences in the quality of the summaries. To further explore rater variability, mean ratings and standard deviations were calculated for each rater, as shown in Figure 1.

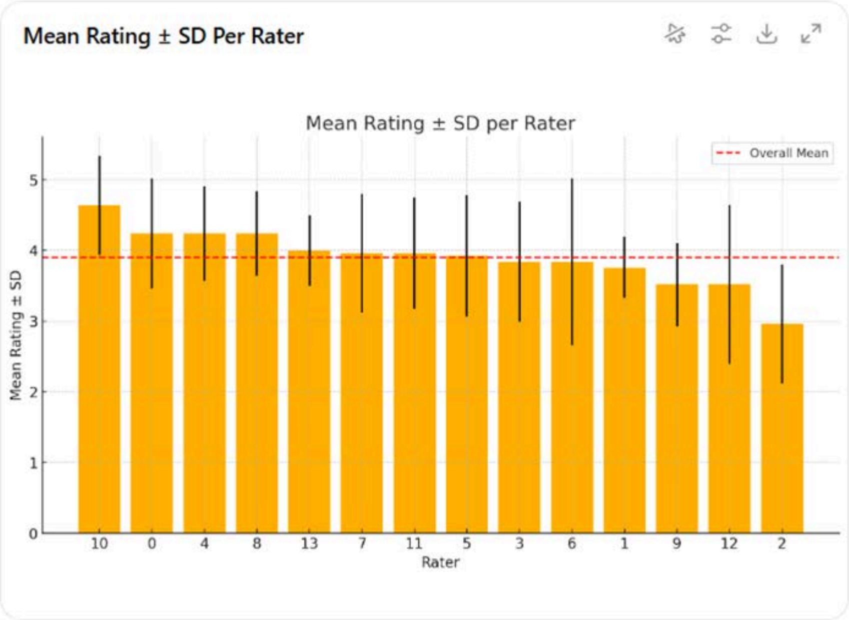

Fig. 1. Mean Rating and Standard Deviation per Rater.

It was found that Participant 10 was consistently more lenient (M = 4.64), while Participant 2 was systematically stricter (M = 2.96). Additionally, Participants 6 and 12 demonstrated the highest rating variability (SD > 1.1). A one-way ANOVA confirmed significant differences across raters ($p < 0.001$), and Tukey's HSD post-hoc tests revealed that multiple rater pairs differed significantly in their mean ratings. These results indicate that participants applied different internal criteria for assessing the quality of the summaries, and that the overall ratings cannot be considered highly reliable as an objective measure of summary quality.

To address this issue, analytical approaches were employed to reduce the influence of rater bias and obtain a more stable estimate of central tendency (i.e. what most raters thought on average even if they disagreed on the details). The data was reanalyzed using median ratings per summary, which revealed that ChatGPT-4 (Mdn = 4.0) and ChatGPT-4.5 (Mdn = 4.0) summaries received comparable evaluations, while Human2 summaries tended to receive slightly higher ratings (Mdn = 4.1) and Human1 and Human3 summaries slightly lower ratings (Mdn = 3.8). Grouping summaries into three categories (Human, ChatGPT-4, ChatGPT-4.5) showed no significant differences. ChatGPT-4 and ChatGPT-4.5 summaries both received median ratings of 4.0, while Human summaries received a median rating of 3.9. Although Human summaries displayed greater variability (SD = 0.39) than ChatGPT-4 and ChatGPT-4.5 (SD = 0.00 for both), the mixed-effects model revealed that these differences were not statistically significant. These findings suggest that ChatGPT-4 and ChatGPT-4.5 produced summaries that were evaluated as being of comparable quality to human-authored summaries, with no significant performance differences between the two models. However, the low inter-rater agreement raises questions about the subjectivity of expert evaluations.

In addition to rating quality, experts were also asked to judge whether each summary was generated by AI, and to offer additional insights into how they differentiated between human and AI-authored texts. Their overall detection accuracy was calculated using a majority vote criterion (≥50% of raters identifying a summary as AI). The result was 56%, suggesting that experts performed only slightly better than chance (50%). Detection accuracy differed by true authorship, whereby experts correctly identified 70% of AI-authored summaries but only 47% of human-authored summaries. This indicates that participants were substantially more accurate at detecting AI-generated summaries than at correctly identifying human-written ones, suggesting the ChatGPT outputs displayed identifiable characteristics.

Finally, a linear regression analysis was conducted to examine whether perceived AI authorship influenced quality ratings. The results revealed a significant effect of perception, whereby summaries perceived as AI-generated received ratings that were, on average, 0.32 points lower than those perceived as human (or uncertain) ($p < 0.001$). This suggests that bias associated with perceived AI authorship negatively impacted ratings regardless of the true authorship of the summary, and may have contributed to the low inter-rater agreement among the participants.

Qualitative Analysis Results

Responses to the open-ended survey questions were analyzed using thematic analysis following Braun and Clarke's [30] framework. An inductive, semantic coding approach was adopted, meaning that codes were generated directly from participants' responses without imposing pre-existing categories, and the analysis focused on the explicit content of the data. Codes were then iteratively grouped into broader themes through comparison and refinement. To structure the analysis and enhance interpretability, themes were organized around four analytical cases: (1) correct classification of human-written texts, (2) misclassification of human-written texts, (3) correct classification of ChatGPT-authored texts, and (4) misclassification of ChatGPT-authored texts. This approach provided a systematic way to capture both the criteria raters used in their judgments and the implications of those judgments for evaluating ChatGPT's effectiveness as a research assistant.

Cases When Human-written Texts Were Correctly Classified

Within the first case, several themes emerged that reflect the criteria raters relied on to distinguish human from AI authorship: information density, word choice and style, structure and organization, credibility and relevance, and the presence of errors.

Information density. Most responses (n = 9) in this category focused on the amount of information presented. Six participants highlighted omissions in methodology or results as indicators of human authorship: "*Not very detailed (no mention of Likert scales, etc.)*" (ART2 HUM3 P7); "*The results and conclusions… are not precise and comprehensive enough*" (ART3 HUM3 P11). Three participants, however, regarded conciseness as a strength: "*Includes appropriate specifics without information overload*" (ART5 HUM3 P13). Overall, participants tended to expect ChatGPT to provide extensive descriptive detail, whereas human authors were expected to exercise greater critical selectiveness. This indicates that ChatGPT is effective for descriptive summarization but less aligned with academic expectations of concise and evaluative writing, which aligns with our findings from the qualitative analysis of the ChatGPT-generated summaries.

Word choice and style. Informal expressions, hedges, and clumsy phrasing were frequently interpreted as markers of human authorship. For example: "*The report hedges at the end with 'The results of this study appear to be credible,' which I think an AI would avoid*" (ART1 HUM1 P8). Language errors further reinforced this perception: "*Also, there are some language mistakes, i.e. wrong choices*" (ART1 HUM3 P4). By contrast, ChatGPT's strengths lie in grammatical accuracy, consistent formality, and the correct use of disciplinary terminology, but its lack of hedging and imperfection may undermine perceptions of authenticity.

Structure and organization. Human-written texts were often perceived as disorganized: "*The summary is very chaotic, not clear(ly organized)*" (ART1 HUM1 P7). In contrast, ChatGPT's rigid adherence to structural conventions can be advantageous for clarity and for facilitating the accurate interpretation of information, though this same rigidity may reduce the stylistic naturalness typically associated with human writing.

Credibility and relevance. Engagement with personal research was a strong indicator of human authorship: "*The conclusion demonstrates actual engagement with how this relates to the writer's own research*" (ART5 HUM1 P13). ChatGPT, by contrast, lacks the ability to establish genuine connections between articles and a researcher's individual project, underscoring a core limitation of the tool as a research assistant.

Errors. Spelling and grammatical mistakes were typically attributed to human authorship: "*Contains small mistakes like 'inlcuded' instead of 'included'*" (ART4 HUM3 P13). ChatGPT, by contrast, largely eliminates such errors, which can enhance efficiency in research-related tasks. However, this very perfection paradoxically makes its outputs more easily identifiable as AI-generated.

Cases When Human-written Texts Were Misclassified

In the second case, themes similar to those identified in the first case re-emerged, particularly regarding word choice, information density, relevance, and errors. This highlights the overlapping criteria participants used and the difficulties of reliably distinguishing between human and AI summaries.

Word choice. Repetitive phrasing and vague expressions led participants to attribute some human-authored summaries to AI: "*Great summary, but wording such as 'well-grounded' and 'valuable insights' are vague… sounding professional*" (ART3 HUM2 P12). This illustrates that stylistic overlap exists, as humans occasionally produce "AI-like" phrasing. For ChatGPT as a research assistant, this suggests that its outputs can blend into established academic conventions, but at the same time may reinforce perceptions of formulaic or overly generic language.

Information density and relevance. A lack of detail regarding how an article related to personal research was often taken as a sign of AI authorship: "*It is impersonal and shows that AI can summarize well but does not relate anything to personal interest*" (ART1 HUM2 P3). When human authors omitted such evaluative context, their summaries were sometimes misclassified as AI-generated. This highlights that while ChatGPT cannot yet convincingly perform evaluative tasks, human writers may also neglect this dimension, contributing to overlap and misclassification.

Errors. Factual mistakes and grammatical slips were at times attributed to AI: "*The summary… combination of generic phrasing and minor grammatical inconsistencies suggests AI*" (ART1 HUM1 P13). This reflects a bias in which participants expected AI to produce errors resembling those made by humans, illustrating how perceptions of AI unreliability can distort evaluations and lead to misclassification.

Cases When ChatGPT-authored Texts Were Correctly Classified

In the third case, participants highlighted themes similar to those observed in the first and second cases, particularly word choice, structure, credibility, and errors, though here these features were more consistently associated with AI authorship.

Word choice and style. Six participants described AI language as "mechanical" or characterized by "typical AI phrasing." For example: "*This one appears to be fully generated by AI. It sounds much more mechanical*" (ART1 ChatGPT-4 P10). Others observed: "*The text is repetitive*" (ART1 ChatGPT-4.5 P12). Word choice was the most frequently cited criterion across all cases, indicating that while ChatGPT ensures consistency and correctness, its predictability and limited lexical variety remain its most prominent stylistic weaknesses.

Structure and information density. ChatGPT was frequently recognized for its rigid organization: "*Extremely systematic progression through all elements*" (ART2 ChatGPT-4 P13). It was also described as detailed and clear: "*Very well organized, detailed and clear*" (ART1 ChatGPT-4 P7). This consistency represents a strength, making ChatGPT reliable for descriptive tasks. However, its rigidity also reduces naturalness, reinforcing perceptions of mechanical style.

Credibility. Formulaic credibility sections were frequently identified as AI markers: "*The final paragraph about credibility reads like a standard AI template*" (ART4 ChatGPT-4 P13). While ChatGPT can reliably insert credibility indicators, they often lack nuance and may appear artificial, limiting its usefulness in evaluative dimensions of research assistance. This observation highlights and reinforces the findings reported earlier, where both models demonstrated limited critical selectiveness in assessing credibility.

Syntactic complexity and punctuation. Participants frequently noted the simplicity of sentence patterns and the presence of "AI-like" punctuation: "*Sentences are mostly simple and to the point… typical AI punctuation*" (ART5 ChatGPT-4 P12). These perceptions indicate that while ChatGPT's preference for clarity and formula supports readability, it also reduces stylistic authenticity and makes its outputs more readily identifiable as AI-generated.

Cases When ChatGPT-authored Texts Were Misclassified

In the fourth case, the only recurring theme concerned credibility and relevance. The main reason AI-generated summaries were mistaken for human was when they included statements about personal research: "*This one has a reference to 'my research'*" (ART3 ChatGPT-4.5 P11). This shows that ChatGPT can mimic evaluative engagement, but such personalization remains surface-level rather than genuine.

Additional insight emerged from participants' comments, indicating that personal attitudes toward AI influenced their evaluations. One participant explicitly acknowledged: "*My obviously negative attitudes towards the use of AI… influence my evaluation of abstract quality*." This demonstrates that predispositions toward AI can bias perceptions, undermining the fairness of comparative evaluations. For ChatGPT as a research assistant, this highlights that user trust is as critical a factor as textual quality in determining its acceptance and effectiveness.

5 Discussion

The purpose of this study was to evaluate the performance of ChatGPT-4 and ChatGPT-4.5 when summarizing applied linguistics (AL) research papers and to examine if these models could function as reliable research assistants. The study addressed two research questions: (1) To what extent do ChatGPT-4 and ChatGPT-4.5 produce accurate summaries that capture key findings while adhering to the required structural conventions? and (2) In what respects do ChatGPT-generated summaries differ from human-authored summaries?

With respect to Research Question 1, both ChatGPT-4 and ChatGPT-4.5 produced summaries that were factually accurate and faithful to the original texts, with no hallucinations when the articles were uploaded individually and the models were guided with explicit, step-by-step prompts. This suggests that the newer versions of ChatGPT have improved in terms of accuracy and supports earlier observations that procedural prompting can reduce hallucinations [21] [22]. ChatGPT-4 tended to provide more detailed accounts of methodology and results, while ChatGPT-4.5 generated shorter and more concise outputs. Both models consistently followed the annotated bibliography structure, demonstrating their reliability for descriptive summarization and structural conformity. However, they showed limited critical selectiveness, often relying on surface-level credibility markers and offering only generalized statements of relevance, confirming limitations highlighted in prior studies [6] [17].

Turning to Research Question 2, the findings indicate that ChatGPT-generated summaries were rated as being of comparable quality to human-authored ones, with no significant performance differences observed across the groups. This suggests that, in terms of overall perceived quality, ChatGPT-4 and ChatGPT-4.5 can already perform at a level similar to human researchers when tasked with descriptive summarization in applied linguistics. This aligns with Hake et al.'s [5] study in medicine, where experts judged ChatGPT-3.5 summaries as high in accuracy and low in bias. However, the low inter-rater agreement highlights the subjective nature of expert evaluations, suggesting that raters relied on diverse criteria and that judgments of summary quality are not straightforward. Thematic analysis helped clarify these underlying criteria. Participants often associated information density, repetitive or mechanical phrasing, rigid structure, and formulaic credibility statements with ChatGPT, while hedging, stylistic variety, evaluative engagement, and occasional errors were taken as indicators of human authorship. This reflects broader concerns raised in applied linguistics and corpus studies that ChatGPT outputs risk appearing formulaic or lacking nuance [8].

The detection task further clarified these distinctions. Experts were considerably better at recognizing AI-authored summaries (70%) than human-authored ones (47%), indicating that ChatGPT output retained identifiable stylistic features that distinguished them from human writing. While this ability to detect AI-authored texts may safeguard against over-reliance on ChatGPT outputs, it also suggests that the models are not yet fully capable of blending seamlessly with human writing styles in academic contexts. Importantly, regression analysis revealed that perceived AI authorship negatively influenced ratings. This bias may be an impediment toward using ChatGPT as a research assistant, echoing concerns in acceptance studies that user perceptions and attitudes significantly shape the adoption of AI tools [15] [16].

The advantages and disadvantages of ChatGPT, as revealed in the findings, become especially clear when compared with human-authored summaries. As shown in Table 4, ChatGPT demonstrates notable strengths in accuracy, structural consistency, grammatical correctness, and efficiency, making it particularly effective for descriptive tasks and the initial stages of literature review. At the same time, its limitations, most notably its inability to engage critically with the material, restrict ChatGPT's usefulness for tasks that demand selectiveness and nuanced interpretation. These results align with prior studies emphasizing ChatGPT's potential as a time-saving tool [31], while confirming its insufficiency for tasks requiring deeper evaluation and theoretical integration [6] [17].

Table 4. Advantages and Disadvantages of using ChatGPT in AL Research Summarization

Dimension	ChatGPT (4 / 4.5)	Human Authors
Accuracy & Faithfulness	Generally accurate and faithful to source texts, few hallucinations when prompted properly.	May omit details or include errors due to oversight or subjectivity.

Dimension	ChatGPT (4 / 4.5)	Human Authors
Structure & Organization	Consistently follows annotated bibliography format. Is rigid and systematic.	Structure sometimes inconsistent or disorganized, but allows flexibility and nuance.
Information Density	Provides comprehensive descriptive detail, reliable for factual coverage.	More selective and concise, exercising critical judgment in deciding what to include.
Word Choice & Style	Grammatically correct, formal, and terminologically precise, but mechanical, repetitive, and formulaic.	Varied, natural, and authentic, but prone to hedging, vagueness, and clumsy phrasing.
Credibility Evaluation	Reliably cites surface-level markers (peer review, methodology, author info). Formulaic and lacking nuance.	Offers more critical selectiveness and nuanced evaluation of credibility.
Relevance & Engagement	General, impersonal statements. Cannot genuinely connect findings to personal research.	Can link articles to own research goals and contexts, showing authentic engagement.
Errors	Few to none (spelling, grammar, factual slips rare). Perfection can make outputs identifiable as AI.	Occasional spelling, grammar, or factual mistakes. Paradoxically perceived as more "human."
Expert Evaluation	Rated as comparable in quality to human summaries (Mdn = 4.0). Identifiable stylistic markers make AI easier to detect. Bias against perceived AI lowered ratings.	Rated similarly (Mdn = 3.9), but more variable in quality. Harder to detect as human when vague or impersonal.
Overall Usefulness	Efficient for descriptive summarization, initial scanning, and standardized outputs.	Stronger in evaluative depth, contextualization, and authentic voice.

6 Conclusion

This study investigated the extent to which ChatGPT-4 and ChatGPT-4.5 can serve as research assistants by generating annotated bibliographies of research articles related to applied linguistics (AL). The findings showed that both models are highly reliable in terms of factual accuracy and structural adherence, as they produced summaries that were comparable in quality to those written by human experts. Their ability to faithfully summarize topics, methodologies, and findings demonstrates they hold potential for tasks that require descriptive summarization and standardized formats.

At the same time, both models exhibited limitations when providing critical evaluation. While they consistently highlighted peer-review status, methodological detail, or author credentials as markers of credibility, the evaluations were formulaic and lacked critical selectiveness. Similarly, statements about the studies' relevance were typically general and impersonal, and failed to demonstrate genuine engagement with research contexts. These weaknesses suggest that even though ChatGPT can assist in providing accurate summaries of the methods and results of AL research articles, it cannot engage critically with them nor provide nuanced reasoning as to why they might be relevant for the researchers' own work. Therefore, human expertise remains invaluable for theory-driven interpretation.

7 Limitations and Further Research

Several limitations of the present study should be acknowledged. First, the dataset consisted of only five articles per model, which does not capture the full range of ChatGPT's performance across different types of articles or subfields of AL. Second, although 14 experts participated in the evaluation, inter-rater reliability was low, which reflects the subjective nature of expert judgments and the influence of biases toward or against AI authorship. This limits the generalizability of the quantitative findings and highlights the need for further research into the criteria experts use when assessing summary quality and identifying AI authorship. Third, the study only tested the summarization of annotated bibliographies. Thus, testing how ChatGPT summarizes other formats such as structured abstracts or critical reviews may reveal different strengths and weaknesses. Fourth, prompting strategies were standardized but not systematically varied, even though prior research suggests that prompting strongly influences ChatGPT's performance [21] [22]. Finally, the analysis was restricted to ChatGPT-4 and ChatGPT-4.5, and other iterations of LLMs may address some of the limitations observed here.

Future research should expand the dataset to include a larger and more diverse corpus of AL research articles and experiment with multiple summary formats. Studies should also systematically vary prompting strategies to assess whether evaluative depth and critical selectiveness can be improved. Further research should also explicitly investigate the criteria on which experts base their perceptions of AI-generated versus human-authored texts, as understanding these criteria would help clarify sources of bias and improve both model development and evaluation practices. Finally, comparative studies across disciplines would help determine whether the findings observed in applied linguistics generalize to other fields, thereby clarifying the broader potential and limitations of ChatGPT in academic research.

Disclosure of Interests.

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., & Ramesh, S. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. npj *Digital Medicine*, *6*(1), 75. doi:10.1038/s41746-023-00819-6
2. O'Connor, S., & ChatGPT. (2023, January). Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? *Nurse Education in Practice, 66*, 103537. doi:10.1016/j.nepr.2022.103537
3. OpenAI. (2023). GPT-4 technical report. https://openai.com/research/gpt-4
4. OpenAI (2025). Introducing GPT-4.5. https://openai.com/index/introducing-gpt-4-5/
5. Hake, J., Crowley, M., Coy, A., Shanks, D., Eoff, A., Kirmer-Voss, K., Dhanda, G., & Parente, D. J. (2024). Quality, accuracy, and bias in CHATGPT-based summarization of Medical Abstracts. *The Annals of Family Medicine, 22*(2), 113–120. https://doi.org/10.1370/afm.3075
6. Salleh, H. M. (2023). Errors of commission and omission in artificial intelligence: Contextual biases and voids of chatgpt as a research assistant. *Digital Economy and Sustainable Development, 1*(1). https://doi.org/10.1007/s44265-023-00015-0
7. Bae, H. (2023). *CHATGPT as a Research Assistant in Experimental Linguistics*. https://doi.org/10.2139/ssrn.4585546
8. Uchida, S. (2024). Using early LLMS for corpus linguistics: Examining chatgpt's potential and limitations. *Applied Corpus Linguistics, 4*(1), 100089. https://doi.org/10.1016/j.acorp.2024.100089
9. Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in CHATGPT: Implications in scientific writing. *Cureus*. https://doi.org/10.7759/cureus.35179
10. Jarrah, A. M., Wardat, Y., & Fidalgo, P. (2023). Using CHATGPT in academic writing is (not) a form of plagiarism: What does the literature say? *Online Journal of Communication and Media Technologies, 13*(4). https://doi.org/10.30935/ojcmt/13572
11. Meyer, J. G., Urbanowicz, R. J., Martin, P. C., O'Connor, K., Li, R., Peng, P.-C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & Moore, J. H. (2023). CHATGPT and large language models in academia: Opportunities and challenges. *BioData Mining, 16*(1). https://doi.org/10.1186/s13040-023-00339-9
12. Rahardyan, T. M., Susilo, C. H., Iswara, A. M., & Hartono, M. L. (2024). CHATGPT: The future research assistant or an academic fraud? [A case study on a state university located in Jakarta, Indonesia]. *Asia Pacific Fraud Journal, 9*(2), 275–293. https://doi.org/10.21532/apfjournal.v9i2.347
13. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of Information Technology. *MIS Quarterly, 13*(3), 319. https://doi.org/10.2307/249008
14. Venkatesh, V., Morris, M.G., Davis, G.B., & Davis, F.D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425. https://doi.org/10.2307/30036540
15. Dahri, N. A., Yahaya, N., Al-Rahmi, W. M., Aldraiweesh, A., Alturki, U., Almutairy, S., Shutaleva, A., & Soomro, R. B. (2024). Extended Tam based acceptance of ai-powered chatgpt for supporting metacognitive self-regulated learning in education: A mixed-methods study. *Heliyon, 10*(8). https://doi.org/10.1016/j.heliyon.2024.e29317
16. Balaskas, S., Tsiantos, V., Chatzifotiou, S., & Rigou, M. (2025). Determinants of CHATGPT adoption intention in higher education: Expanding on Tam with the mediating roles of trust and risk. *Information, 16*(2), 82. https://doi.org/10.3390/info16020082
17. Rahman, M., Terano, H., Rahman, N., Salamzadeh, A., & Rahaman, S. (2023). CHATGPT and academic research: A review and recommendations based on practical ex-

amples. *Journal of Education, Management and Development Studies, 3*(1), 1–12. https://doi.org/10.52631/jemds.v3i1.175

18. Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 74–81). Association for Computational Linguistics.
19. Goyal. T., Li, J. J., & Durrett, G. (2022). *News summarization and evaluation in the era of gpt-3*. (arXiv:2209.12356). https://arxiv.org/abs/2209.12356
20. Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (2023). *Exploring the limits of ChatGPT for query or aspect-based text summarization* (arXiv:2302.08081). arXiv. https://arxiv.org/abs/2302.08081
21. Zhang, H., Liu, X., & Zhang, J. (2023). *Extractive summarization via ChatGPT for faithful summary generation* (arXiv:2304.04193). arXiv. https://arxiv.org/abs/2304.04193
22. Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X. (2023). *AI vs. human – Differentiation analysis of scientific content generation*. arXiv. https://doi.org/10.48550/arXiv.2301.10416
23. Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics, 9,* 391–409. https://doi.org/10.1162/tacl_a_00373
24. Biber, D., & Conrad, S. (2019). *Register, genre, and style*. Cambridge University Press. https://doi.org/10.1017/9781108686136
25. Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2022). ‘Look, I can speak correctly’: Learning vocabulary and pronunciation through websites equipped with automatic speech recognition technology. *Computer Assisted Language Learning, 37*(5–6), 1335–1363. https://doi.org/10.1080/09588221.2022.2080230
26. Dennis, N. K. (2024). Using AI-powered speech recognition technology to improve English pronunciation and speaking skills. *IAFOR Journal of Education: Technology in Education, 12*(2), 107–126. https://doi.org/10.1016/j.caeai.2024.100230
27. Hsu, H.-W. (2024). An examination of automatic speech recognition (ASR)-based computer-assisted pronunciation training (CAPT) for less-proficient EFL students using the Technology Acceptance Model. *International Journal of Technology in Education, 7*(3), 456–473. https://doi.org/10.46328/ijte.681
28. Lyu, J., & Andi, H. K. (2024). The role of technology-enhanced learning in improving English pronunciation and language proficiency. *Sciences of Conservation and Archaeology, 36*(4), 96–102. https://doi.org/10.48141/sci-arch-36.4.24.9
29. Mohammadkarimi, E. (2024). Exploring the use of artificial intelligence in promoting English language pronunciation skills. LLT Journal: *A Journal on Language and Language Teaching, 27*(1), 98–115. https://doi.org/10.24071/llt.v27i1.8151
30. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa
31. Patel, S. B., & Lam, K. (2023). Chatgpt: The future of discharge summaries? *The Lancet Digital Health, 5*(3). https://doi.org/10.1016/s2589-7500(23)00021-3