Preprints.org

Article

Not peer-reviewed version

Finding Your Voice: Using Generative AI to Help International Students Improve Writing

Leon Sterling [*] , Chunchun Ye , Haoxuan Ying , Zhe Chen

Posted Date: 20 February 2025

doi: 10.20944/preprints202502.1524.v1

Keywords: Generative Artificial Intelligence; Academic Writing; international Students; LLMs

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2025

doi:10.20944/preprints202502.1524.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Finding Your Voice: Using Generative AI to Help International Students Improve Writing

Leon Sterling [1,*,†], Chunchun Ye [2,†], Haoxuan Ying [2,†] and Zhe Chen [3,†]

[1] Computing and Information Systems, University of Melbourne, Australia
[2] CS Student, University of Melbourne, Australia
[3] Graduate, University of Melbourne, Australia
[*] Correspondence: leonss@unimelb.com.au
[†] These authors contributed equally to this work.

Abstract: Students are faced with a wide range of writing tasks during their studies, including literature reviews, summarising papers and reflective reports. The writing tasks present a challenge for students who are not writing in their native language due to studying overseas. Indeed, students writing in their native language have a distinct advantage in assignments involving writing. The rapid emergence of Generative Artificial Intelligence (Gen-AI) over the past three years has the potential to significantly impact the quality and efficiency of writing of non-native English speakers by providing international students with an opportunity to minimise the language barrier when writing in academia. This paper reports on a series of structured exercises undertaken by the authors to determine how using Gen-AI tools built on large language models (LLMs) such as ChatGPT and Claude might improve their writing in the context of computing degrees. We analyse some issues to be aware of when using Gen-AI tools and make suggestions as to their effective use. The key underlying message is that students need to develop their own distinct voice.

Keywords: generative artificial intelligence; academic writing; international students; LLMs

---

1. Introduction

Gen-AI tools built on LLMs, pioneered by ChatGPT, have rapidly emerged over the last couple of years [11]. The tools present both opportunities and challenges for the entire academic community, including students and teachers. For students, the main opportunities are improving their ability to write reports of high quality, and becoming more efficient when undertaking writing tasks. The challenge is to ensure that students develop their writing capabilities and not just rely on a tool. Students need to ensure that the writing reflects their own voices. A secondary challenge is to use Gen-AI resources ethically.

For teachers, the challenge is to develop new ways of teaching while preserving academic integrity. Teachers need to be aware of the capabilities of Gen-AI tools and learn to set assignments that develop capabilities of students. There have been some initial attempts to point out when writing is from a Gen-AI tool, such as [2], but no definitive method exists for determining whether or not a student's work submitted for an assignment involved writing by a Gen-AI tool. Teachers need to be able to advise students on how to use the Gen-AI tools effectively. All stakeholders including students, teachers and educational institutions need to ensure that the use of Gen-AI tools is transparent and governed by clear and reasonable policies. In general, academics should teach communication skills as well as discipline-specific skills. Students need to learn research methods, and writing.

If used well, Gen-AI tools can help increase the quality of students' work, and improve the efficiency of assessments. Students will be able to complete more complex tasks in shorter time. If used poorly, Gen-AI tools can lead to an explosion of mediocrity and loss of skills for students [1]. Many papers have discussed the strengths and limitations of ChatGPT and other Gen-AI tools. Examples are

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

[2,10,11]. In this paper, we advocate learning how to use the systems to improve student performance and learning.

The primary objective of this paper is to help international students with academic writing, though we believe that the guidance on using Gen-AI tools concerning the nuance of expressions should also be valuable for native English speaking students. We use the term *international student* to denote students whose native language is not English. Sometimes such students are called *non-native English speaking students* or *ESL students*, for *English as a Second Language*. For simplicity, we use the term international students throughout the paper.

Writing is a key communication skill needed by students who are embarking on academic study and/or tackling university life. It is also important for professional careers [17]. Speaking is another essential communication skill but discussing it is beyond the scope of this paper. We believe that Gen-AI tools offer valuable assistance to help international students with writing tasks, because they can quickly generate first drafts and help students overcome the intimidating Blank Page Syndrome [12]. The drafts can provide a starting point or a skeleton for their work.

Discussing how to use Gen-AI tools is a moving target as there are continual efforts to improve the quality of output generated by Gen-AI tools. Indeed, Gen-AI tools have improved significantly since we started our research in the second half of 2024. The challenge is to provide strategies for students to use Gen-AI tools for a variety of writing tasks, while maintaining academic standards, ensuring academic integrity, and preserving students' unique voices. The exercises described in this work show potential for improving the use of Gen-AI tools by international students for academic writing.

The landscape of Gen-AI tools is diverse, with many options such as ChatGPT, Claude, Meta AI, etc., each possessing unique strengths and limitations. A list of Gen-AI tools and the corresponding LLMs used in this research is given in an appendix. In our exercises, Claude 3.5 Sonnet stands out for its ability to conduct literature reviews and follow complex instructions due to its strong contextual handling capabilities. Claude 3.5's token limit and context length far exceed that of GPT-4, allowing it to process and output long texts without losing critical information. This became particularly evident during long-text testing, where GPT-4 occasionally struggled to fully read and process extended inputs, whereas Claude 3.5 was able to handle the entire context effectively.

ChatGPT-4 is often preferred for its balanced performance across a wide range of academic tasks, benefiting from its fine-tuned contextual awareness and general-purpose adaptability. During Ian's internship, he experimented with connecting these three models to an Agent for executing complex instruction sets. The results showed that GPT-4 outperformed the other models in terms of precision in executing specific instructions, making it the most suitable choice for tasks requiring exact follow-through.

Meanwhile, Llama 3, though newer, is optimised for token efficiency, allowing it to handle longer-form writing tasks with less computational overhead. According to a study by Novita AI, Llama 3 70B can be up to 50 times cheaper and 10 times faster than GPT-4 when used through cloud API providers [14]. This makes it particularly advantageous for tasks that require processing large amounts of data or generating long-form content without excessive computational costs. These differences in performance, context-sensitivity, and task specialisation make it more complex for students to choose the right model for their specific writing needs. Effectively utilising each model's strengths while compensating for its limitations further complicates the process of matching the right tool to the task at hand.

Given the variety of Gen-AI tools, a critical question emerges: How can students best leverage these technologies to enhance their writing skills? There is also a financial question as students may need to consider which system is the best value for money. This paper aims to explore effective strategies for international students to utilise current AI technologies in improving their academic writing, while maintaining their unique voice and adhering to academic integrity standards.

This paper is organised as follows. The next section discusses our approach to helping international students find their voice. We then describe the exercises we undertook. Section 4 contains a discussion where we look at previous work on helping international students with writing. We include recommendations for students on how to use Gen-AI tools more effectively. The final section concludes.

2. Method

This paper takes a non-standard methodological approach. The authors conducted a series of exercises using Gen-AI tools for specific writing tasks. The results were discussed and analysed each week over a period of four months. Initial meetings were weekly but became more sporadic after the teaching semester finished. Specific insights are shared in the Results section. More discussion of the validity of our approach is given in the ***Discussions*** section.

The rationale for why students may benefit from using Gen-AI tools is as follows. International students in academic settings often face significant challenges (especially in project-based subjects) in both comprehending and expressing ideas that are complex [15]. Before the advent of Generative AI models like ChatGPT, international students relied on tools such as Google Translate and DeepL to help them navigate through complex academic content. Because people feel more comfortable with processing information in their native languages, a common approach is that reading materials or assignment specifications are translated into their native language, and all the thinking and written work are performed in the native language too. The finished work will be translated back into English for submission. However, traditional translation tools have limitations, such as they typically perform mechanical translations solely based on dictionary definitions, failing to account for abstract concepts or contextual nuances [13]. This may result in students struggling not only to understand assignment requirements but also to articulate their own ideas clearly and accurately.

With Generative AI tools, international students now have access to a more dynamic and efficient way of engaging with academic content. Instead of relying solely on traditional translation tools, students can use generative AI to first summarise complex texts, helping them grasp the core ideas more quickly. Through interactive conversations with an AI, students can have confusions clarified, such as contextual ambiguities or highly specialised academic terminologies and vocabularies, which can be particularly challenging to international students. As a result, students can have an opportunity to save time on Googling and asking questions on forums unfriendly to beginners, resulting in a more in-depth understanding in academic materials. In this way, the Gen-AI tool acts not only as a means for summarising information but also as a sophisticated search engine, capable of offering tailored explanations and context-driven insights.

Unlike traditional search engines, Generative AI models offer ongoing, context-aware conversations, retaining the context of previous interactions to create a more coherent and iterative learning experience. This is particularly valuable when students encounter complex, technical topics, as the AI can provide detailed explanations and simplify difficult concepts. Essentially, generative AI functions as a personalised, interactive learning assistant, helping students bridge the gap between understanding abstract academic material and expressing their ideas effectively. This not only significantly reduces the time spent searching for answers online where relevant information may not always be immediately available or may require waiting for others to respond. The tools also deliver customised, real-time solutions to questions.

The rationale for the project is rooted in the academic experience of the first author. He has read numerous reports over the past thirty years from international students. It was a struggle to understand what was being said. More importantly from a teaching perspective it was difficult to determine whether the student understood the concept but lacked the ability to express it in English, or misunderstood the concept. This is known from many studies on teaching English to international students over many years, [5,8,9,15].

As grammar checkers such as Grammarly have improved, some of the English expression difficulties can be mitigated. Students should be expected to write effectively. Indeed good communication skills are a requirement for most professional jobs.

The emergence of Gen-AI tools has changed the experience of creating and assessing reports. ChatGPT was the first Gen-AI tool used extensively for writing reports. Other tools such as Claude and Llama 3 quickly followed that could also be used for writing tasks. It makes sense for international students to use such tools to write reports. However, it creates a new problem for teachers to navigate, namely whether the student has independently done the work.

In an undergraduate capstone project subject [1] taught in 2023, the first author set an exercise where students needed to reflect on their project experience. They were explicitly allowed to use Gen-AI tools as long as they were transparent about their use. He read over 300 reflection reports, most of which were from non-native English speakers. Reading the reports was insightful and the experience directly influenced this paper.

Overall the reports written with the assistance of Gen-AI tools were easier and quicker to read. However, there were two major problems that are connected. One is that the reports largely sounded the same. It was hard to distil students' unique experience, a phenomenon being increasingly understood. Prakhar Mehrotra's article *ChatGPT and the Magnet of Mediocrity* [1] discusses how AI-generated content, such as by ChatGPT, tends to be derivative and mediocre. Mehrotra illustrates this by stating that while ChatGPT can generate *"decent"* blog posts or articles, it lacks the originality and creative flair that human writers can bring to their work, such as in the creation of new narrative structures or interesting ideas. The second problem concerns the authenticity of the experience being reported. There were at least ten reports that compared the experience of leading a software team to being a conductor of an orchestra. The gist of the comparison was that it was necessary to get diverse instruments/people to work together and coordinate them effectively. Clearly ten people did not come up with the analogy independently. Furthermore, why was the analogy even appropriate in the first place? Did the students who used this analogy have the experience of being in an orchestra at all? Did the reader have or can relate to such an experience? On reflection it was not an appropriate analogy. It did not reflect the student's voice. There were also several analogies comparing leading a software team to being the captain of a boat on the seas. The same issue of authenticity exists.

In reading the reflections, the first author wanted to hear the students' voices which were largely getting lost in the use of Gen-AI tools. He proposed a research project topic on how students can find their voice when writing reflective reports using Gen-AI tools. The second and third authors volunteered to undertake the project and conducted the initial exercises, the results of which are described here. The fourth author joined the project for later stages of analysis.

The second, third and fourth authors are international IT students studying computing at the University of Melbourne who desire to improve both their writing skills and the use of Gen-AI tools. Communication can be an issue for international students [18]. While the focus is on exercises which are meaningful for IT students due to the experience and expertise of the authors, we believe that the insights shared are more broadly applicable.

As the project began, we undertook weekly writing exercises. An exact list of the exercises is available from the authors. Weekly meetings were very efficient and beneficial. The discussion of the outcomes were insightful for all the authors. These insights are shared here. There is scope for a systematic study, but in our opinion it is valuable to report the results, as the tools themselves are evolving. Even if a Gen-AI tool has been tinkered with to improve outputs, we believe our observations and reflections may provide insight to the readers.

The results in the next section are presented as a series of writing exercises followed by insights gained by the authors while discussing what was produced in the writing exercise.

---

[1] IT Project, University of Melbourne. Available at: https://handbook.unimelb.edu.au/2023/subjects/comp30022

3. Results

3.1. Exercise 1: Writing Personal Reflections with AI Assistance

The first exercise we undertook set the scene. A simplified reflective report was proposed. Students were asked to write 100 words about what they had learned during their undergraduate studies. To make the exercise more systematic it was expanded. The participants were asked to take three different approaches to write the simplified reflection.

1. Independent writing: Participants independently authored their responses without external assistance.
2. AI-assisted writing: Participants used ChatGPT to polish self-written responses in approach 1.
3. AI-generated content: Participants relied on ChatGPT to generate the responses entirely based on prompts provided.

The exercise was conducted by two of the authors and some of their friends. Twelve students overall undertook the exercise. They were given a week to respond. ChatGPT was used for the exercise due to its high popularity in the student community. It is a mainstream LLM with the most paid monthly active users on the market, as of 2024.

The three versions of a simplified reflective report of each of the participants were discussed at our second weekly meeting. The student authors gave a qualitative reaction to the changes introduced by ChatGPT. The consensus was that the ChatGPT version sounded more "formal" than the initial response. The sense of formality was due to both choice of words and sentence structure. It was helpful to discuss together and was a valuable learning exercise.

Each of the changes that were made by ChatGPT were discussed as to whether they still accurately represented the intention of the author. The consensus was that ChatGPT often struggled to capture subtle semantic nuances, resulting in deviations from the original ideas it aimed to express.

For instance, a response (approach 1) was: *"The first thing that I learned is to be brave"*. After being refined (approach 2) by ChatGPT, it became: *"First, I've learned the importance of bravery"*. The revised version, while grammatically correct, altered the original tone. The phrase *"to be brave"* is more personal and direct, whereas the refined version sounded more abstract and detached, which is less suited for a personal reflection. Interestingly that distinction would not have been initially apparent to the student authors.

Another issue we identified was a tendency towards over-boasting. For example, when students used modest and humble expressions such as *acquired the skill* or *capable of doing*, the Gen-AI tools often replaced these terms with more exaggerated ones such as *comprehensively* or *swiftly extract vital information*. The AI-generated responses would describe outcomes in a more confident and affirmative manner, which often misrepresented the student's original intention. Specific examples include:

- *Know* replaced by *Master*
- *Use* replaced by *Leverage*
- *More* replaced by *Significantly*
- *Are prepared* replaced by *Well prepared*
- *Capable of doing* replaced by *Have a good command of*
- *Have developed* replaced by *Gain a deep understanding of*
- *Have learned* replaced by *Have gained proficiency*

This overconfidence in language use can distort the students' intended meaning and tone, jeopardising students' academic integrity in academic writing, where modesty and accuracy are priorities.

We were able to extract more significant findings from the refined responses (approach 3). Participants may "lose their voice" after responses are refined. Firstly, international students may have limited ability to determine the semantic and contextual appropriateness in word choices in English, especially when thesauruses are suggested by LLMs. For instance, a response mentioned the participant's ability to utilise some academic resources provided by the university to improve the academic performance. In the refined response, the word *utilise* was replaced by *leverage*. It may not be challenging for a native

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2025 doi:10.20944/preprints202502.1524.v1

English speaker to determine the inappropriateness to use *leverage* based on years of experience in using English. However, many international students may fail to realise implications of this word choice. Additionally, the fact that Gen-AI tools tend to write in an overconfident tone and choose more advanced words (regardless of appropriateness) may create an illusion for international students that Gen-AI tools are professional in writing, yielding excessive trust in their expression.

Interestingly, as we were writing the paper, the grammar checker provided by the LaTeX editor we used, Overleaf, became a focus of attention. We would discuss the suggestions and realised that we did not want to automatically accept the changes because the changes could change the tone. It was not solely an issue of correcting grammar.

An interesting case in point was the suggestion to change the phrase *"excel in"* to *"excel at"*. This prompted an interesting discussion and Internet search as to which was correct usage. There was no clear answer, and the international students were exposed to a subtlety of the English language that would not have been apparent if the change was automatically accepted. Indeed, discussions about points of language were consistently insightful and do not come about in standard feedback on assignments.

Another change concerned whether to use *collected responses* or *responses collected.* To explain the subtle difference between these phrases: *"Collected responses"* functions as an adjective phrase where *"collected"* modifies *responses*, suggesting responses that have been gathered or assembled. *"Responses collected"* is a past tense verb phrase indicating the action of collecting responses. It often appears in constructions like *"the number of responses collected"* or *"responses collected during the survey"*. Both are grammatically correct but serve slightly different functions in a sentence. *"Collected responses were analysed"* is appropriate. The responses collected showed interesting trends. The above explanation is edited from Claude 3.5 Sonnet, which then asked *"Would you like me to provide some example sentences to illustrate the difference in usage?"*

Here is another interesting little subtlety. For the following posting *"Definitely something <person1> and <person2> need to refer to."* The grammar checker suggested adding a comma after *"Definitely"*: *"Definitely, something <person1> and <person2> need to refer to"*. The "corrected" version has the tone changed, as *"Definitely"* in the former sentence described the extent of certainty, while "Definitely" in the latter sentence was simply an exclamation. However, it would be hard for an international student not to accept the change as they may not appreciate the subtle distinction.

3.2. Exercise 2: Comparing Gen-AI Tools

Heartened by the discussions stemming from analysing the responses to the first exercise, we were encouraged to undertake more exercises. One of the respondents in the first exercise had used a different Gen-AI tool. The modifications were different. We were all curious as to how effective the various Gen-AI tools were. The detailed information about the LLMs we used is provided in Appendix A1.

Based on the responses of the survey in exercise 1, the second exercise sought to explore qualities (including credibility, semantic consistency and answer integrity) of generated texts from various Gen-AI tools. The question for which a response was requested was: "Why do I need to understand and learn artificial intelligence, and why is artificial intelligence important?" Participants in the exercise were asked to write a 200-word response to this question.

We experimented with multiple Gen-AI tools. We received responses from the question and then refined the responses using three different LLMs: ChatGPT-4, Claude 3.5 Sonnet, and Llama 3 with the same prompt: *"refine it"*. A complete set of the responses are available on request.

By comparing the generated results from different LLMs with same prompt, we found that performance and output quality could be significantly affected by technical limitations, such as token limits and the maximum context length, also referred to as the context window. For instance, ChatGPT-4's maximum token limit is 4096 tokens, while Claude 3.5's is 8132 tokens and Llama 3 even supports 128k tokens. The token limit difference can significantly impact the performance when dealing with larger contexts. ChatGPT-4 tended to omit instructions or generate truncated responses when longer prompts are provided, whereas Claude 3.5 and Llama 3 128k handled the same prompts better because of their larger token capacity.

During the discussion, the first author shared a response he had received from ChatGPT to the following prompt: *"Write a 100-word description of the fire at my father's paint factory when I was 9."*

The response was: *"When you were nine, a blazing inferno engulfed your father's paint factory, painting the sky with hues of fiery orange and billowing smoke. Flames danced voraciously, consuming the structure, while the acrid scent of burning chemicals filled the air. Panic gripped the scene as firefighters battled tirelessly, their efforts highlighted by flickering red and blue lights. Amidst the chaos, memories of your father's hard work and dedication mingled with fear for his safety. Despite the devastation, the community rallied, offering support and solidarity. The aftermath left an indelible mark, a testament to resilience and the fragility of human endeavour."*

The report is well written, but it is not an account of the fire. How could it be, as it had no details about the actual fire. ChatGPT generated an essay about a (fictional) memory, rather than describing a memory. They are not the same thing.

Interestingly, we entered the same prompt to Claude and received a more suitable response. Claude responded that it could not give a memory about the fire. However, it could give a 100-word description about a fire, which was a much better and useful response.

The discussions surrounding this exercise suggest that when students refine their prompts using Gen-AI tools, they should carefully consider the trade-offs between tool capacity and task requirements, especially when engaging in long-form or detailed writing. This led to the next exercise.

3.3. Exercise 3: Prompt Tuning Insights

In the third exercise, we built on the results of the second exercise and conducted further experiments. This time, we compared different prompt structures and tested multiple Gen-AI tools to assess how well they could handle personal writing tasks. Our motivation stemmed from observing in previous exercises that the level of detail in prompts influenced the balance between creativity and standardisation in AI-generated responses. The Gen-AI tools used in this exercise remained the same, but the prompts ranged from simple instructions to more detailed ones that included personal background, writing context, and even bullet points. The three prompt structures we used were:

1. Minimal context: The original question plus *"refine it."*
2. Moderately detailed context: The original question plus *"refine it and check grammar while keeping the same tone."*
3. Highly detailed context: The original question plus *"refine it as a paragraph, academic writing, as an undergraduate student, major in CS, University of Melbourne, 200-words, answering 'Why do I need to understand and learn artificial intelligence, and why is artificial intelligence important?'"*

We conducted more tests using multiple LLMs (ChatGPT-4o, Claude 3.5 Sonnet, Llama 3 70B) and tested each of the three prompts under controlled conditions. There were ten participants in this exercise, largely consisting of students majoring in computer science, ensuring a consistent background to allow for meaningful comparisons. For each prompt, the AI-generated responses were collected, grouped by prompt type, and then analysed for:

- **Relevance:** The degree to which the response addresses the prompt.
- **Diversity:** The linguistic, structural, and content variation between responses.
- **Tone Preservation:** The degree to which the response retains the distinctive tone and style implied by the original question.

By comparing the responses generated from different prompt styles and Gen-AI tools, we identified several key issues:

Our findings indicate that while detailed prompts can improve relevance and specificity, they also increase the risk of generating uniform and predictable responses. Comprehensively, the more detailed the prompt, the more uniform and standardised the responses became. For example, when the prompt explicitly mentioned the student's background as *"a CS student at the University of Melbourne"*, ChatGPT-4o tended to produce highly similar responses across different participants.

Here are two notable sets of similar generated paragraphs:

Set 1 **paragraph 1** As an undergraduate student majoring in Computer Science at the University of Melbourne, understanding and learning artificial intelligence (AI) is crucial for both my academic and professional development.

paragraph 2 As an undergraduate student majoring in Computer Science at the University of Melbourne, I recognise the critical importance of understanding and learning artificial intelligence (AI).

Set 2 **paragraph 3** Understanding and learning artificial intelligence (AI) is crucial for several reasons, particularly from the perspective of an undergraduate student majoring in Computer Science at the University of Melbourne.

paragraph 4 Understanding and learning artificial intelligence (AI) is essential for me as an undergraduate student majoring in Computer Science at the University of Melbourne, as it aligns with both my personal interests and future career aspirations.

In the two pairs of examples above, it is obvious that when given enough detailed information as a prompt, ChatGPT-4o is more likely to translate or paraphrase the prompt than to totally refine it based on the information provided by the prompt. Thus, with the same prompt, ChatGPT-4o will likely produce very similar generated content.

Similarly, if detailed bullet points were provided in the prompt, the AI would incorporate all of them into the response, often without much variation. This raises concerns in contexts like a classroom where many students may use similar prompts.

In such cases, responses could become repetitive. For instance, if most students identify as Computer Science students, ChatGPT might consistently generate phrases such as "version control systems" or other highly specific technical terms, which would lead to homogeneity in responses. This is problematic, as the goal of academic writing is to encourage original thoughts and varied perspectives.

We found that simpler prompts such as *"refine it"* or *"check its grammar and keep the same tone."* were preferred when students were writing personal reflections. These instructions allowed them to maintain control over their writing while enhancing their English fluency without losing their voice.

3.4. Exercise 4: Summarisation Ability of LLMs

We used ChatGPT-4o and Claude 3.5 Sonnet to test Gen-AI tools' performance on different scenarios where international students face challenges. We observed that Gen-AI tools perform better on summarising content in very specific contexts, rather than generating content for open-ended questions in general contexts.

To begin with, we simulated a scenario that can challenge many international students – reading and understanding long assignment specifications. We collected many undergraduate-level Computer Science assignment specifications from the University of Melbourne, cleansed the text via removing irrelevant content such as headers and footers, and supplied them to both Gen-AI tools as the context with a prompt *"Summarise it"*. As a result, they successfully extracted the main points from the specifications in a straightforward and accurate manner. Minor errors did exist, such as ChatGPT-4o made an assumption that the total mark of an assignment was 100, while the specifications never explicitly stated so. We then gave a more challenging scenario: *"I am an international student struggling to understand this specs, the deadline is approaching, and I don't even know where to start"*. ChatGPT-4o and Claude 3.5 Sonnet were both capable of accurately identifying the implementation priorities and extracting the grading emphases, and gave proper advice on earning the most essential marks in a time-sensitive manner. The advice was articulated concisely in bullet points, with subheadings such as *"Immediate steps to take"* and *"Technical must-haves"*, minimising the effort required to navigate through

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2025 doi:10.20944/preprints202502.1524.v1

the information. At the end, both Gen-AI tools offered further assistance by asking *"Let me know if you need help with specific coding tasks!"* or *"Would you like specific guidance on where to begin with either the UML design or the code implementation?"*.

We also tested Gen-AI tools' ability to assist with GRE (Graduate Record Examinations) questions, such as questions requiring to extracting main ideas from given reading materials, which are typically several passages. These passages consist of substantial information with an incredible amount of details, and are written in a clear structure. Using such reading materials to assess Gen-AI tools' summarisation capabilities conveniently provides an accurate way to identify and uncover potential issues, such as oversimplification, confusion of subjects, incorrect chronological order, etc. For testing, we randomly selected three long passages from GRE reading comprehension questions which appeared in actual exams. Both ChatGPT-o1 and Claude 3.5 Sonnet showed tendencies towards simple restatements, where the AI merely paraphrased the original text rather than gathering and synthesising information scattered around. Additionally, ChatGPT-o1 occasionally presented information "made out of thin air", generating details that were not present in the original passage. For example, for the following sentence extracted from an actual exam question:

Original: *Recently, an unusually high number of dolphins have been found dead from infectious diseases, and most of these had abnormally high tissue concentrations of certain compounds that, even in low concentrations, reduce dolphins' resistance to infection. The only source of these compounds in the dolphins' environment is boat paint. Therefore, since dolphins rid their bodies of the compounds rapidly once exposure ceases, their mortality rate should decline rapidly if such boat paints are banned.*

ChatGPT-o1's summary: *A recent surge in dolphin deaths from infectious diseases has been linked to elevated concentrations of infection-weakening compounds in their tissues (Made out of thin air: The original article did not mention the recent increase in concentrations of toxic compounds). These chemicals come exclusively from boat paint. Because dolphins quickly eliminate these substances once exposure stops, prohibiting the use of such paints should rapidly lower their death rates.*

Meanwhile, Claude 3.5 Sonnet sometimes failed to provide a comprehensive summary, omitting essential information. Overall, these Gen-AI tools are not yet fully reliable as reading assistants, and we still need to manually verify the accuracy via referring to the original text.

In contrast, both ChatGPT-4o and Claude 3.5 Sonnet performed worse on open-ended questions with broad contexts. For instance, we attempted to seek career advice such as *"What would a career as a data science engineer be like"*. The responses were relatively too general and they simply repeated the information that can be easily found on the Internet. An analogy could be, if the previous *"a struggling student coping with the approaching deadline"* situation was provided, instead of advising on which specific tasks to prioritise and implementation steps to follow, suggestions such as *"calm down and meditate yourself"*, *"make sure that you read the assignment specs thoroughly"* and *"reviewing the lecture content and the learning materials would help"* were given. In the real world, if an open-ended question with a broad context was asked, the person being asked may seek more background information to better understand the intention and provide a more tailored response. However, both ChatGPT-4o and Claude 3.5 Sonnet provided general responses without asking any follow-up questions for clarification.

To verify the integrity and reliability of LLMs, we collected more than 20 assignment specifications from 11 computer science subjects at the University of Melbourne, converted them into the Markdown format with redundant text (such as headers and page numbers) cleansed, summarised these specifications using LLMs, then evaluated and compared the results using metrics such as Semantic Similarity, Bert Score, ROUGE-1 (n-grams), ROUGE-2, and ROUGE-L Scores, adopting and modifying the metrics used by a group of researchers to study [16]. Semantic Similarity measures the semantic closeness between the generated and original text, indicating the extent of meaning preservation. BERT Score evaluates the contextual and semantic alignment using BERT embeddings, providing measurements in precision, recall, and F1 scores. ROUGE-1 and ROUGE-2 calculate the overlap of unigrams and bigrams, reflecting lexical coverage and local coherence. ROUGE-L assesses global structure similarity

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2025 doi:10.20944/preprints202502.1524.v1

based on the LCS (longest common subsequence). These metrics allow for a multidimensional analysis of the summary's quality, combining lexical and semantic perspectives.

We used ChatGPT-4o and Claude 3.5 Sonnet as test objects and employed 4 different prompts to test the summary output of the LLMs:

- *"Summarise this"*
- *"Summarise in detail"*
- *"I am a student and my assignment is due soon, but I don't have time to read everything thoroughly. Please summarise this and extract the key points."*
- *"I am a student and I am struggling with figuring out where to start, while the deadline is approaching. Could you give me some hints on what to do?"*

We used BLEU and ROUGE as metrics to evaluate the summary accuracy and compare different models and prompt variations.

Table 1. Performance Metrics for Claude 3.5 Sonnet and ChatGPT-4o.

Metrics	ChatGPT-4o	Claude 3.5 Sonnet
ROUGE-1 (Precision)	0.8973	0.8453
ROUGE-2 (Precision)	0.2556	0.2765
ROUGE-L (Precision)	0.5000	0.4792
BERTScore_F1	0.5585	0.5519

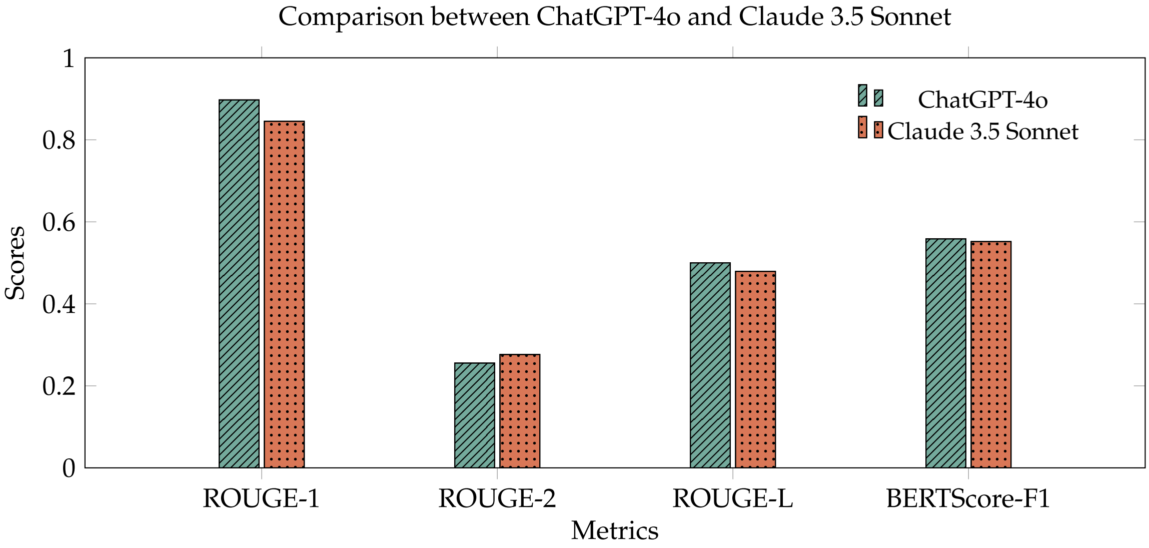

Figure 1. A full-width bar chart comparing ChatGPT-4o and Claude 3.5 Sonnet performance scores on different metrics. ChatGPT-4o bars have diagonal lines, while Claude 3.5 Sonnet bars have dots for better accessibility.

According to the experimental data, ChatGPT-4o and Claude 3.5 Sonnet performed almost equally in summarising, especially in keyword extraction and semantic information retention, (as indicated by high ROUGE-1 and BERTScore metrics), and were capable of general tasks. However, their ability to retain local coherence and global structure, indicated by ROUGE-2 and ROUGE-L scores, needs improvement, especially when dealing with complex logical relationships or highly structured content. In general, LLMs are suitable for quickly generating summaries, but for summarisation tasks that require high precision and complex semantics, we cannot rely solely on LLMs.

To evaluate summarisation ability further, we also conducted an exercise asking ChatGPT-4o and Claude 3.5 Sonnet to summarise an article by the first author [4]. Both LLMs produced coherent summaries. The accuracy of the sumamries and their coherence was a surprise to the first author.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2025 doi:10.20944/preprints202502.1524.v1

4. Discussion

Teaching communication skills is often an afterthought in computing subjects. Yet communication skills are vitally important for success in the workforce. Furthermore, computing academics do not typically seek advice from education experts. Engineering teachers “do it themselves”. The writing exercises were designed from the experience of the first author rather than through consultation with the education literature. Writing this paper is an opportunity to link with other work. An early example of seeing how international students learn to write in English dates from 1987 and is authored by Arndt [5]. There is a good survey of the writing approaches in [6]. The paper [7] discusses particular writing strategies. We have taken a social construction approach with the exercises.

4.1. Language Difficulties for International Students

The discussions about how to use Gen-AI tools have been beneficial for all the authors. In this AI era, the use of AI in the student community is inevitable and unstoppable. While some universities have been tempted to outlaw students’ AI usage, there is little justification because competency in using AI tools will be required when students enter their professional lives. A comparable situation can be, in some countries with an exam-oriented education system, students are strictly prohibited from dating because parents and teachers consider dating a distraction that worsens students’ academic performance. Ironically, the stance on dating suddenly shifts to the other extreme when students enter the universities. Dramatically, rather than being prohibited, students are constantly urged to date and eventually get married. Students, who were strictly forbidden from dating due to high-pressure exam-oriented education, suddenly face social pressures to start dating. They struggle with this abrupt shift and lack dating experience. Therefore, it is essential and beneficial to practise how to use the tools in a learning environment.

Staff need to specify assessment so that the instructions to students about how they can use the Gen-AI tools is clear. There should also be an opportunity to reflect. The weekly discussions where we reviewed the exercises undertaken with the Gen-AI tools were valuable. Discussions about the nuances of words were insightful and open-ended. Providing an opportunity for such open-ended discussions is important for learning. Oftentimes, the demand for assessment rubrics mitigates against such discussions. In other words, the push for precise, measurable assessments to quantify students’ academic performance can overshadow or discourage the kind of free-form, in-depth conversations.

From conducting the studies and our regular conversations, we observed that international students may encounter difficulties in the following four areas:

1. **Keeping a personal voice:** Ensuring that their unique voice and tone are retained while adhering to formal academic writing conventions is often challenging.
2. **Over-reliance on AI:** Non-native English speakers often overly trust AI tools due to lack of experience and comparative knowledge, but they’re more discerning in their mother tongue.
3. **Challenges in literature review:** Effectively finding and synthesising relevant research remains a considerable obstacle for many students, especially when overcoming language barriers.
4. **Difficulty in summarising texts:** International students often spend a great deal of time summarising complex academic material, a situation further exacerbated by the fact that computer science specifications often contain advanced vocabulary or abstract concepts.

There is no doubt that using Gen-AI tools can help non-native English speakers increase the quality of their writing. At the very least, grammar and flow should improve. However, the challenge is to ensure that students learn rather than just accepting the suggestion of a Gen-AI tool. It is important for students to be aware of the need to develop a voice, instead of submitting writing which sounds like “everyone else”.

4.2. Suggestions for Improvement

Advice on the advantages and pitfalls in using Gen-AI tools has begun to appear both in academic journals [10,11,13], and in the popular press [1–3]. A common theme in the articles is that text generated by Gen-AI tools have predictable patterns, an observation that we noted while discussing our exercises. Example papers are [1–3]. Here are three tips that emerged from the writing exercises.

1. **Translate to your native language** To mitigate the understanding difficulties, we propose an alternative approach: after generating English content with AI, students should translate it back into their native language. Since students are more confident in their native language, this method allows them to assess whether the generated English text aligns with their original intent. By translating the content back and forth, students can ensure that their personal tone and meaning are preserved, resulting in a final output that better reflects their ideas.
2. **Experiment with different Gen-AI tools**: As discussed in the results section, different Gen-AI tools produced different results. Discussing the differences led to useful insights. Claude was more honest that it couldn't make up history.
3. **Experiment with different prompts and summarising ability**:
   In a writing task, a structured workflow can be adopted. First, the entire task is broken down into multiple stages, each focusing on a specific subset of literature. Then, a Gen-AI tool can be used to generate a concise summary of each document. RAG technology can also help international students quickly access and summarise relevant academic resources. It uses retrieval technology to extract information from existing research, making it easier to conduct a comprehensive review without extensive manual searching.

To make AI-assisted content more personalised, students can adopt several strategies. Incorporating personal experiences, describing research motivations, and incorporating one's own opinions and comments when discussing literature are all effective.

5. Conclusions

Gen-AI tools can provide valuable assistance to international students in their academic writing. In order for students to use the tools effectively, they need to understand the changes that the tools make to their original thoughts and drafts, and how best to provide prompts.

The research was undertaken to help improve English writing skills. For the students involved, it was an opportunity to improve their writing and make better use of Gen-AI tools. For the teacher involved, it was to learn how to better allow the use of Gen-AI tools while still improving the writing of students. At the end of the exercise, the importance of improving writing skills has only become more apparent.

By streamlining the literature review process and improving writing structure, these tools enable students to focus more on the content of their writing submissions and less on mechanical aspects.

There is potential to further explore the role of AI in critical analysis, research question generation, and enhancing personal writing style. These tools could evolve to meet the specific needs of international students, providing more personalised assistance in the future.

Author Contributions: All authors contributed equally to this research

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: We used Google Drive and OneDrive to store our research data. We will not disclose detailed data for the privacy of the volunteers who participated in the research questionnaire. For the data that can be disclosed, this article has included it in the table.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Supplementary Data

Appendix A.1. Model Details

The following shows the specific parameters of the models we used throughout the research process, so as to compare the advantages and disadvantages of different models.

Table A1. Details of the large language model involved in this paper.

Model Name	Parameter Quantity	Context Length	Time of Knowledge Cutoff	Token Limit
ChatGPT-4o	200b	128k	Oct 2023	4096
ChatGPT-o1	200b	200k	Oct 2023	4096
Claude 3.5 Sonnet	175b	200k	April 2024	8192
Llama 3-70B	70b	128k	Dec 2023	4096

References

1. Mehrotra, P. Chat-GPT and the Magnet of Mediocrity. The Edge, March 2023
2. Da Money Hacker, 11 Signs That AI Wrote It, https://medium.com/the-writers-pub/15-signs-that-ai-wrote-it-9bc37e165973, Medium magazine, Dec. 2024
3. Ghate, A. The AI panic is real — More so for a writer, Medium.com, Dec. 17, 2024
4. Sterling, L. The effect of Chat-GPT on Relationshios, in Interpersonal Relationships in the Digital Age (ed. Ferreira, J.), Intech Open, 2025.
5. Arndt, V. Six writers in search of texts, Elt journal, Vol, 41, pp. 257-267, 1989
6. Mu, Congjun (2005) A Taxonomy of ESL Writing Strategies. In Proceedings Redesigning Pedagogy: Research, Policy, Practice, pages pp. 1-10, Singapore
7. De Mello, G., Omar, N., Ibrahim, R., Ishak, N. and Rahmat, N. A Study of The Use of Writing Strategies from The Social Cognitive View, International Journal of Academic Research in Business and Social Sciences, 2023
8. Ghafar, Z. Teaching writing to Students of English as a foreign language: The Challenges Faced by Teachers, JOURNAL OF DIGITAL LEARNING AND DISTANCE EDUCATION, (2), pp. 483-490, 2023
9. Vejayan, L. and Yunus, M. Writing Skills and Writing Approaches in ESL Classroom: A Systematic Review, International Journal of Academic Research in Business and Social Sciences, (12), 2022
10. Ahmad, N., Murugesan, S. and Kshetri, N. Generative Artificial Intelligence and the Education Sector, Computer Journal, Vol. 56, pp. 72-76, 2023
11. Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jayaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawy, M., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., and Wright, R., o what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy, International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
12. Korostyshevskiy, V. Spoken language and fear of the blank page. Adult Learning, 29(4), 170-175., 2016
13. Palanichamy Naveen, Pavel Trojovský, Overview and challenges of machine translation for contextually appropriate translations, iScience, Volume 27, Issue 10, 2024, https://doi.org/10.1016/j.isci.2024.110878.
14. Marketing Novita AI. (2023). Llama 3 3.70B vs. Mistral Nemo: Which is suitable for multilingual chatbots? Retrieved from https://medium.com/@marketing_novita.ai/llama-3-3-70b-vs-mistral-nemo-which-is-suitable-for-multilingual-chatbots-8c84233bd8eb
15. Hilton-Jones, U., Project-Based Learning for Foreign Students in an English-Speaking Environment. Paper presented at the Annual Meeting of the International Association of Teachers of English as a Foreign Language (22nd, Edinburgh, Scotland, April 11-14, 1988
16. S. R. Bogireddy and N. Dasari, "Comparative Analysis of ChatGPT-4 and LLaMA: Performance Evaluation on Text Summarization, Data Analysis, and Question Answering," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10725662

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2025 doi:10.20944/preprints202502.1524.v1

17. Graham, S. and Perin, D. Writing Next: Effective Strategies to Improve Writing of Adolescents in Middle and High Schools – A Report to Carnegie Corporation of New York, Alliance for Excellent Education, Washington, DC, 2007.
18. Amin, M., Afzal, S. and Kausar, F. N. Undergraduate Students' Communication Problems, their Reasons and Strategies to Improve the Communication, Journal of Educational Research & Social Science Review, Vol. 2(2), pp. 109-117, 2022

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.