International Journal of Basic and Applied Sciences, 14 (7) (2025) 33-45

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/g2apsg21
Research paper

Evaluating The Efficacy of A Large Language Model in Scaffolding Research Report Writing for EFL Learners

Taj Mohammad [1], Mohd Nazim [2], Ali Abbas Falah Alzubi [3] *, Soada Idris Khan [4]

[1] Associate Professor, English Skills Department, Preparatory Year, Najran University, Najran, Saudi Arabia
[2] Associate Professor, Department of English, College of Languages and Translation, Najran University, Saudi Arabia
[3] Assistant Professor of Applied Linguistics, Department of English, College of Languages and Translation, Najran University, Saudi Arabia
[4] Teacher and trainer, Sanabel Al Noor International School, Najran, Saudi Arabia
**Corresponding author E-mail: aliyarmouk2004@gmail.com*

Received: October 13, 2025, Accepted: October 22, 2025, Published: November 2, 2025

Abstract

This study examines how the Generative Pre-trained Transformer (GPT) architecture can assist English as a Foreign Language (EFL) learners in writing research reports. In a quasi-experimental design, 60 undergraduates were divided into two groups: a control group and an experimental group, where the latter utilized ChatGPT as a writing assistant. The intervention tested the model's performance in generating and improving text in five areas: organization and structure, research and evidence, analysis and critical thinking, writing quality and presentation, and content and relevance. Writing skills were measured with a 20-item achievement test before and after the intervention. Additionally, 15 participants took part in semi-structured interviews to provide qualitative feedback. The groups had similar skills before the intervention (p > 0.05). Afterward, the experimental group improved in all research report writing skills, particularly in language accuracy, organization, and content relevance. Participants praised ChatGPT for its user-friendly interface, time-saving features, and ability to improve grammar, vocabulary, and report structure. However, some raised concerns about plagiarism and reliance associated with the tool. These findings revealed that large language models (LLMs) can greatly improve the structure and language of student reports. Still, human oversight is needed to ensure critical thinking and factual accuracy.

***Keywords**: ChatGPT; EFL writing; Efficacy; Large Language Models; Research report writing skills.*

1. Introduction

Writing is essential for academic advancement, professional communication, and personal expression [1]. As a reflective process, it encourages critical thinking and creativity while requiring structured thought, clarity of communication, and a solid mastery of language aspects to create effective learning outcomes [2]. Writing has long been at the forefront of scholarly investigation, and EFL teachers in Saudi Arabia have constantly articulated a wide range of opinions and concerns, particularly pedagogical methods and issues in the writing classroom [3], [4]. Teachers believe that academic writing is crucial to the language development of English language learners, necessitating expertise in areas such as organization, coherence, grammar, and vocabulary [5]. Proficient writing abilities enable students to effectively communicate their ideas, clarify their viewpoints, and achieve academic success in a range of professional situations [6].

One of the primary goals of EFL teaching is to build good writing skills in English. EFL teachers have always concentrated on numerous strategies and techniques to help EFL students improve their writing skills. In the age of rapid artificial intelligence (AI) development, using AI-powered writing tools such as ChatGPT presents a viable option to enhance EFL learners' writing skills [7].

Research report writing, as an integral part of academic writing, is a critical skill in EFL education, as it equips students with the ability to articulate complex ideas, synthesize information, and present arguments coherently in a second language [8]. Mastering this skill is essential for academic success and professional communication, enabling EFL learners to engage with global discourse effectively [9]. The consequences of generative AI have extended throughout modern civilization, and education is not immune [10]. The advent of technology and AI has significantly transformed research report writing in EFL contexts by providing tools that enhance drafting, editing, and research processes. AI-powered tools, ChatGPT, can help EFL learners overcome common challenges of poor vocabulary and organizational issues by offering real-time feedback, improving linguistic accuracy, and assisting in the creation of coherent reports [11]. ChatGPT helps EFL learners improve their writing abilities by creating contextually suitable content, suggesting structural modifications, and providing rapid feedback on grammar and style, allowing them to create well-organized and polished research reports [12].

However, it is undeniable that ChatGPT presents numerous obstacles. To successfully include these AI-based writing tools into educational frameworks, it is critical to balance ChatGPT's support with the development of students' natural writing talents properly [13], [14]. Furthermore, Kohnke et al. [11] stated that ChatGPT's answers are not unique and instead contain paraphrased content from uncited sources, which is plagiarism. Meniado [15] also identified concerns about ChatGPT's use in ESL/EFL training, such as the likelihood of academic integrity issues and incorrect responses. Additional systematic examinations of ChatGPT research in the education sphere discovered similar concerns [16], [17], [18].

From a technical standpoint, ChatGPT is an application of OpenAI's GPT (Generative Pre-trained Transformer) architecture. These models are trained on vast datasets through unsupervised learning, enabling them to predict subsequent words in a sequence and generate coherent, contextually relevant text. The model's ability to assist in research report writing can be conceptualized as a function of its natural language generation (NLG) capabilities, which allow it to produce structured text, and its in-context learning, where it adapts its output based on the user's prompt. This study examines the efficacy of a specific AI model (GPT-4) in a constrained, high-stakes domain—academic research reporting—assessing not only learner outcomes but also the model's proficiency in handling complex tasks requiring evidence integration, logical structure, and analytical depth.

While existing literature explores various dimensions of how ChatGPT can enhance EFL learners' general writing skills, few studies specifically examine its effectiveness in improving EFL students' research report writing within the context of the current study. Hence, understanding ChatGPT's role is crucial in properly implementing it into the academic writing curriculum. Given these conditions, the multiple objectives of this study are to evaluate the effectiveness of ChatGPT in improving EFL students' research report writing skills. First, it aims to evaluate the equivalence of experimental and control groups before the intervention by comparing their pre-test scores, ensuring that any observed variations in results are due to the intervention rather than pre-existing discrepancies. Second, the study will compare the significant difference in post-test achievement scores between the experimental group, which uses ChatGPT during the research report writing process, and the control group, which receives traditional teaching, to ascertain the tool's impact on writing performance. Finally, the project plans to document EFL students' experiences and observations of utilizing ChatGPT as a learning tool for research report writing, gathering qualitative insights into its usability, effectiveness, and perceived benefits in enhancing their ability to effectively rewrite research reports.

1.1 Research questions

1. What is the role of ChatGPT in developing EFL students' research report writing skills?
2. What are EFL students' experiences of using ChatGPT in enhancing research report writing skills?

2. Theoretical Framework

This study is based on Computer-Assisted Language Learning (CALL) theory, which provides a foundation for understanding how technology might aid in second and foreign language acquisition. CALL emphasizes the use of digital resources in language learning to provide learners with authentic input, immediate feedback, meaningful engagement, and opportunities for self-directed learning [19]. In the context of this study, CALL is especially relevant to the use of ChatGPT as an AI-powered writing assistant to improve EFL students' writing skills. ChatGPT is an interactive application that enables students to develop, edit, and perfect their writing, in line with the CALL principles of improving linguistic accuracy, fluency, and learner autonomy [20]. Through ChatGPT, students can practice writing activities similar to those in the real world. It reflects CALL's emphasis on contextualized practice and task authenticity, both of which are required for skill transfer beyond the classroom [21]. Furthermore, CALL highlights the importance of giving students control over the writing process [22]. ChatGPT would make this easier by allowing students to plan, compose, revise, and receive feedback at their own pace, which boosts their confidence and proficiency in business communication [20]. In this regard, CALL theory not only explains ChatGPT's pedagogical integration but also underlines its role in connecting traditional EFL education to technology-enhanced learning environments. This study will examine how AI tools can be effectively utilized to improve precision, clarity, and professionalism in students' critical research report writing skills in academic achievement.

While the CALL theory provides the pedagogical context for this investigation, the mechanism of the intervention is rooted in computational linguistics and human-computer interaction (HCI). The students' interaction with ChatGPT represents a feedback loop where human prompts guide the model's transformer-based neural network to generate, revise, and refine text. This study, therefore, operates at the intersection of applied linguistics and the applied science of AI, using the EFL writing classroom as a live laboratory to evaluate the functional utility and limitations of a state-of-the-art LLM in a real-world educational task.

2.1 Literature review

The use of AI in EFL writing has been appreciated for its potential to improve task completion. Cheong and Hong [23] and Domenech [24] found that ChatGPT and other AI technologies may significantly improve students' writing performance in EFL contexts by providing individualized prompts and comments. However, these studies predominantly emphasize measurable performance outcomes, offering limited insight into how AI feedback engages with learners' cognitive and emotional dimensions, a crucial gap that future research must address. Previous research indicates that technology is frequently employed and has become an integral component of language teaching [25], [26], [27]. EFL teachers have adapted and employed a wide range of computer-mediated, internet-based, and digitally focused technology. However, as more sophisticated technologies, such as AI, became more widely used in classroom instruction, generative AIs, such as ChatGPT, emerged as game changers in a variety of fields, including education [28], [10], [29], particularly in language instruction [11]. Still, as Kohnke et al. [11] and Chen and Xu [30] caution, the pedagogical promise of AI is tempered by persistent technical and ethical challenges such as factual "hallucinations," data privacy concerns, and the potential erosion of learner autonomy when applied uncritically. This tension between innovation and oversight echoes an ongoing debate in applied linguistics about AI's role as either a cognitive partner or a substitute for human scaffolding. The capabilities of these tools are built upon foundational breakthroughs in Natural Language Processing (NLP), specifically the transformer architecture, which enabled the efficient training of LLMs on massive text corpora [31]. These models, including the GPT series that powers ChatGPT, function as powerful pattern-matching systems that generate text by predicting probable next tokens in a sequence, a process that yields high fluency and coherence but does not equate to human-like understanding or reasoning [32].

This distinction underscores the growing call among scholars for critical literacy towards AI-generated text. Although ChatGPT can replicate advanced discourse structures, it often lacks semantic nuance and pragmatic sensitivity, resulting in content that may be rhetorically convincing yet factually flawed [33]. Gayed et al. [34] conducted a study to improve EFL students' writing skills and help them overcome cognitive challenges. The findings showed that the AI-powered language learning application significantly improved students' writing skills and reduced the anxiety that they had when completing writing assignments. While these affective benefits hold value, they may obscure an excessive dependence on AI-generated text—a concern among some researchers who argue it can hinder the development of independent thought [35].

Özçelik and Ekşi [36] demonstrate the context-dependent value of AI in writing pedagogy. Their investigation into ChatGPT's ability to teach English writing registers yielded mixed results. On one hand, the AI was "extremely effective" as a learning assistant for the formal register, aiding in paragraph structure and providing constructive feedback. On the other hand, the study identified significant shortcomings, as students found it less effective for casual and neutral registers. This highlights a critical limitation, and the researchers recommend that educators strategically integrate ChatGPT to leverage its strengths while actively compensating for its inadequacies in less formal writing contexts.

This reflects a broader scholarly consensus that AI usefulness is highly context-dependent. Although it proves effective in structured academic writing, its performance tends to falter in tasks demanding socio-pragmatic sensitivity or stylistic adaptability [17].

Abdullayeva and Musayeva [37] investigated the impacts of ChatGPT on EFL students' writing skills and discovered that it was effective in offering prompts, instant feedback, and revision recommendations. Yet, their findings—echoing those of [38]—suggest a potential motivational bias: students frequently overrate the reliability of AI-generated output, largely due to its fluency, a tendency associated with automation bias [32]. This underscores the pedagogical imperative for deliberate and critical engagement with AI feedback, rather than its unexamined acceptance. Nazari et al. [38] undertook a real-world experimental investigation to compare the effects of AI-assisted language learning on EFL learners' writing skills. Findings revealed that students who used the AI-powered application performed better in their writing. AI learners demonstrated high levels of behavioral, cognitive, and emotional involvement in AI-assisted writing exercises. Nevertheless, their study did not investigate whether this engagement leads to lasting skill transfer once AI assistance is withdrawn—an aspect still insufficiently examined in the current EFL-technology literature.

Kohnke et al. [11] investigated ChatGPT's features in relation to language learning. Their evaluation of present technology has produced preliminary insights into ChatGPT's ability to support language teaching and learning by providing a set of learning activities that the platform could use. They found from their review that ChatGPT might help in language acquisition as it simulates real-world communication.

Yet, the authenticity of this simulation remains a point of contention. Although ChatGPT can emulate human conversation, it falls short of capturing the sociocultural and interpersonal nuances that characterize genuine communicative exchanges [13]. This limitation underscores the need for pedagogical frameworks that pair AI integration with reflective dialogue and teacher facilitation, ensuring that human interaction remains central to language learning. Collectively, the reviewed studies reveal both the instructional potential and conceptual fragility of incorporating ChatGPT into EFL writing contexts. While many highlight gains in grammar, structure, and learner confidence, fewer delve into critical literacy, factual accuracy, or ethical considerations. This disparity indicates a prevailing emphasis on performance outcomes at the expense of epistemic and ethical engagement—dimensions that are increasingly essential in AI-supported education.

The literature review demonstrates that there has been plenty of research on the functionality of ChatGPT in boosting several parts of EFL learners' writing skills. However, there is essentially little research on the role of ChatGPT in developing students' research report writing skills as part of academic tasks, particularly at Najran University in Saudi Arabia. As a result, this study aims to fill this research gap. By doing so, it not only builds on existing empirical research but also enriches the growing critical discourse surrounding AI-mediated writing support, particularly its alignment with pedagogical oversight, ethical responsibility, and contextual sensitivity in EFL environments. To further corroborate the research findings, the study investigates learners' experiences and observations about the effectiveness of AI-assisted language learning training.

3. Methodology

3.1 Research design

This study used a quasi-experimental approach with two groups of EFL students from Najran University: an experimental group that received ChatGPT-assisted instruction in research report writing, and a control group that received traditional instruction. A total of 60 students were chosen and distributed equally between the two groups. A 20-item achievement test (validated through expert review, item analysis, and reliability testing with Cronbach's Alpha = 0.95) was used as a pre- and post-test to assess changes in students' writing performance across domains such as content and relevance, organization and structure, research and evidence, analysis and critical thinking, writing quality, and presentation. Normality tests confirmed the applicability of parametric analyses, and independent-samples t-tests with effect size estimations were performed utilizing SPSS. In addition to quantitative metrics, semi-structured interviews were conducted with a sample of experimental students to document their experiences and observations while using ChatGPT as a learning tool. Thematic analysis was used to examine qualitative data [39], providing more information about learners' impressions of the tool's usability and effectiveness.

3.2 Population and sample

The study included 90 undergraduate students aged 18-22 participating in the Reading and Writing course, chosen through purposeful sampling from PY in the second semester of the 2025 academic year. A total of sixty students were selected from this population, divided into two groups: experimental and control. The sample belongs to Arabic-speaking students who aspire to be engineers, doctors, and computer science experts, and must then pass PY to continue their studies in their respective disciplines. They come to PY after finishing upper secondary school and meeting the enrollment requirements for Najran University. They completed their upper secondary schooling at Najran. They studied English as a foreign language and were classified as PY level between elementary and upper intermediate according to a diagnostic test.

To strengthen the internal validity of the quasi-experimental design, several measures were implemented to control for potential confounding variables. First, the purposeful sampling of students from the same course and proficiency level ensured a homogeneous population at the outset. The group assignment followed existing classroom sections and avoided selection bias while maintaining ecological validity.

Crucially, a pre-intervention needs analysis survey was administered to both groups to assess key variables such as prior experience with AI tools and self-reported motivation for writing. The survey results indicated minimal and statistically comparable levels of prior AI exposure between the control group (M= 1.2, SD= 0.4) and experimental group (M= 1.3, SD= 0.5), confirming that this variable was unlikely to skew the results. Furthermore, both groups were taught by the same instructor, followed the identical curriculum, used the same textbook and evaluation rubrics, and received the same total instructional time. The only systematic difference was the integration of ChatGPT for writing assistance in the experimental group.

The study's ethical approval number is 0076-00076-DS. Before providing consent, the student participants were fully informed about the research process. If they agreed, participants might withdraw or skip any questions at any time. In addition, participants were free to ask questions about the research. They were advised that participating in the study would provide no direct or indirect benefits. Participants were assured that all information obtained would be kept strictly confidential and used only for the study. Moreover, they were given the research team's contact information in case they needed any additional information or clarification.

Furthermore, this research utilized Generative AI to assist with the initial analysis of qualitative data, search for authentic academic sources, proofreading, and editing. It was also used to rephrase and organize ideas, and draft and format sections such as results and references.

3.3 Instruments

The researchers administered a test and semi-structured interview, with the test developed based on a review of relevant literature and their own teaching experience [40]. The test assessed the efficacy of the intervention program by comparing pre- and post-test outcomes. The five main themes of the test were content and relevance, organization and structure, research and evidence, analysis and critical thinking, writing quality, and presentation. Using the rubrics, the students were required to compose a research report on a “local start-up”. According to the rubrics, students were expected to write the research report in the form of a paragraph for evaluation, which took about twenty minutes on average to complete, and was administered by the researchers to the same population in the PY building.

3.3.1 Semi-structured interview

A semi-structured interview was conducted to get feedback from EFL learners on the efficacy of ChatGPT-mediated instruction in enhancing research report writing skills. The interviews took place at Najran University during the second semester of the 2024–2025 academic year, in a designated lecture room within the Language Skills Unit of the Deanship of Preparatory Year building. This setting offered a professional, comfortable, and distraction-free environment conducive to focused discussion. A total of 15 participants were purposefully selected from the experimental group to reflect varied proficiency levels and engagement patterns, ensuring diverse perspectives despite the limited sample size. The researcher questioned each participant individually, while another colleague served as an observer, noting nonverbal cues such as reluctance, confidence, and enthusiasm. The interviews lasted approximately 10-15 minutes per student and were performed two weeks after the post-test.

To mitigate potential biases such as social desirability, students were assured that their responses would remain confidential and have no bearing on their academic evaluation. The interviewer employed neutral, open-ended questions and actively encouraged candid reflections, emphasizing that both positive and negative experiences with ChatGPT were equally valuable to the study. Follow-up prompts were also used to clarify responses and minimize the likelihood of participants offering overly favorable accounts. The interviews centered on challenging aspects of research report writing, such as content and relevance, organization and structure, research and evidence, analysis and critical thinking, writing quality, and presentation. To maintain academic integrity, semi-structured questions were created based on past research and the researcher's teaching experience [11], [13], [14], [12].

The interview protocol was based on two key categories: experiences and observations.

- Experiences: Students were asked to describe how ChatGPT assisted them in creating, organizing, and revising research reports by writing paragraphs, as well as how it influenced their confidence and motivation to write. The inquiries included the ones that followed: "How did ChatGPT change the way you approach research report writing?”, "Which aspects of your writing improved the most (e.g., grammar, vocabulary, organization, research)?" Alternatively, "What challenges did you encounter while using ChatGPT?"
- Observations: Students were asked to reflect on how ChatGPT influenced their writing habits, approaches, and learning process. These included the following inquiries: "What differences do you notice in your writing before and after using ChatGPT?”, "Did ChatGPT make research report writing easier or more difficult for you?”, and finally: "How did ChatGPT affect the way you edit and review your work?"

Throughout the interviews, several themes emerged. Many students had favorable experiences, claiming that ChatGPT helped them overcome difficulties with language choice, sentence transitions, researching the topic, and logical flow. Others cited increased autonomy in revising drafts, a higher awareness of grammatical accuracy, and improved conceptual organization. However, a few participants expressed concerns about over-reliance on AI and an inclination to accept its recommendations without a critical analysis. Finally, the researcher reviewed interview transcripts using Braun and Clarke's [39] thematic content analysis method, which allowed for the identification of common themes, patterns, and discrepancies among student responses.

3.4 Intervention

For four months (February-May 2025), the experimental group (N = 30) and the control group (N = 30) met twice a week for 200 minutes (100 minutes per class) in the same university computer laboratory. The intervention aimed to see if ChatGPT-enhanced writing instruction improved Preparatory Year (PY) students' argumentative paragraph/essay writing when compared to traditional instruction. Both groups followed the same curriculum, class objectives, activities, and evaluation rubrics; the only difference was that the experimental group used ChatGPT for 50 minutes per session. They received instruction based on the program's mandated writing textbook sections, which included research report writing with a focus on content and relevance, organization and structure, research and evidence, analysis and critical thinking, and writing quality and presentation. The content emphasized providing a thorough and exhaustive examination of individuals, covering all aspects of their lives and accomplishments with precision and depth, while also ensuring the report was well-structured with a clear introduction, body, and conclusion presented logically and persuasively. The research requirements included the use of reliable sources, relevant citations, and a thorough investigation backed by critical thinking and analysis that made available a range of viewpoints and convincing evidence in conclusions. Standards for writing quality and presentation required that work be well-written, grammatically correct, and presented in a way that was engaging, professional, and consistent. To protect privacy, no personal information was uploaded, and anonymized task IDs were used. Students mainly used OpenAI ChatGPT-4 through institutional logins on lab desktop PCs, even though personal mobile devices were used in cases where a workstation was unavailable or malfunctioning. To support this, the instructor experimented with model lessons and assisted students when necessary. The instructor also led a two-hour micro-workshop on prompt design, safety and ethical considerations, troubleshooting, and aligning AI outputs with textbook rubrics.

The AI intervention utilized the GPT-4 architecture, accessed via the ChatGPT interface. The experimental protocol was designed to test the model's few-shot and zero-shot learning capabilities, where students provided iterative prompts to generate outlines, draft sections, and refine arguments based on the provided rubric. Students were guided to initiate their interactions with broad, task-oriented prompts (e.g., "Generate a report outline on the impact of social media on academic writing") and progressively refine their queries through iterative scaffolding. Examples included requests such as "Revise the introduction to include a thesis statement and transition sentence" or "Provide three examples of academic evidence that could support paragraph two." This process of prompt refinement was integral to the training, as students actively compared outputs, evaluated clarity and tone, and reworded prompts to enhance coherence and academic formality (e.g., "Rephrase this paragraph to sound more formal and academic without changing meaning"). This process effectively positioned students as human evaluators within a human-in-the-loop system, providing real-time data on the model's ability to produce academically valid and structurally sound research reports.

To promote methodological transparency, students were provided with a prompt checklist that categorized examples into four key areas: (1) Content generation (Draft a paragraph explaining the significance of the topic), (2) Structural enhancement (Suggest transitions between the literature review and analysis sections), (3) Stylistic refinement (Edit this text to use an objective academic tone), and (4) Critical reflection (List three weaknesses in this paragraph's argument and suggest improvements). Each prompt and its corresponding AI-generated response were documented in a digital log, allowing systematic analysis of interaction patterns between students and the AI tool.

Stage 1: Pre-Intervention (Week 1–2)

The planning sessions aimed to establish baselines, standards, and preparation for using AI tools. Students first completed a 40-minute pre-test on writing argumentative paragraphs without the help of AI. They also took a brief survey assessing device access, confidence, prior AI experience, and motivation towards writing. Next, students attended a 30- to 40-minute ethics and orientation session. Here, they learned about AI's limitations, the importance of avoiding copy-paste submissions, and the need to rewrite outputs in their own words while citing and explaining modifications. Privacy was emphasized by using anonymized task codes and instructing students not to include names, email addresses, or personal identification numbers in prompts. Finally, a 20–25-minute micro-training session explained the process, including prompt development, outline creation, sample generation, critique, and revision.

Stage 2: During-Intervention (Weeks 3–14)

The experimental group participated in an extra 50 minutes of ChatGPT-assisted practice in addition to the two weekly 100-minute textbook-based classes. This practice focused on reinforcing five specific aspects of writing research reports: topics and relevance, organization and structure, research and evidence, analysis and critical thinking, and writing quality and presentation. Students critically analyzed sample reports for content depth and relevance during a 30- to 40-minute model-analysis exercise at the start of each session. This was followed by 40- to 50 minutes of supervised practice on report structure, argument planning, and creating cohesive paragraphs with smooth transitions. After that, students worked independently for 50 minutes at lab PCs or mobile devices, using a prompt checklist and one-on-one guidance from the instructor to improve how they integrated evidence, citation styles, and analytical commentary. For instance, students were encouraged to engage in iterative dialogue with ChatGPT, beginning with prompts such as "What makes this argument weak?" and following up with queries like "How can I improve it using more critical reasoning?" or "Suggest a citation format for this paraphrased evidence." This recursive feedback loop enhanced the internalization of revision strategies and fostered greater critical awareness of the strengths and limitations of AI-generated content. In contrast, the control group finished comparable paper-based assignments that included feedback from the teacher. Using a rubric based on the five targeted skills, both groups spent 30 to 40 minutes editing and peer reviewing. Each session concluded with a 15-to-20-minute exit ticket that included a suggested improvement and a reflective question. ChatGPT was utilized in the experimental group to help with idea generation, micro-explanations, alternative wording, and paragraph-level writing. However, students were specifically expected to edit and customize AI's outputs, recording any modifications. Additional resources included a QuickStart guide, mini-demos that addressed common report-writing problems, and one-on-one coaching to assist students in reframing ambiguous prompts into specific ones.

Stage 3: Post-Intervention (Weeks 15–16)

The post-test was administered using the same rubric and conditions as the pre-test: a 40-minute writing exercise devoid of AI assistance. The findings showed that, in comparison to the pre-test, students' writing abilities had improved statistically significantly. A stratified subset of the experimental group participated in 10-to-15-minute semi-structured interviews to gather qualitative insights. Students reported improved writing skills and confidence as well as generally positive experiences with the intervention, despite a few minor technical difficulties. All AI logs were stored under anonymized IDs and used exclusively for research purposes. The control group's students spent the same amount of time on textbook exercises, teacher-led drills, and manual drafting and revision as they did on the textbook modules, peer review, and guided practice.

3.5 Validity and reliability

3.5.1 Test validity

When the test was first created, it was shown to ten validators, including English language experts from the teaching faculty. It was their responsibility to assess the test's appropriateness, the suitability of questions for the test's intended use, the formulation of questions and domains in accordance with test construction standards, the linguistic phrasing of items, and the clarity of instructions. The final test included five domains based on the validators' feedback, which included item additions, changes, and deletions. Here are both versions, before and after, of the modifications to the validators.

Table 1: Experts' Comments

Stage	Prompt(s)	Domain Assessed
Test before modification	Write a research report on a local startup.	Contents and relevance, organization, research and evidence, analysis, writing quality
Test after modification	Write a research report on a local startup. Or Write a research report on a famous businessman.	Contents and relevance, organization and structure, research and evidence, analysis and critical thinking, writing quality, and presentation

3.5.2 Internal consistency

The test was administered to a pilot sample of 20 students from outside the study sample. Pearson correlation coefficients were calculated between the score of each question and the total score, as well as between the questions and the domain they belong to. Table 2 illustrates the results.

Table 2: Pearson Correlation Coefficients Between Question Scores and Total Test Score, and between Questions and Their Respective Domains

Question	dom1	dom2	dom3	dom4	dom5	Total
A1	0.789**					0.823**
A2	0.653**					0.652**
A3	0.691**					0.665**
A4	0.679**					0.624**
A5	0.876**					0.850**
A6	0.574**					0.587**
A7		0.759**				0.762**
A8		0.758**				0.685**
A9		0.817**				0.796**
A10		0.758**				0.642**
A11			0.781**			0.661**
A12			0.653**			0.652**
A13			0.856**			0.766**
A14			0.908**			0.940**
A15				0.746**		0.810**
A16				0.792**		0.661**
A17				0.797**		0.748**
A18				0.862**		0.748**
A19					0.866**	0.724**
A20					0.862**	0.719**

*Note: ** Correlation is significant at the 0.01 level (2-tailed).*

All test items had statistically significant positive correlations with their respective domains and the overall test score at the 0.01 level according to the Pearson correlation analysis results, demonstrating strong internal consistency. Most of the items displayed moderate to very high correlations, with correlation coefficients ranging from 0.574 to 0.940.

For instance, item A6 displayed the lowest but still satisfactory correlations (0.574 with its domain, 0.587 with the total score), whereas item A14 had the highest correlation with both its domain (0.908) and the total score (0.940). These results demonstrate that the test items are reliable indicators of the domains they are designed to measure a significant contribution to the overall concept under evaluation. Pearson correlation coefficients between the domains and the total test score appear in Table 3.

Table 3: Pearson Correlation Coefficients between Domains and the Total Score

Domains	Person correlation
1. Contents and relevance	0.939**
2. Organization and structure	0.933**
3. Research and evidence	0.945**
4. Analysis and critical thinking	0.929**
5. Writing quality and presentation	0.835**

All five domains demonstrated high, positive, and statistically significant correlations with the overall test score, ranging from 0.835** to 0.945*), according to the results in Table 3. Although still at a high level, Domain 5 exhibited the lowest correlation (0.835**) and Domain 3 the highest (0.945**). These findings support the test's general validity and internal consistency.

3.5.3 Test reliability

The test reliability was calculated using Cronbach’s Alpha, as shown in Table 4.

Table 4: Cronbach’s Alpha Coefficient

No.	Domain	Reliability Coefficient
	Total Score	0.95

With Cronbach’s Alpha coefficient of 0.95 for the overall score, the results in Table 4 reveal that the achievement test had a high degree of reliability. This value shows that the test items are consistent and reliable for gauging students' performance, as it is higher than the generally accepted cutoff of 0.70.

3.5.4 Test item analysis

A pilot sample of 20 students not part of the study sample was given the test, and the test questions were examined to verify their validity by computing discrimination and difficulty coefficients.

3.5.5 Difficulty coefficients

The difficulty coefficient for the test questions was calculated according to Odeh [48] using the following formula:

Difficulty Coefficient = Total Score of the Question / (Number of Students * Question Score)

The difficulty coefficients for the test questions based on the pilot sample results appear in Table 5.

Table 5: Difficulty Coefficients for the Achievement Test Questions

Questions	Difficulty Coefficient	Questions	Difficulty Coefficient
1	0.50	11	0.65
2	0.55	12	0.60
3	0.50	13	0.55
4	0.55	14	0.55
5	0.60	15	0.55
6	0.60	16	0.60
7	0.60	17	0.55
8	0.45	18	0.60
9	0.55	19	0.55
10	0.45	20	0.65

Table 5 shows the test questions' difficulty coefficients ranged from 0.45 to 0.65. Odeh [48] notes that items with coefficients between 0.20 and 0.80 are appropriate and should be retained. The test's average coefficient was 0.56, indicating moderate difficulty.

3.5.6 Discrimination coefficients

The discrimination coefficients for the objective test items were calculated according to the following formula [48]: Discrimination Coefficient = (Nu – NI) / N

- Nu = Number of students in the upper group who answered the question correctly.
- NI = Number of students in the lower group who answered the question correctly.
- N = Number of students in one of the groups.

The students were split into two groups to determine the discrimination coefficient for each question: a lower group, which included 50% of the students with the lowest scores (10 students), and an upper group, which included 50% of the students with the highest scores (10 students). According to Odeh [48], measurement professionals created the following reference values for evaluating test items:

- Items with a negative discrimination coefficient are discarded.
- Items with a discrimination coefficient less than 0.20 are recommended for deletion.
- Items with a discrimination coefficient of 0.20 or higher are acceptable.

Table 6 presents the discrimination coefficients for the test questions.

Table 6: Discrimination Coefficients for the Achievement Test Questions

Questions	Discrimination Coefficient	Questions	Discrimination Coefficient
1	0.80	11	0.50
2	0.30	12	0.60
3	0.60	13	0.70
4	0.70	14	0.90
5	0.80	15	0.70
6	0.60	16	0.80
7	0.60	17	0.70
8	0.50	18	0.80
9	0.70	19	0.70
10	0.50	20	0.70

According to Table 6, the discrimination coefficients for every test item fell between 0.30 and 0.90, which is within the acceptable range (≥0.20) recommended by [48]. This suggests that the items were successful in differentiating between students who performed well and those who did not. Item 14 had the highest discrimination (0.90), and item 2 had the lowest acceptable value (0.30). Overall, the findings support the test items' strong discriminatory power and suitability for gauging students' academic performance.

3.5.7 Normality of distribution

The Kolmogorov-Smirnov test was used to verify the normality of the distribution of the study sample's scores on the test in the pre- and post-applications, as shown in Table 7.

Table 7: Kolmogorov-Smirnov Test for Normality of Distribution of Study Sample Scores in Pre- and Post-Applications

	group	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	Df	Sig.	Statistic	Df	Sig.
Pre	control	0.105	30	0.200*	0.958	30	0.597
	experiment	0.101	30	0.200*	0.987	30	0.668
Post	control	0.138	30	0.132	0.923	30	0.124
	experiment	0.169	30	0.67	0.917	30	0.112

For both the control and experimental groups, the Kolmogorov-Smirnov and Shapiro-Wilk tests produced non-significant values ($p > 0.05$) in the pre- and post-tests, as shown in Table 7. This suggests that the students' scores were distributed normally, meeting the normality assumption and enabling additional analysis using parametric statistical tests, the independent samples t-test.

3.6 Data analysis

The researchers used the Statistical Package for the Social Sciences (SPSS) version 26 to answer the study questions, extracting the following:

- Pearson correlation coefficient for internal consistency validity.
- Cronbach's Alpha for reliability.
- Kolmogorov-Smirnov test.

- T-test for independent samples.
- Eat squared for effect size.

4. Results

4.1 Research question 1: What is the role of ChatGPT in developing EFL students' research report writing skills?

The t-test for independent samples was used to determine the significance of differences between the mean scores of the control and experimental groups in the pre-intervention, as shown in Table 8 and Figure 1.

Table 8: T-Test for Independent Samples to Show Significance of Differences between Mean Scores of Controls and Experimental Groups in Pre-intervention

Domains	Group	N	Mean	Std. Deviation	t	Df	Sig. (2-tailed)
1. Content and relevance	Control	30	1.02	0.61	0.465	58	0.644
	Experiment	30	0.95	0.5			
2. Organization and structure	Control	30	0.80	0.39	0.419	58	0.677
	Experiment	30	0.85	0.53			
3. Research and evidence	Control	30	0.85	0.27	1.015	58	0.314
	Experiment	30	0.75	0.47			
4. Analysis and critical thinking Writing	Control	30	0.90	0.33	1.577	58	0.120
	Experiment	30	0.70	0.61			
5. Writing quality and presentation	Control	30	0.60	0.20	0.803	58	0.425
	Experiment	30	0.55	0.27			
Total	Control	30	5.30	0.79	1.347	58	0.183
	Experiment	30	4.97	1.1			

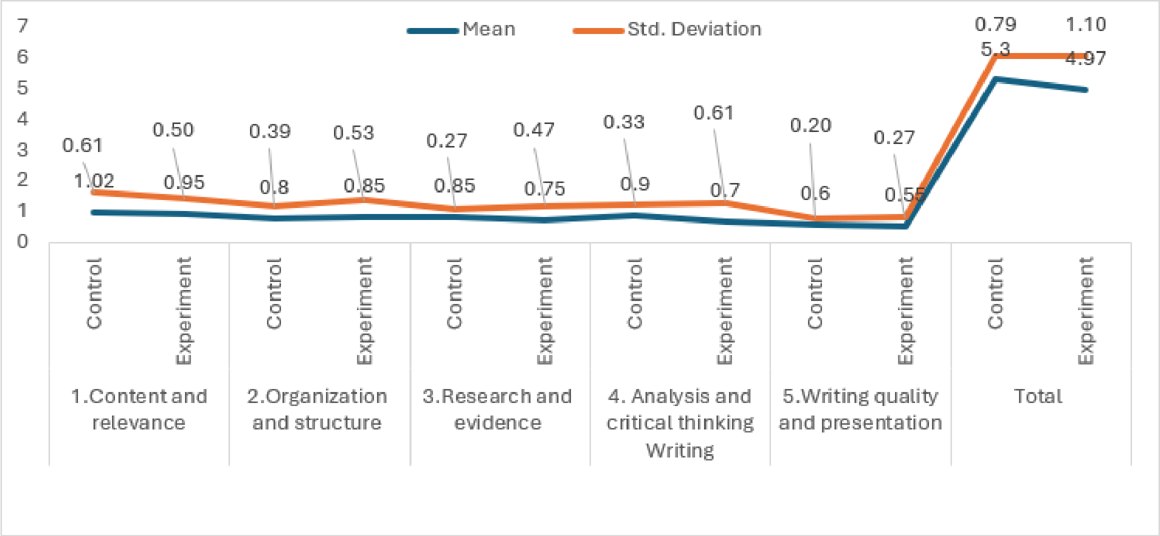

Fig. 1: Comparison of Pre-intervention Scores for Control and Experimental Groups

In every pre-test domain, there were no statistically significant differences ($p > 0.05$) between the experimental and control groups, as depicted in Table 8. With the experimental group's overall mean score of 4.97 and the control group's at 5.30, the two groups' mean scores were comparatively close. By confirming that the groups were equal at baseline, these results guarantee that any discrepancies observed in the post-test can be ascribed to the intervention and not to underlying inequalities.

The t-test for independent samples was used to determine the significance of differences between the mean scores of the control and experimental groups in the post-test, as shown in Table 9.

Table 9: T-Test for Independent Samples to Show Significance of Differences between Mean Scores of Control and Experimental Groups in the Post-Intervention

Domain	Group	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Eta Squared	Level
1. Content and relevance	Control	30	3.37	0.59	11.982	58	0.000	0.712	Large
	Experiment	30	5.17	0.58					
2. Organization and structure	Control	30	2.52	0.45	8.084	58	0.000	0.530	Large
	Experiment	30	3.42	0.42					
3. Research and evidence	Control	30	2.30	0.31	10.179	58	0.000	0.641	Large
	Experiment	30	3.33	0.46					
4. Analysis and critical thinking Writing	Control	30	1.95	0.24	18.422	58	0.000	0.854	Large
	Experiment	30	3.02	0.21					
5. Writing quality and presentation	Control	30	.77	0.25	6.547	58	0.000	0.425	Large
	Experiment	30	1.25	0.32					
Total	Control	30	11.80	1.03	17.476	58	0.000	0.840	Large
	Experiment	30	16.67	1.12					

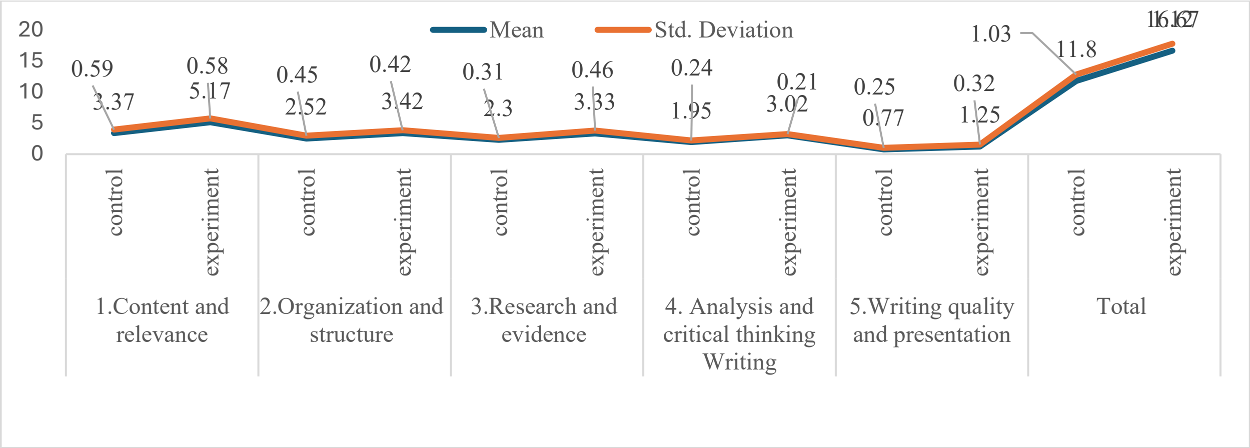

Fig. 2: Comparison of Post-intervention Scores for Control and Experimental Groups

The findings in Table 9 demonstrate that the experimental group outperformed the control group in the post-test across all domains, with statistically significant differences ($p < 0.001$) between the two groups. M = 5.17 vs. M = 3.37 in domain 1 (contents and relevance) and M = 16.67 vs. M = 11.80 in total scores are just two examples of the consistently higher mean scores attained by the experimental group. In comparison to traditional education, the use of ChatGPT as an intervention significantly improved students' research report writing skills, as evidenced by effect sizes (Eta-squared) ranging from 0.425 to 0.854, all at a large level.

The practical significance of these results is underscored by the large effect sizes (η^2). According to Cohen's benchmarks, an η^2 of 0.14 is considered a large effect. Our findings far exceed this threshold, with the intervention having a particularly strong impact on Analysis and Critical Thinking ($\eta^2 = 0.854$) and the Total score ($\eta^2 = 0.840$). This indicates that the ChatGPT-assisted instruction was not merely statistically superior but also made a substantial, real-world difference in the students' writing capabilities. For instance, the mean total score for the experimental group (16.67) was over 41% higher than that of the control group (11.80), demonstrating a marked improvement in overall research report writing performance. The domain of Writing Quality and Presentation, while still showing a large effect ($\eta^2 = 0.425$), had the most modest gain, suggesting that while ChatGPT helped polish the final product, its greatest impact was on the more complex, higher-order thinking processes involved in research and analysis.

4.2 Research question 2: What are EFL students' experiences of using ChatGPT in enhancing research reports? Writing skills?

The data generated from interviews were analyzed using Braun and Clarke's [39] six phases of thematic analysis. The analysis uncovered both experiential and attitudinal dimensions of students' engagement with ChatGPT, providing insights into how the tool supported their research report writing skills and influenced their learning perceptions.

The first phase involved familiarization with the data, which required reading and re-reading the transcripts to gain an in-depth understanding of how students described their experiences. Initial observations indicated that students consistently highlighted ChatGPT's role in improving report content, structure, grammar, spelling, and overall writing quality, while also emphasizing its accessibility and motivational value.

During the second phase, related codes were systematically collated into broader themes, which are: writing skill improvement (content development, sentence structure and sequencing, grammar and language accuracy), confidence and motivation (boosting confidence, encouraging engagement), and user experience and tool perception (interest and engagement, ease of use, educational usefulness).

During the third phase, the research identified three major themes. First, improved report writing skills developed because of greater content, structure, and grammatical accuracy, demonstrating ChatGPT's involvement in writing quality. Second, students reported satisfaction with its utility and usability, praising both its learning value and ease of use. Third, ChatGPT increased confidence and motivation, facilitating both effective and linguistic progress. Overall, the tool benefited EFL learners cognitively, technologically, and emotionally.

The themes were thoroughly examined in the fourth phase in comparison to the dataset to guarantee their uniqueness, coherence, and clarity. A single theme emphasized content relevance, and the efficient use of evidence was developed from the overlap between the codes about research and evidence, as well as content and relevance. The codes associated with spelling and grammatical enhancements were also verified to be connected and grouped under the more general heading of writing precision and presentation quality. This process of improvement made sure that the themes preserved clear analytical bounds while encapsulating the core of the students' experiences.

Three opinion-based themes—positive views of ChatGPT's usefulness, usability, and accessibility, and pedagogical integration under teacher guidance—and five experiential themes—content relevance and evidence use, organization and structure, analysis and creativity, writing accuracy and presentation quality, and confidence and motivation—were finalized in the fifth phase. When combined, they show how students view AI-assisted learning as well as their skill-based gains.

In the sixth phase, five themes were used to record students' experiences using ChatGPT. It addressed the essential features of report writing (content, organization, and accuracy), encouraged creativity and critical thinking, and was praised for its usability and educational usefulness. Students also emphasized the importance of teacher-guided integration and how it can improve confidence and reduce fear. ChatGPT was viewed as a useful and motivating tool for writing research reports.

Overall, the thematic analysis demonstrates that students' experiences of using ChatGPT in research report writing are multifaceted. On a practical level, ChatGPT directly enhanced the technical aspects of their writing, such as content, structure, grammar, and spelling. On an affective level, it fostered positive perceptions, increased confidence, and promoted motivation. However, the analysis also highlights the need for teacher-mediated integration to ensure students use the tool critically and effectively.

5. Discussion

The quantitative results unequivocally demonstrate that ChatGPT-mediated instruction led to a major leap in writing performance, especially the analytical depth and overall coherence of the reports [41], [42], [30]. According to qualitative findings, students perceived ChatGPT as an accessible, time-saving tool that boosted their confidence, improved grammar, vocabulary, and report structure, and facilitated research by providing relevant content and feedback, though its effectiveness was maximized with teacher guidance [43], [44], [45]. Nevertheless, concerns about possible plagiarism due to paraphrased information that was not cited, and periodic errors highlighted the necessity of ethical use and instructor supervision to strike a balance between AI assistance and autonomous skill development [11], [33], [35]. These results are strengthened by the triangulation of quantitative and qualitative findings, which is backed by consistent research. This validates ChatGPT's usefulness as an EFL writing instruction. These results suggest that to guarantee long-term skill development in EFL environments, instructors should utilize ChatGPT with explicit rules to optimize its advantages while encouraging critical thinking and academic integrity.

The quantitative findings, when combined with qualitative findings and preceding literature, give strong evidence that ChatGPT had a transformative impact on EFL students' research report writing. The quantitative results showed that the experimental group that received ChatGPT-mediated instruction outperformed the control group in all five domains of writing—content and relevance, organization and structure, research and evidence, analysis and critical thinking, and writing quality and presentation—with highly significant differences and large effect sizes. It demonstrates that ChatGPT support was a significant contributor to better writing performance.

These statistical findings were reinforced by qualitative interview data, in which students constantly described ChatGPT as a useful learning companion that helped them with numerous aspects of writing. Students specifically indicated that it gave them quick access to relevant knowledge and research ideas, assisted them in logically sequencing sentences and organizing reports, and increased their creativity and analytical thinking by considering alternative viewpoints or argument structures. Furthermore, students emphasized linguistic benefits, such as improved grammar, vocabulary, spelling, and general writing mechanics, which increased their confidence in completing writing tasks.

They also commended its usability, claiming it is simple to use, time-saving, and easy to incorporate into their writing process. However, several of them pointed out that its efficacy was enhanced when combined with teacher assistance to promote critical thinking and prevent blind acceptance of AI recommendations.

These results are generally in line with earlier empirical data when compared to the body of current literature. Like the improvements in students at Najran University, studies by Wei et al. [42] and Andargie et al. [41] proved that students exposed to ChatGPT or other AI-supported environments made significant progress in coherence, lexical richness, and organized writing. In line with the stated qualitative themes of enhanced sentence sequencing and paragraph cohesion, Chen and Xu [30] also emphasized ChatGPT's capacity to promote logical flow and organizing abilities. Like this study's findings of increased analytical depth, Darwin et al. [17] revealed that AI tools boosted students' critical thinking by encouraging them to assess the reliability of sources and organize arguments more convincingly.

The mechanical and grammatical enhancements seen here are consistent with the results of Behforouz and Al Ghaithi [44] and Alkamel [43], who recognized the value of ChatGPT in assisting with proofreading, grammar, and punctuation. However, the qualitative data gathered from this study and a wider body of literature also highlight significant inadequacies. Concerns with plagiarism, excessive dependence, and random factual errors are like those expressed by [33], [11], and [35], who emphasized that ChatGPT outputs are occasionally quoted from uncited sources or susceptible to "hallucinations." Although some students acknowledged that they have been inclined to accept ChatGPT's recommendations without question, the study's training sessions and teacher supervision helped to reduce these risks and supported more general academic recommendations that AI be integrated with clear instructions on critical evaluation, citation, and ethical use.

Together, the triangulated evidence shows that ChatGPT is more than just an additional tool; it is a catalyst for substantial advancements in EFL academic writing, especially when it comes to research report assignments that require linguistic accuracy, coherence, and evidence integration. While the constant warning regarding ethical and pedagogical protections emphasizes the necessity for balanced use, the convergence of strong quantitative improvements, encouraging qualitative feedback, and confirmatory findings from the literature indicates the dependability of these outcomes. Thus, ChatGPT becomes a dynamic teaching tool whose full potential can only be achieved in conjunction with ethical awareness, independent critical thinking, and instructor support.

According to the triangulation of findings, students' opinions and experiences with ChatGPT were generally favorable, supporting the quantitative evidence of its effectiveness and aligning with a substantial portion of the literature on writing with AI. According to the qualitative interviews, students frequently highlighted how ChatGPT helped them get beyond recurring obstacles when writing research reports by offering ideas for pertinent information, assisting with sentence structure, boosting creativity and critical thinking, and improving organization. Additionally, they reported observable language improvements in spelling, grammar, vocabulary, and general writing mechanics, all of which helped to produce reports that were more polished and freer of errors.

Importantly, students described ChatGPT as accessible, time-saving, and motivating, stating that it increased their confidence in engaging in academic writing tasks, particularly when used under teacher supervision to guide responsible use. These experiences supplement the quantitative results, which show that the experimental group outperformed the control group in all five writing domains, implying that the subjective improvements students described were also objectively measurable in their performance outcomes.

The results are highly consistent with previous studies when compared to the body of existing literature. In line with students' self-reported increases in linguistic accuracy, Behforouz and Al Ghaithi [44] and Alkamel [43] discovered that ChatGPT significantly enhanced grammar knowledge, proofreading, and mechanics. In a similar vein, Adeshola and Adepoju [45] and Amin [46] emphasized ChatGPT's impact on vocabulary development, confidence, and engagement, which aligns with the students' focus on skill development and motivation in this study. In keeping with students' findings that ChatGPT improved creativity and analysis in report writing, Darwin et al. [17] also noted that ChatGPT encouraged critical thinking and argument refinement. Nonetheless, the interviews and the literature both emphasize serious viewpoints.

In line with Kohnke et al. [11] and Perkins [13], who cautioned that ChatGPT could encourage plagiarism or diminish independent problem-solving if used without supervision, a small percentage of students voiced concerns about over-reliance and their tendency to accept AI-generated suggestions without question. Moreover, contrary to the overwhelmingly positive opinions expressed by most students in this study, critiques in the literature [47], [35] highlight the dangers of "hallucinated" information and the possible impediment of developing deeper critical thinking abilities.

While the quantitative results showed corresponding, tangible gains in writing performance, the triangulated evidence overall demonstrates that ChatGPT was viewed as a useful and inspiring writing assistant with obvious advantages for grammar, organization, research support, and confidence. However, the literature and student reflections warn that ChatGPT raises the possibility of academic dishonesty, uncritical dependence, and occasional errors in the absence of explicit teacher guidance. This shows that even though ChatGPT can significantly improve the writing experiences of EFL students, its incorporation needs to be supported by moral instruction, critical assessment procedures, and a fair focus on encouraging independence to ensure long-term skill development.

The significant improvements observed in the experimental group, particularly in organization and structure, writing quality, and presentation, highlight the GPT-4 model's core strength: generating well-formed, grammatically correct text based on patterns in its training data. This suggests its architecture is highly effective for tasks requiring syntactic and structural coherence.

Conversely, the model's relative weakness in fostering deeper analysis and critical thinking—a domain where human guidance was most crucial—points to a fundamental technical limitation. While LLMs can summarize and synthesize existing information, their transformer architecture, based on statistical prediction, does not genuinely 'understand' or reason about content. This can lead to 'hallucinations' or superficial analysis, as noted in student concerns about factual accuracy.

Bender et al. [32] draw attention to the ethical, social, and cultural complexities surrounding LLMs. Despite their capacity to produce fluent, human-like text, these models operate without genuine understanding or empathy. They frequently mirror cultural biases embedded in their training data, pose environmental concerns due to their intensive computational requirements, and risk fostering user over-reliance by conflating linguistic polish with factual accuracy or moral soundness. Furthermore, the risk of plagiarism underscores the model's operation as a 'stochastic parrot' that recombines training data without inherent attribution.

These findings have direct implications for the applied development of educational AI. For LLMs to be more effective in advanced academic contexts, future iterations may require specialized fine-tuning on high-quality, academic corpora and the integration of fact-checking modules and attribution mechanisms. Our study demonstrates that while current LLMs are powerful tools for scaffolding the mechanics of writing, achieving depth and originality still necessitates a hybrid approach that leverages human critical thinking to guide and validate the AI's output.

However, it is important to recognize that this study was conducted at Najran University and involved a relatively homogeneous cohort of Preparatory Year EFL students. This institutional and demographic specificity inevitably constrains the generalizability of the findings. To assess the broader applicability of AI-integrated writing interventions, future research should explore more diverse educational contexts—such as institutions with varying academic traditions, language proficiency levels, and disciplinary orientations. Cross-institutional and cross-cultural comparative studies would offer valuable insights into how factors like linguistic background, educational culture, and digital literacy shape the integration and pedagogical impact of ChatGPT. Addressing these contextual variables will be essential for developing a more nuanced and globally relevant understanding of AI-mediated instruction in EFL settings.

6. Conclusion

The purpose of this study was to determine how ChatGPT improves the research report drafting skills of EFL students at Najran University. The experimental group that received ChatGPT-mediated instruction performed significantly better than the control group in all five writing domains: content and relevance, organization and structure, research and evidence, analysis and critical thinking, and writing quality and presentation, according to the results of the quasi-experimental design. The robustness of these gains was validated by large effect sizes. The quantitative results were supplemented by qualitative insights, as students reported that ChatGPT helped them create pertinent content, organize reports logically, enhance their vocabulary and grammar, and increase their confidence when writing assignments. These results were further supported by triangulation with the existing body of literature aligned with studies that emphasize ChatGPT's beneficial effects on linguistic accuracy, critical thinking, and organizational skills while also raising issues of factual errors, plagiarism risks, and students' over-reliance on AI-generated text.

The study has some limitations in addition to its encouraging findings. First, the applicability of findings to more extensive EFL contexts is limited because the sample was selected from a single institution (Najran University) and concentrated on a comparatively homogeneous group of Preparatory Year students. Secondly, the 14-week intervention period may not have adequately captured the long-term impacts of integrating ChatGPT on sustained writing development. Third, the research focused on a single AI tool, ChatGPT, which is popular but only one aspect of AI-assisted learning tools. Fourth, although the interviews yielded valuable insights, they may have been influenced by social desirability bias, with students potentially emphasizing positive experiences over negative ones. Finally, limitations inherent to the AI tool itself must be acknowledged. ChatGPT and similar LLMs are prone to generating "hallucinations"—factually inaccurate or misleading content, which can compromise the reliability of student outputs [32], [11]. These challenges underscore the importance of instructor oversight and the critical evaluation of AI-generated responses to safeguard academic accuracy and integrity.

The results support the importance of Computer-Assisted Language Learning (CALL) theory by showing that AI-powered writing assistants can improve real input, provide instant feedback, and encourage students to take charge of their own academic writing tasks. ChatGPT helps students write, edit, and think about their writing, which is in line with CALL's ideas of authentic tasks and contextualized practice. The study suggests that ChatGPT can be a valuable tool in EFL instruction, aiding brainstorming, model structures, and grammar-focused feedback. However, it emphasizes the need for ethical use, supervision, and AI literacy training to prevent plagiarism and promote critical thinking.

This study demonstrates that the application of an advanced LLM like ChatGPT can significantly enhance the precision of a specific language task. By framing our research within the context of applied computational linguistics, we move beyond viewing AI as a mere black-box tool and begin to critically assess its architectural capabilities and constraints. We recommend that future work in this area actively collaborate between linguists, educators, and computer scientists to develop more refined, domain-specific models and to create optimized prompt libraries that can better leverage existing model architectures for educational purposes. Furthermore, these insights extend beyond EFL writing and hold relevance across a range of educational and computational contexts. Comparable strategies could be applied using other AI tools, such as Bard, Gemini, or domain-specific language models, to support tasks in STEM education, professional development, and creative writing. A nuanced understanding of each model's capabilities and limitations, paired with structured human oversight, can inform best practices for AI integration. Such practices have the potential to improve learning efficiency and task precision across diverse disciplinary settings. This interdisciplinary approach is essential for harnessing the full potential of basic AI research in solving complex, applied problems in human communication.

Future research should explore the long-term impact of ChatGPT on EFL writing development, including whether improvements persist even after AI support is removed, and compare different AI tools or platforms.

Future research could adopt longitudinal designs to track students' writing development across multiple semesters or implement cross-cultural comparisons to assess the efficacy of ChatGPT-mediated instruction in diverse EFL contexts. Employing mixed-methods approaches, combining quantitative writing assessments with qualitative interviews or think-aloud protocols, would yield richer insights into both skill acquisition and learner experience. Investigating technical improvements to large language models is also essential, including fine-tuning for academic writing, enhancing pragmatic and cultural competence, and integrating fact-checking modules to reduce hallucinations. Additionally, systematic analysis of student drafts could help document instances of plagiarism or over-reliance on AI, such as direct copying, minimal revisions, or repeated patterns of uncritical acceptance. These granular evaluations would inform the development of pedagogical guidelines that promote ethical and effective integration of AI tools in educational practice. Diverse samples, longitudinal interviews, and mixed-methods studies could enhance further generalizability.

Acknowledgment

The authors are thankful to the Deanship of Graduate Studies and Scientific Research at Najran University for funding this work under the Growth Funding Program grant code (NU/GP/SEHRC/13/182-4).

References

[1] Pratama, Y. D. (2020). The investigation of using Grammarly as an online grammar checker in the process of writing. English Ideas: Journal of English Language Education, 1(1), 46–54.

[2] Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. Frontiers in psychology, 14, 1260843.

[3] Grabe, W. (2001). Notes toward a theory of second language writing. In T. Silva & P. K. Matsuda (Eds.), On second language writing (pp. 39-57). Lawrence Erlbaum Associates.

[4] Matsuda, P. K. (2019). The myth of linguistic homogeneity in U.S. college composition. In Landmark essays on ESL writing (pp. 65-84). Routledge.

[5] Campbell, C. (2019). Teaching second language writing: Interacting with text. Routledge.

[6] Mohsen, M. A. (2022). Computer-mediated corrective feedback to improve L2 writing skills: A meta-analysis. Journal of Educational Computing Research, 60(5), 1253-1276.

[7] Han, J., Yoo, H., Kim, Y., Myung, J., Kim, M., Lim, H., Kim, J., Lee, T. Y., Hong, H., & Ahn, S. (2023). RECIPE: How to integrate ChatGPT into EFL writing education. In Proceedings of the 10th ACM Conference on Learning @ Scale (pp. 416–420). ACM. https://doi.org/10.1145/3573051.3596200

[8] Hyland, K. (2009). Academic discourse: English in a global context. Continuum.

[9] Nation, P. (2007). The four strands. International Journal of Innovation in Language Learning and Teaching, 1(1), 2–13. https://doi.org/10.2167/illt039.0

[10] Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. The International Journal of Management Education, 21(2), 100790. https://doi.org/10.1016/j.ijme.2023.100790

[11] Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. RELC Journal, 54(2), 537–550. https://doi.org/10.1177/00336882231162868

[12] Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. Education and Information Technologies. https://doi.org/10.1007/s10639-023-11742-4

[13] Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching and Learning Practice, 20(2), 07. https://doi.org/10.53761/1.20.02.07

[14] Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. Journal of Applied Learning and Teaching, 6(1), 1–9. https://doi.org/10.37074/jalt.2023.6.1.17

[15] Meniado, J. C. (2023). The impact of ChatGPT on English language teaching, learning, and assessment: A rapid review of literature. Arab World English Journal, 14(4), 3–18. https://doi.org/10.24093/awej/vol14no4.1

[16] Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. Contemporary Educational Technology, 15(4), ep464. https://doi.org/10.30935/cedtech/13605

[17] Darwin, Rusdin, D., Mukminatien, N., Suryati, N., Laksmi, E. D., & Marzuki. (2024). Critical thinking in the AI era: An exploration of EFL students' perceptions, benefits, and limitations. Cogent Education, 11(1), 2290342.

[18] Vargas-Murillo, A. R., de la Asuncion, I. N. M., & de Jesús Guevara Soto, F. (2023). Challenges and opportunities of AI-assisted learning: A systematic literature review on the impact of ChatGPT usage in higher education. International Journal of Learning, Teaching and Educational Research, 22(7), 122–135. https://doi.org/10.26803/ijlter.22.7.7

[19] Chapelle, C., & Chapelle, C. A. (2001). Computer applications in second language acquisition. Cambridge university press.

[20] Xiao, Y., Li, D., & Guo, K. (2025). Generative AI-powered non-player characters in digital storyline-based learning: an innovative approach to EFL writing. Innovation in Language Learning and Teaching, 1-15.

[21] Hampel, R., & Stickler, U. (2005). New skills for new classrooms: Training tutors to teach languages online. Computer Assisted Language Learning, 18(4), 311–326.

[22] Warschauer, M. (1996). Computer-assisted language learning: An introduction. In S. Fotos (Ed.), Multimedia language teaching (pp. 3-20). Logos International.

[23] Cheong, W., & Hong, H. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research. Journal of Educational Technology and Innovation, 37–45.

[24] Domenech, J. (2023). ChatGPT in the classroom: Friend or foe? In Proceedings of the 9th International Conference on Higher Education Advances (HEAd'23) (pp. 339–347). Universitat Politècnica de València. https://doi.org/10.4995/HEAd23.2023.16179

[25] Chen, H. (2024). Educational applications of ChatGPT: Ethical challenges and countermeasures. English Language Teaching and Linguistics Studies, 6(3), 100–110. https://doi.org/10.22158/eltls.v6n3p100

[26] Ulla, M. B., Perales, W. F., & Tarrayo, V. N. (2021). Integrating Internet-based applications in English language teaching: Teacher practices in a Thai university. Issues in Educational Research, 31(1), 283–301.

[27] Ulla, M. B., & Perales, W. F. (2021). Facebook as an integrated online learning support application for English language teaching: A case study in a Thai university. THAITESOL Journal, 34(2), 1–24.

[28] Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. Computers and Education: Artificial Intelligence, 2, 100017. https://doi.org/10.1016/j.caeai.2021.100017

[29] Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. Applied Sciences, 13(9), 5783. https://doi.org/10.3390/app13095783

[30] Chen, X., & Xu, L. (2024, December). Effectiveness of ChatGPT in education: a meta-analysis. In 2024 5th International Conference on Information Science and Education (ICISE-IE) (pp. 428-431). IEEE.

[31] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).

[32] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, 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). https://doi.org/10.1145/3442188.3445922

[33] Cooper, K. (2021, March 25). *OpenAI GPT-3: Everything you need to know*. Springboard. https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/

[34] Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing assistant's impact on English language learners. Computers and Education: Artificial Intelligence, 3, 100055. https://doi.org/10.1016/j.caeai.2022.100055

[35] Welborn, A. (2023, March 7). ChatGPT and fake citations. Duke University Libraries Blogs. https://blogs.library.duke.edu/scholcomm/2023/03/07/chatgpt-and-fake-citations/

[36] Özçelik, N., & Yangın Ekşi, G. (2024). Cultivating writing skills: the role of ChatGPT as a learning assistant—a case study. Smart Learning Environments, 11(1), 10.

[37] Abdullayeva, M., & Musayeva, Z. M. (2023). The impact of ChatGPT on students' writing skills: An exploration of AI-assisted writing tools. International Conference on Educational Research and Innovation, 4, 61–66.

[38] Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of artificial intelligence powered digital writing assistant in higher education: Randomized controlled trial. Heliyon, 7(5), e07014.

[39] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

[40] Chui, H. C. (2023). ChatGPT as a tool for developing paraphrasing skills among ESL learners. Journal of Creative Practices in Language Learning and Teaching (CPLT), 11(2), 85-105.

[41] Andargie, A., Amogne, D., & Tefera, E. (2025). Effects of project-based learning on EFL learners' writing performance. PLoS ONE, 20(1), e0317518. https://doi.org/10.1371/journal.pone.0317518

[42] Wei, P., Wang, X., & Dong, H. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. Frontiers in Psychology, 14, 1249991. https://doi.org/10.3389/fpsyg.2023.1249991

[43] Alkamel, M. A. A., & Alwagieh, N. A. S. (2024). Utilizing an adaptable artificial intelligence writing tool (ChatGPT) to enhance academic writing skills among Yemeni university EFL students. Social Sciences & Humanities Open, 10, 101095. https://doi.org/10.1016/j.ssaho.2024.101095

[44] Behforouz, B., & Al Ghaithi, A. (2024). Grammar gains: Transforming EFL learning with ChatGPT. Educational Process: International Journal, 13(4), 25–41. https://doi.org/10.22521/edupij.2024.134.2

[45] Adeshola, I., & Adepoju, A. P. (2024). The opportunities and challenges of ChatGPT in education. Interactive Learning Environments, 32(10), 6159-6172.

[46] Amin, M. Y. M. (2023). AI and ChatGPT in language teaching: Enhancing EFL classroom support and transforming assessment techniques. International Journal of Higher Education Pedagogies, 4(4), 1–15. https://doi.org/10.33422/ijhep.v4i4.602

[47] Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., Heathcote, L., & Sun, M. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. Journal of Applied Learning and Teaching, 6(1), 41–56. https://doi.org/10.37074/jalt.2023.6.1.4

[48] Odeh, R. (2005). Examining the difficulty coefficient in educational assessments. Journal of Educational Research, 98(4), 200–210. https://doi.org/10.1080/00220670509597512