Smith, J. (2025). ENHANCING ESP STUDENT CRITICAL THINKING SKILLS and VOCABULARY ACQUISITION THROUGH a GENAI-BASED PROJECT. *Teaching English with Technology*, *25*(2), 82–107. https://doi.org/10.56297/vaca6841/ZDDD4208/HNVJ7828

ISSUES

ENHANCING ESP STUDENT CRITICAL THINKING SKILLS and VOCABULARY ACQUISITION THROUGH a GENAI-BASED PROJECT

Jorden Smith[1,2]

[1] Department of Education and Educational Innovation, Faculty of Law, Education and Humanities, Universidad Europea de Madrid, [2] Department of English Studies, Faculty of Philology, Universidad Complutense de Madrid

Keywords: Critical thinking, Vocabulary learning, GenAI, Project-based learning, ESP, Multimodal text creation

https://doi.org/10.56297/vaca6841/ZDDD4208/HNVJ7828

Teaching English with Technology

Vol. 25, Issue 2, 2025

UNESCO (2023) has called for educators to use pedagogical approaches that prioritize human agency and a responsible interaction between humans and generative artificial intelligence (GenAI). The present study, therefore, sought to design and implement an innovative 5-stage GenAI-based project with Sports Science English for Specific Purposes (ESP) students. Specifically, the objective of the project was for students to create multimodal texts for promotional purposes while developing both critical thinking skills and vocabulary range. Upon completing the project, 42 students participated in a survey, analyzed with both quantitative and qualitative methods, to gauge the impact of the project. Regarding critical thinking skills, 95.2% of students revealed they learned it was important to use these skills when using chatbots, emphasizing the need to verify GenAI-generated content and external sources a chatbot uses. Concerning vocabulary learned, 85.7% of the students, irrespective of their self-perceived level, reported learning useful topic-based and general vocabulary, thus strongly implying generalized intentional and incidental vocabulary learning. The study's main conclusion is that GenAI-based projects, when designed to primarily assess student critical reasoning, can create optimal conditions for vocabulary acquisition and critical thinking skill development. This approach can also have significant implications for GenAI-related teaching practice and student assessment.

1. Introduction

International standard-setting bodies have proposed the responsible integration of generative artificial intelligence (GenAI henceforth) into education by offering guidelines that prioritize human agency and an instructionally sound engagement between GenAI and humans (OECD, 2023; UNESCO, 2023). However, a lack of global consensus on standardized GenAI use in higher education is manifest, with some courses embracing it while others impose partial or total bans (Eaton, 2024; World Economic Forum, 2024). This implies that usage policies are being enforced at course level rather than institution level. Since teacher unpreparedness is a central reason for GenAI not being implemented effectively at Spanish universities (Cabellos et al., 2024; Escario et al., 2024), it can be inferred that many students are not being shown GenAI's capabilities for enhancing academic or professional work while promoting individual growth. This would

consequently indicate that the development of key competencies that underpin its principled use, such as critical thinking (CT) skills, is being hindered.

The English for Specific Purposes (ESP) discipline has long focused on developing critical thinking skills alongside more traditional language skills and strategic competencies involved in vocabulary acquisition. However, implementing GenAI in the ESP classroom will require a diversification of pedagogical approaches (Kostka & Toncelli, 2023) to enable the fostering of ESP-related GenAI skills that could make students more attractive propositions for employers. To this end, a GenAI-based project in which native Spanish-speaking undergraduate students created a multimodal text for promotional purposes is the focus of this paper. A key aim of the study was to identify a correlation between learners with varying levels of English proficiency and 1) CT skill development, and 2) key vocabulary learned through doing the project. Accordingly, survey results used to measure the impact of the project will be discussed along with their implications for teaching practice and further research involving GenAI tools.

2. Literature review

2.1. Fostering ESP students' critical thinking skills and vocabulary growth through GenAI-based pedagogy

According to Hyland (2022), the ESP classroom assumes a key role in enhancing students' professional or workplace literacies. However, accelerating GenAI use for professional purposes due to its information synthesis and content generation capabilities is reshaping traditional vocational literacies, meaning developing professional GenAI-related skills, such as CT skills, is of ever-increasing importance (Harvard Business Publishing Corporate Learning, 2025). Consequently, practical, interactive and context-based learning approaches that use AI to simulate profession-related environments have been suggested to address GenAI-related paradigm shifts in ESP (He et al., 2025).

A viable way to build ESP GenAI-related skills could be through GenAI-integrated project-based learning (PBL). Due to PBL's emphasis on collaboration, sustained inquiry and real-world challenges (Johnsen et al., 2023), the skills related to CT, defined in this article as "reasonable, reflective thinking that is focused on what to believe or do" (Ennis, 2011, p. 1), can seemingly be embedded in a PBL process integrating GenAI, as students are required to think deeply, analyse information and make reasoned decisions throughout (Du & Han, 2016). In today's context, through the integration of GenAI-powered information-retrieving assistants, such as Microsoft Copilot, project-based approaches can ostensibly develop CT skills, regarded

as key for the 21st century (Yu & Zin, 2023), since the proliferation of digital technology use and online information has resulted in the need to foster students' digital-related literacies (Smith & Storrs, 2023).

Furthermore, the intrinsic nature of project-based methodologies integrating GenAI can favor field-specific vocabulary interiorization in multiple ways. Firstly, using vocabulary in a meaningful context can facilitate vocabulary comprehension (Godwin-Jones, 2018). Secondly, the iterative nature of PBL ensures repeated exposure to new language, thus aiding long-term retention (Nation, 2001). Thirdly, students can deepen mastery by applying new vocabulary creatively (Maley & Peachey, 2015). Additionally, projects incorporating multimodal pedagogy through visual, audio or hands-on learning activities, such as when creating digital presentations for professional purposes, can positively impact language learning (Aghaei & Gouglani, 2016) and therefore cater for the vocabulary acquisition of students with diverse learning styles (Macedonia, 2015).

2.2. Previous studies on critical thinking and vocabulary development using GenAI

Limited studies to date have determined the impact of GenAI-based projects on ELT learner CT skill and vocabulary enhancement. However, in related research that has explored CT development, its relationship with GenAI appears to be both complementary and complex. While GenAI can enhance CT by offering varied viewpoints, generating ideas and promoting analysis and refinement of GenAI-output, over-reliance risks reducing cognitive effort and independent critical thought (Lee et al., 2025; Sardi et al., 2025). As a result of the latter, various authors stress the judicious application of GenAI involving higher-order thinking skills to help ESP students grasp its implications (Lytovchenko et al., 2025; Muchibbuddin & Zakiyyah, 2024). Nevertheless, in a relevant study whose results emphasize the positive CT gains to be made from GenAI use, Liu and Wang (2024) measured the critical thinking levels of 90 Chinese English literature university students pre-and post-intervention. In the experimental group, students used GenAI tools such as ChatGPT for both pre-class and in-class activities, whereas the control group followed more traditional methods, such as pre-class reading and holding debates during class time. The findings of the study showed that learners in the experimental group scored significantly higher on the Critical Thinking Disposition Inventory assessment tool post intervention, and qualitative data revealed this was due to GenAI tools' ability to provide immediate feedback, personalized learning paths and scaffolded learning experiences.

Various studies have measured ELT student perceptions of GenAI's impact on CT skills. Examining 39 students' views of integrating AI into an EAP module, Ngo and Hastie (2025) found increased criticism of GenAI output post module, and concluded that activities boosting AI literacy should also

promote CT. In their study involving 137 ESP engineering students at various academic levels, Sinkus and Ozola (2024) found that 90% of the students used AI tools to do academic work and 83.2% believed AI would become more prevalent. The results revealed a need to critically adopt AI tools to avoid issues such as plagiarism, cheating, overreliance and decreased quality of learning. Additionally, in a study by Avsheniuk et al. (2024), 31 ESP students in Ukraine shared their thoughts before and after using ChatGPT to review and evaluate their work. The analysis of results showed ChatGPT enhanced learner CT skills such as analysis and decision-making, seemingly by interpreting student output and providing personalized feedback.

With regards to GenAI-influenced vocabulary improvement, chatbots' ability to provide tailored learning experiences marks significant advancement in aiding student vocabulary learning (Kostka & Toncelli, 2023; Romadhon, 2024), along with fostering vocabulary-learning related benefits such as autonomous learning (Peláez-Sánchez et al., 2024), reducing learning anxiety (Mohamed & Alian, 2023) and catering for visual learners (Dokukina & Gumanova, 2020). Furthermore, the advantages of chatbots' ability to give instant feedback on student production have been highlighted (Hong, 2023), and, in their systematic review of human–chatbot comparison research, Jeon and Lee (2024) found that chatbots elicit student vocabulary comparable to human–human interactions yet provide less detailed vocabulary explanation.

Several studies comparing conventional vocabulary teaching to student–chatbot interaction found the latter to be more effective. For example, Silitonga et al. (2024) found Indonesian ESP students' posttreatment vocabulary scores were higher for those using chatbots. This was because those students better negotiated new word meanings through enhanced noticing and attention. In Zhang and Huang's (2024) experiment with 52 Chinese high school students in an intermediate English course, participants learned target words with or without a customized chatbot. Posttreatment assessments showed the group using chatbots performed better, providing evidence that chatbot use can aid receptive, productive and incidental vocabulary acquisition. Qasem et al. (2023) examined the impact of chatbot use on the vocabulary learning of 40 Saudi ESP students. Posttreatment, students using the dialogue chatbot scored higher on a vocabulary test due to receiving quick and effective explanations of new meanings. Additionally, in the study by Waziana et al. (2024), 100 Indonesian EFL students gave their perceptions about how chatbots had enhanced their writing abilities. Results showed that chatbots improved English proficiency by enriching student vocabulary through feedback containing unknown synonyms. The researchers concluded this could improve writing through expanding lexical repertoires.

Therefore, as no research has yet explored the effects of a GenAI-based PBL approach on ESP student CT skill development and vocabulary acquisition, this study seeks to assess the impact of said pedagogical approach on these cognitive processes and detect trends among Spanish ESP students with different levels of English proficiency. Accordingly, the research questions to be addressed are:

> RQ1: What was the impact of the GenAI-based project on the CT skill development of students with varying English proficiency levels?
>
> RQ2: What was the impact of the GenAI-based project on the vocabulary learning of students with varying English proficiency levels?

3. Methodology

3.1. The Gen-AI-based project

3.1.1. EDUCATIONAL CONTEXT

The project was completed by 2nd-year students enrolled in an onsite English course at a Spanish private university, focused on academic skills and sports-related language. The project, aligning with a pillar of the university's academic model, "Data Driven Focus" (Universidad Europea, n.d.), also integrated AI and CT skill development. The project spanned ten 100-minute sessions over the 15-week course and accounted for 50% of the final grade.

3.1.2. PROJECT OBJECTIVES

The project's main goal is to help students create multimodal texts to present academic work or ideas online. Upon completion, students achieve three goals: 1) develop a schema for multimodal text design essential for communicating sports-related ideas; 2) become autonomous, critical assessors of GenAI content; and 3) learn English language essential for their discipline.

3.1.3. PROJECT STAGES

Stage 1 - Brainstorming. Students lay the foundations for analyses undertaken at various stages. Pairs/groups choose their text-type and topic, and brainstorm key text design criteria for five key design aspects: 1) main ideas (and associated authors), 2) layout, 3) idea organization, 4) key language, and 5) language style.

Stage 2 - Chatbot as an information retrieval tool. Students consolidate their knowledge of the five design aspects of their multimodal text by either confirming ideas or filling gaps in knowledge apparent in Stage 1. Initially, students prompt a chatbot to provide key criteria for each aspect along with accompanying sources, and they subsequently consider which information

is reliable using the CRAAP Test (Blakeslee, 2004). This tool was chosen for its exhaustive criteria, suitable for students with minimal experience in evaluating sources (Sye & Statton Thompson, 2023). To ensure a rigorous evaluation of chatbot data, students complete a table with information about their chosen criteria and corresponding sources (Appendix 1). If the chatbot does not provide a source or if there is unclear source attribution for a design criterion, the information is processed in a second table (Appendix 2).

Stage 3 - Creating the multimodal text. Following the finalized specific criteria decided in Stage 2, students create their multimodal texts, drafting the text and adding complementary visual information (e.g., images, tables, videos).

Stage 4 - Chatbot as a text revisor. Groups prompt the chatbot to improve the written part of their multimodal text. Once the chatbot returns its upgrade, groups compare it to their Stage 3 text using an analytical checklist (Appendix 3). Afterwards, students explore a linguistic feature (collocations) before identifying topic-specific examples in their chatbot-revised text.

Stage 5 - Finalizing the multimodal text. Based on prior comparative and linguistic analyses, groups make final adaptations to their drafted texts (Stage 3). Changes could be made regarding key ideas (and authors), the organization of ideas in the text, key language and language style.

3.1.4. PROJECT ASSESSMENT

Assessment consists of two presentations (after Stages 2 and 5, respectively) and submitting a "digital dossier". Instructions for each presentation are found in Appendices 4 and 5. The digital dossier, shared via a Google Drive folder, comprises chatbot information evaluation tables (Stage 2); the initial multimodal text (Stage 3); the comparative analysis checklist (Stage 4), a personalized list of collocations learned with translations (Stage 4), and the finalized multimodal text (Stage 5).

3.1.5. RESPONSIBLE AI-USE DISCLOSURE AND PEDAGOGICAL MAPPING

Students could choose between Microsoft Copilot (version unspecified due to continuous updates) and OpenAI's ChatGPT (either GPT-3.5 or the GPT-4 paid version, depending on the subscription), between September and December 2024 to retrieve design criteria (Stage 2) and revise their written texts (Stage 4). In Stage 3, they were shown various GenAI tools they could experiment with to create suitable multimedia material: for example, text-to-image and text-to-video applications, such as Freepik (https://www.freepik.com/) and Invideo (https://invideo.io/), respectively. Groups then used Canva (www.canva.com) graphic design tool to integrate textual and multimodal elements.

Table 1. CEFR levels of students in the study

A2	B1	B2	C1	Total
2	7	24	9	42

In accordance with institutional guidelines (Vicerectorado de Profesorado e Investigación, 2024), students were made aware of the risk of sharing personal data and they were encouraged to read each GenAI application's privacy policy. As an integral part of the project was for students to critically consider GenAI output, students applied the CRAAP Test (Stage 2) and worked through analysis tables (Stages 2 and 4, respectively) to mitigate risks of bias or misinformation.

3.2. Participants

The participants were 68 undergraduate students from three separate ESP Sports Science groups. Forty-two students provided data, as 26 students were absent on data-collection day. According to Spanish educational legislation (Ministerio de Educación, Formación Profesional y Deportes, n.d.) the language competences achieved by students finalizing post-compulsory secondary education (Bachillerato), which are created in line with the Common European Framework for References of Languages (CEFR) (Council of Europe, 2001) level descriptors, suggest high school-leavers graduate with an approximate B1-B2 level. However, after two weeks of classes (the project began in the semester's third week), a broad range in proficiency levels among the cohorts became increasingly perceptible. As a result, students were asked to self-report CEFR English levels to anchor the focus of this study's research questions. As seen in Table 1, levels ranged from A2 to C1.

3.3. Research design and instruments

The project was carried out between September and December 2024. In January 2025, ethical clearance was obtained from the university's ethics committee. Later that same month, following this approval, students were invited to complete a Google Forms questionnaire (Appendices 6-11) during class time upon providing informed consent. Data was gathered using a mixed-method approach to gain a comprehensive understanding of the impact of the project. The questionnaire was in English and Spanish to guarantee comprehension, and students could respond in either language. Answers given in Spanish have been translated into English in this paper. The six questions used for this study were part of a broader set of 19 questions employed to obtain student perceptions of the project. The questionnaire also aimed to discover topics that could inform future methodological adjustments. The questions covered the following themes: learner motivation, perceptions of assessment, linguistic resources, the project's impact, and its challenging aspects.

For quantitative data, Likert scale questions, each with the four response options "Strongly agree", "Agree", "Disagree" and "Strongly Disagree", addressed student perception of 1) the importance of using CT skills with GenAI output (Appendix 6), 2) learning outcomes in terms of CT skills (Appendix 7) and 3) new vocabulary acquisition (Appendix 8). Open-ended follow-up questions were employed to gather qualitative data for both CT skill development (Appendix 9) and vocabulary learning (Appendix 10) quantitative responses, and a further open-ended question asked students what the most valuable skill learned from doing the project was (Appendix 11).

3.4. Reproducibility and open-source materials

To enable the re-use of project materials and the transparency of survey instruments and data, all study materials can be found at https://osf.io/7sfyk/ (Smith, 2025).

3.5. Data analysis

Likert scale responses were analysed using descriptive statistics to determine the distribution of responses. To reveal general agreement or disagreement on the project's impact on CT and vocabulary development and to mitigate data sparsity (including a zero-count column), Likert responses were merged into "Agree" versus "Disagree". Subsequently, a 4×2 chi-square test of independence was performed to examine the association between CEFR levels and student responses. Effect size was measured using Cramér's V.

Qualitative data was analysed using reflexive thematic analysis (RTA) (Braun & Clarke, 2012, 2020) in an iterative process. Given that knowledge and meaning are contextually constructed through interaction (Berger & Luckmann, 1991), and in this study interaction was between students, the teacher and GenAI, a social constructionist epistemology underpinned the theoretical assumption of the analysis. Occupying a dual role of teacher and researcher, peer debriefing complemented reflexive analytic memo writing throughout the RTA process to mitigate the influence of familiarity with students and the pedagogic approach on the analysis. Interpretations and reflections in the memos became the basis for debriefing sessions with two experienced colleagues whose considered feedback informed subsequent memo entries. This allowed a re-examination of assumptions and clarification of theme boundaries and furthermore supported a reflexive and transparent analytical process. A predominantly open-coded, inductive analytical approach was used to be as true as possible to student-expressed ideas in the data, but an element of deductive analysis was also employed to ensure the themes generated were relevant to each RQ. No preference was given to either semantic or latent coding as both explicit and implicit interpretations were treated as socially constructed expressions of meaning.

Table 2. Six-phase process of Braun & Clarke's (2012, 2020) TFA

Phase	Description
1. Familiarization with data	Students' comments read through several times and preliminary trends noted.
1. Generating initial codes	Semantic coding carried out initially before latent coding applied on subsequent readings of data.
1. Constructing initial themes	Codes for each RQ collated and classified into emerging themes and sub-themes.
1. Reviewing themes	Themes reviewed undergoing a process of moving between codes, patterns identified in the data and corresponding student quotes anchoring said themes and sub-themes.
1. Defining and naming themes	Themes finalized after incidences of thematic overlapping addressed e.g., checking chatbot output (source integrity vs general inaccuracies).
1. Report writing	Theme reports for each RQ written ensuring ideas presented aligned closely with the data while meeting research aims.

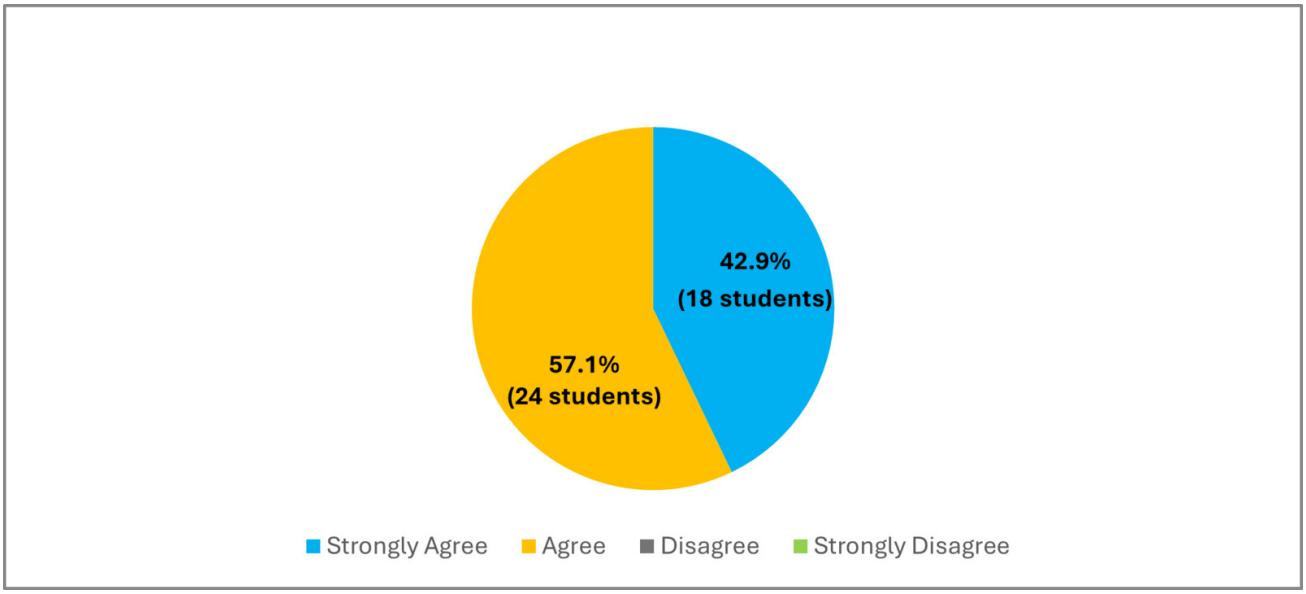

Figure 1. Student perception of the importance of using CT skills with GenAI content

4. Results and analysis

4.1. Critical thinking skill development

When asked if they agreed that it is important to think critically about GenAI output, 57.1% (24 students) said they "agreed", while the remaining 42.9% (18 students) said they "strongly agreed", as shown in Figure 1

Furthermore, when asked if they had learned the importance of employing CT skills when using GenAI because of undertaking the project (Figure 2), 24 (57.1%) students said they "agreed", and 16 (38.1%) "strongly agreed". One student (2.4%) "disagreed", while another learner (2.4%) "strongly disagreed". When broken down into CEFR levels, one student at A2 level, two at B1 level, eight at B2 level, and five students at C1 level "strongly agreed" that the project raised awareness of the importance of critically considering GenAI output. Furthermore, one learner at A2 level, four at B1

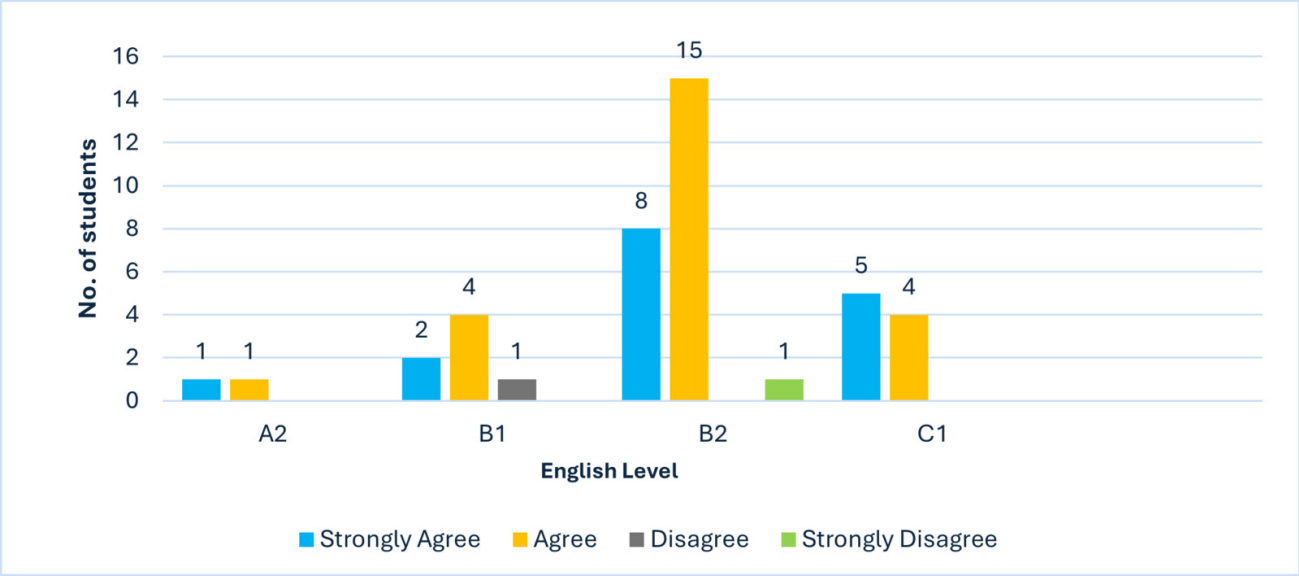

Figure 2. Perception of students across English proficiency levels of importance of using CT skills with GenAI content after undertaking project

level, 15 at B2 level, and four at C1 level said they "agreed" that undertaking the project highlighted the need to critically consider GenAI output. In contrast, one student at B1 level "disagreed" and another at B2 "strongly disagreed" that the learning experience contributed to that realization.

The qualitative data significantly supported the quantitative findings. Moreover, students frequently mentioned the need to apply analytical and evaluative skills when using GenAI, as stated by a B2-level student (1). Additionally, a B2-level student and a C1-level classmate (2 and 3) emphasized the requirement to check it for potential inaccuracies, as they did primarily in Stages 2 and 4 of the project.

(1) *It is necessary to apply critical thinking and personal judgment to evaluate the output offered by AI.*

(2) *I believe it is essential to verify the information provided by a chatbot, as it is not always accurate.*

(3) *GenAI doesn't always understand what you want and can change things for better, but also for worse.*

A C1 learner (4) and a B2-level classmate (5) showed awareness of the repercussions of communicating false information if GenAI is not checked.

(4) *Because it is information we have to present to someone, I think it is important to verify its accuracy.*

(5) *You need to ensure that the information provided by the sources is accurate and up to date in order to present the most accurate information without errors.*

Various students mentioned learning to use GenAI judiciously was the most valuable thing learned from doing the project. This is alluded to by a B1-level student (6) and a B2-level student (7), with an A2 student (8) stating how the learning experience had deepened their understanding of the implications of GenAI use.

(6) *You can use AI to help you, but not to do the entire task.*

(7) *Learning to use AI in a professional and critical manner.*

(8) *AI tools are not always reliable, so we must be very careful when we using* [sic] *AI tools. That, I feel, was the most valuable skill is* [sic] *making decisions based on a lot of information.*

Consequently, these comments could be interpreted twofold: 1) using it responsibly throughout the project, students heightened their awareness of the negative implications of the unethical use of GenAI, and 2) lower-level students can both engage with GenAI output in English and judge its reliability or utility for specific academic or professional purposes.

In general, student contributions emphasized the necessity of evaluating and confirming the reliability of sources a chatbot uses. One B2-level student (9) stressed the significance of verifying chatbot-provided sources and assessing them. Furthermore, another B2-level learner (10) mentioned the implications of source quality for the integrity of their work, and an A2 student (11) admitted the danger of not checking sources.

(9) *The chatbot has the ability to search for information anywhere on the internet, so it is important to ask for sources and verify their reliability.*

(10) *To do a good job, you must thoroughly investigate the sources you rely on.*

(11) *Because if not, we might end up using sources that are not truthful.*

Therefore, it would seem that students have become aware of the importance of using credible sources and how chatbot-invented data can affect the academic integrity of their work.

A B2-level student (12) was the only one to elaborate on their quantitative response of the two respondents who stated they had not learned the importance of using CT skills with GenAI content during the project.

(12) *I was already aware that the information you obtain needs to be verified.*

However, various students elaborated on how the project had taught them to critically consider chatbot-produced output, which would suggest they learned skills that can be useful for more general or academic research purposes. A B1-level student (13) mentioned the usefulness of the CRAAP

Table 3. Percentages of students who learned vocabulary across CEFR levels

CEFR Level	Strongly agree (%)	Agree (%)	Disagree (%)	Strongly Disagree (%)	Total
A2	50	50	0	0	2
B1	28.6	71.4	0	0	7
B2	29.2	54.2	16.7	0	24
C1	11.1	66.7	22.2	0	9
Total	26.2	59.5	14.3	0	42

test to evaluate sources, a B2-level student (14) referred to being able to obtain high-standard GenAI output and a C1 learner (15) said they had learned to check source integrity.

(13) *During the project, I was able to learn, thanks to the CRAP* [sic] *test, new ways of analysing various sources found on the internet, which help me identify the best information at my disposal.*

(14) *The project taught me to apply critical skills to ensure that decisions and outcomes generated with AI are reliable and high quality.*

(15) *I would never check whether the source I was using was reliable or not, and thanks to the project, I now have the tools to determine if a source is trustworthy.*

4.2. Vocabulary learning

As shown in Table 3, when asked if they had learned useful language doing the GenAI-based project, 85.7% of the students agreed. More specifically, 59.5% (25 students) said they "agreed", 26.2% (11 learners) said they "strongly agreed", and 14.3% (six students) said they "disagreed". Concerning proficiency levels (Figure 3), at the A2 level, both students (100%) either strongly agreed or agreed that they learned vocabulary. Similarly, all seven students at B1 level (100%) strongly agreed or agreed they acquired new vocabulary, which means that the project was highly effective for vocabulary learning at lower levels. At B2 level, 20 out of 24 students (83.3%) felt they learned new vocabulary (seven strongly agreed and 13 agreed), whereas a notable minority of four students (16.7%) disagreed. At the C1 level, seven out of nine students (77.8%) felt they learned new vocabulary (one "strongly agreed", and six "agreed"), while two learners (22.2%) "disagreed", which would imply that the methodology involved needs adaptation to better suit higher-level learners.

The adjusted 4×2 chi-square was not significant, $\chi^2(3) = 2.07$, $p = 0.557$, and the effect size (Cramér's $V = 0.137$) suggested a weak association between proficiency levels and students' learning perceptions.

Generally, students at the B2 and C1 levels offered more insightful comments on vocabulary development. Overall, students across all levels commented on learning many new words, both topic-specific and other types of vocabulary.

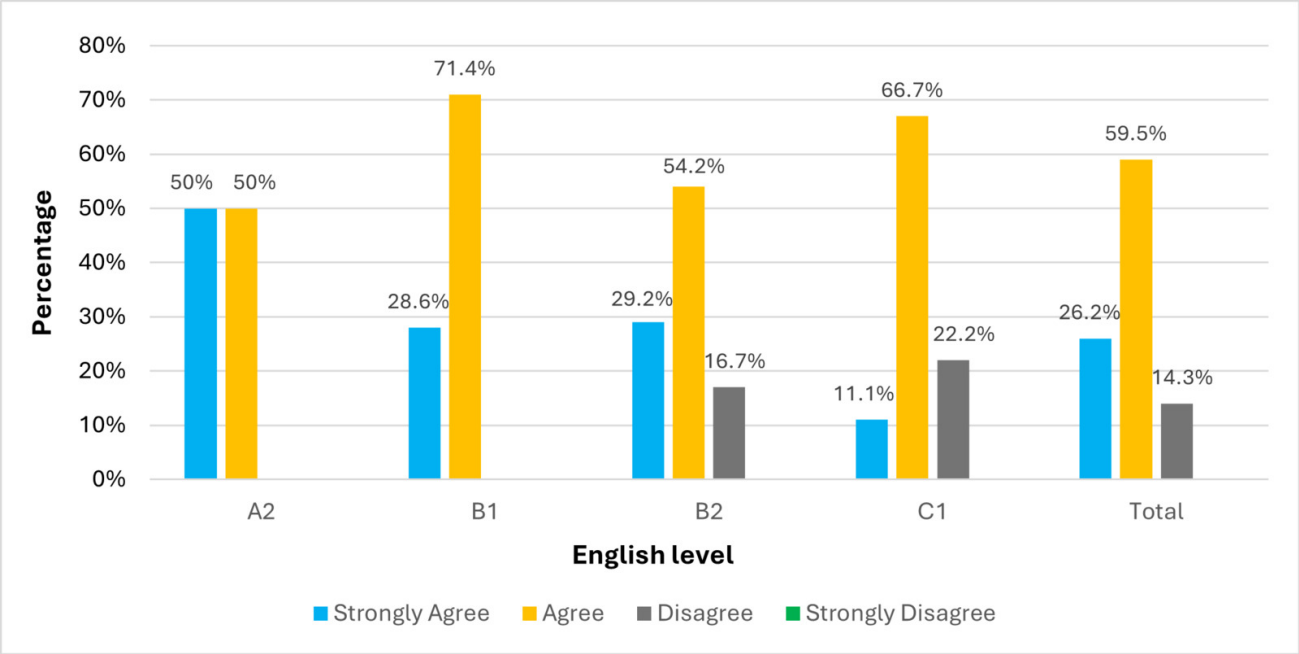

Figure 3. Student perception of vocabulary learning across the CEFR levels

Two B2-level learners (16 and 17) mentioned learning profession-related vocabulary, while another B2-level classmate (18) stated learning topic-specific collocations, which were the linguistic focus in Stage 4. A C1-level student (19) stated learning language they were not familiar with, which could suggest that language related to multimodal text design could represent a significant portion of all new language learned.

(16) *I have learned a lot of new vocabulary on this topic, including more technical and professional language.*

(17) *I consider that I have learned quite a bit of language given that I did not know many words in this field. I found it very useful for the future.*

(18) *By focusing my text on the topic of sports, I learned useful words and collocations for my future.*

(19) *There were a lot of new terms I didn't know, mainly because I don't use them daily, so it was a nice experience to learn new vocabulary in a different aspect.*

Nevertheless, a few students mentioned not having learned any useful vocabulary. One B2-level student (20) believed that learning multimodal text design-related language would not be valuable, and another B2-level learner (21) indicated not learning any new language at all.

(20) *We have learned new language related to multimodal texts, but it is not something that will be very useful for us.*

(21) *Because the words and expressions that we were using in the final text I already knew them* [sic].

These utterances, therefore, highlight potential problems with project-based learning design. Firstly, related to (20), students need to see real value and relevance in the tasks they undertake. Secondly, the teacher may need more control over topic selection to make sure students interact with vocabulary beyond their comprehension to avoid cases like (21).

Students also emphasized improved vocabulary range and mentioned the application of newly acquired vocabulary in both the finalized multimodal text and the presentations they gave. One A2-level student (22) mentioned learning vocabulary (linkers) that was not part of the linguistic focus (Stage 4) of the project, thereby suggesting incidental vocabulary learning took place when analysing chatbot output.

(22) *It has helped me somewhat improve my vocabulary regarding linkers and expressions.*

Furthermore, a C1-level student (23) stated how useful GenAI output was in exposing them to new words they autonomously learned and applied without teacher guidance.

(23) *Using artificial intelligence has provided me with new words I have learned to use that I otherwise would not have learned.*

Additionally, a B2-level student (24) highlighted the benefits of varied encounters with new language throughout the project's stages and assessment.

(24) *When doing the project you learn new language, and when you explain and justify it, it can be used as a resource.*

This could suggest students engage in deeper, more meaningful processing of GenAI-generated new language by judging if it should be included in their finalized multimodal text.

5. Discussion and pedagogical implications

RQ1 examines the impact of the project on the CT skill development of students at different levels of English proficiency. The results show that virtually all students at all CEFR levels (95.2%) felt that carrying out the project had heightened their awareness of the need to critically consider GenAI-generated output and GenAI tool use. Furthermore, student comments imply that through handling GenAI-produced data throughout the project, they have learned it is important to both check it for mistakes and evaluate the sources a chatbot uses. This is to avoid producing work with erroneous data (linguistically or factually) or work that could be deemed as lacking academic rigor, respectively. Therefore, it can be assumed that these students had not engaged critically with GenAI-generated content previously,

thus highlighting the need for policymakers and educators alike to adopt approaches that foster CT skills in the context of GenAI usage (Gonsalves, 2024; Wang et al., 2024).

Additionally, the structured use of chatbots with analysis tables and the CRAAP test (both Stage 2), and the comparative analysis checklist (Stage 4) appears to have raised students' awareness of the integrity and accuracy of chatbot output. This could subsequently suggest that these ESP students have developed key skills related to critical analysis and source evaluation that result in an effective use of GenAI. Moreover, these skills can be transferred to both academic research and more casual information-seeking activities that intend to resolve profession-related doubts or influence occupational decision-making (Li, 2023; Murawski, 2014). Accordingly, using chatbots for information-seeking purposes may imply the need to consolidate research-related reading skills other than idea and feature analysis and source verification in the ESP classroom. As well as more practical strategies such as scanning for specific information or skimming for gist, students will need to recognize text structure, notice implicit or explicit messages, guess the meaning of unknown words, and successfully interpret the tone and style of text, among other more nuanced skills, for a more complete understanding of the variety of chatbot-provided internet sources.

RQ2 examines the project's impact on vocabulary learning for students with varying levels of English proficiency. A significant majority (85.7%) stated that they learned useful words and/or expressions related to sport, multimodal text design, and more general English language. However, some students stated that some vocabulary they learned would not be useful for professional purposes or that they did not learn any topic-specific vocabulary at all, meaning the methodology requires adjustments to better meet higher-level students' needs.

Nevertheless, participant contributions suggest that exposing students to corrected or novel language via GenAI output represented an effective way for students to learn vocabulary without teacher control. For example, from when new language is discovered (Stage 2), used for the first time as part of the written text (Stage 3), possibly corrected or exchanged for more context-appropriate chatbot-derived alternative (Stage 4), conscientiously inserted into the finalized text (Stage 5), to when its inclusion in the finalized text is explicitly justified (assessment), students undergo attentive, high-quality encounters with the new language thus increasing the chances of acquiring it (Webb & Nation, 2017).

Accordingly, if vocabulary can be acquired through students critiquing their own chatbot-reviewed texts, assessment may benefit from focusing on decision-making during the learning or text-creation process and not on the final product itself. However, to do such assessments effectively, learners will need the linguistic resources necessary to demonstrate analyses and

evaluations of enhanced language use. Depending on the layers of complexity of the text genre, this could entail students knowing metalanguage to describe text structure (e.g., introduction, topic sentences, body), grammar and syntax (e.g., clauses, verb tenses, parts of speech), discourse (e.g., cohesive devices, linking works, ellipsis) and lexical features (e.g., register, domain-specific, collocation), among other categories of metalanguage. Consequently, teachers would need to gauge learners' receptive language knowledge levels to ensure they align with the complexity of target text's genre analysis.

Relevant to Zhang and Huang's (2024) finding, previously mentioned, asking students to contrast their own written texts with a chatbot-revised version of it appears to be beneficial for incidental vocabulary acquisition. Webb et al. (2020) define that as the learning of language that is a "by-product of a meaningful task". This is due to the various favourable conditions for incidental vocabulary learning that are created in the project. Firstly, students draft their texts at Stage 3, meaning they will likely have a strong grasp of the meaning of any alternative words suggested by the chatbot at Stage 4. More specifically, it is highly likely that students will understand 95%-98% of the words in the text, the desirable threshold for full comprehension of a text (Hu & Nation, 2000), thus achieving a clearer context or co-text surrounding an unknown word. This can facilitate correct inference that enables its successful acquisition (Webb, 2008). Secondly, it is arguably a task with elevated motivational and cognitive engagement, thus giving it a high involvement load (Laufer & Hulstijn, 2001). In other words, 1) the learners have a strong *need* to learn and use new vocabulary since they want to masterly craft their text on a topic they will showcase; 2) while reviewing the chatbot-revised text in Stage 4, students actively *search* for new words or better alternatives for their texts; and 3) they must simultaneously assess the chatbot's suggestions, comparing them with their original text, and subsequently *evaluate* which words are most appropriate for their context and therefore finalized text (Stage 5). Finally, learners can use images or videos to complement their multimodal text, meaning the combination of text and visual aspects could help support incidental vocabulary learning, too, as discovered by Shahrokni (2009).

Therefore, if created following detailed teacher-controlled prompts, chatbot-produced enhancements of student-written texts could improve learning outcomes at the language focus stage of a teaching procedure for various reasons. First, since the text represents an upgrade of a student's own text, thus giving the sensation of authentic, polished language use, it may prove more engaging for language analysis tasks than cognitively inaccessible textbook display texts (Mishan, 2022) or contrived teacher-made attempts. Secondly, students gain a clear understanding of concepts and unfamiliar vocabulary in the text, as alluded to above, meaning any linguistic analysis carried out can be performed more seamlessly as a result. Finally, teachers can end their dependence on prescribed textbook language exercises since

they can create tailor-made tasks that draw students' attention to linguistic features that are better aligned to learner needs or specific course objectives. For instance, for learners at lower levels, lexical features such as technical terms or jargon may be prioritized, while higher-level students may benefit from studying more nuanced features, such as register.

6. Conclusion

This study has examined the CT skill development and vocabulary learning of ESP students of diverse English proficiency who were engaged in project-based learning involving GenAI. To meet the project's objectives, students had to use GenAI to create multimodal texts on sport-related topics for presentational purposes, with stages of the project involving the analysis and evaluation of research and linguistic content produced by a chatbot. Regarding the project's impact on students across CEFR levels, the findings show no detectable association between the levels and CT skills development because almost all participants at all represented levels reported improvements in CT skills after undertaking the project. Furthermore, the majority of learners across all levels reported learning new vocabulary.

The findings of this paper show that through the responsible use of chatbots, students can harness the benefits of GenAI by critically determining the features of information sources and the effective use of English language. This would consequently suggest they have become aware of how to leverage the technology in a principled manner for both academic and professional settings, thus meeting the recommendations of UNESCO (2023) and OECD (2024). The results, therefore, further underline the potential of the ESP subject to equip learners with GenAI skills that can go beyond the classroom.

More specifically, the study's pedagogical implications suggest that by exposing learners to upgraded chatbot-generated versions of their own texts, there is scope for providing students with a more personalized and thus engaging learning experience. With teacher support, students can be guided to analyze proficient language use for the array of text genres a chatbot can produce. However, an obvious goal must be for students to replicate such linguistic analyses and evaluations in an autonomous manner, such as when required for academic or profession-related matters.

Although the findings of this study contribute valuable information to the nascent study of GenAI-based teaching approaches in ESP, several limitations should be addressed. Firstly, all the participants in this research studied Sports Science-related degrees, and the sample size is relatively small. Furthermore, student CEFR proficiency levels were self-reported. Consequently, future studies should encompass a larger sample with students from varying fields who have assigned CEFR levels based on performance in standardized tests to obtain more robust results. Secondly, students gave their own perceptions regarding CT and vocabulary development, so a longitudinal research design

involving both pre- and posttreatment tests on CT skills and vocabulary learning would yield more objective data regarding individual growth. Thirdly, this study examines the perceptions of students who were present on the day the survey was conducted. Since the survey's non-response rate was 38.2%, the responses obtained (61.8%) may overrepresent students who had favorable views of the project. Therefore, timing data collection to avoid external reasons for irregular attendance, such as end-of-course exam preparation, may improve response rates and consequently enhance data reliability.

It is also recommended that future studies explore the implications of basing student assessment on the decision-making processes involved in creating English-language work using GenAI applications, as carried out in the project this article describes. Furthermore, studies on students critically engaging with chatbot texts versus conventional teaching methods could offer key insights for implementing GenAI effectively and increasing incidental vocabulary learning.

Acknowledgement

I would like to express my gratitude to the peer reviewers for their considered and constructive feedback. Their insights helped strengthen the definitive version of the article and are greatly appreciated.

Funding

This research received no external funding.

Conflict of interest

The author declares no conflict of interest.

Ethics approval

The study received ethical clearance from the Escuela de Doctorado e Investigación (Code No. 2025-41). All participants gave informed consent prior to taking part in the experiment and had the right to withdraw their data/consent at any time during it. Their data were treated anonymously, stored securely, and processed in accordance with applicable data protection laws (GDPR).

Author contributions

The author was solely responsible for all aspects of the manuscript, including its conceptualization, research, writing and editing.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at http://creativecommons.org/licenses/by/4.0 and legal code at http://creativecommons.org/licenses/by/4.0/legalcode for more information.

REFERENCES

Aghaei, K., & Gouglani, F. K. (2016). Multimodal pedagogy and L2 vocabulary retention. *International Journal of English Language & Translation Studies*, *4*(3), 142–153. http://www.eltsjournal.org

Avsheniuk, N., Lutsenko, O., Svyrydiuk, T., & Seminikhyna, N. (2024). Empowering language learners' critical thinking: Evaluating ChatGPT's role in English course implementation. *Arab World English Journal*, *15*(1), 210–224. https://doi.org/10.24093/awej/ChatGPT.14

Berger, P. L., & Luckmann, T. (1991). *The social construction of reality: A treatise in the sociology of knowledge*. Penguin Books.

Blakeslee, S. (2004). The CRAAP Test. *LOEX Quarterly*, *31*(3), 6–7.

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 2: Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57–71). American Psychological Association. https://doi.org/10.1037/13620-004

Braun, V., & Clarke, V. (2020). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238

Cabellos, B., de Aldama, C., & Pozo, J.-I. (2024). University teachers' beliefs about the use of generative artificial intelligence for teaching and learning. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1468900

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Dokukina, I., & Gumanova, J. (2020). The rise of chatbots: New personal assistants in foreign language learning. *Procedia Computer Science*, *172*, 542–546. https://doi.org/10.1016/j.procs.2020.02.212

Du, X., & Han, J. (2016). A literature review on the definition and process of project-based learning and other relative studies. *Creative Education*, *7*, 1079–1083. https://doi.org/10.4236/ce.2016.77112

Eaton, L. (2024). *Syllabi policies for AI generative tools*. Google Docs. https://docs.google.com/document/d/1RMVwzjc1o0Mi8Blw_-JUTcXv02b2WRH86vw7mi16W3U/

Ennis, R. H. (2011). *The nature of critical thinking: An outline of critical thinking dispositions and abilities*. https://education.illinois.edu

Escario, I., García, A. N., Ripollés, M., Arriaga, C., Cáceres González, P. A., & García Cumbreras, M. A. (2024). Impulsando la Educación Superior con IAGen: oportunidades y retos para docentes. [Promoting higher education with GenAI: Opportunities and challenges for teachers]. *REDU. Revista de Docencia Universitaria*, *22*(2), 137–157. https://doi.org/10.4995/redu.2024.22065

Godwin-Jones, R. (2018). Contextualized vocabulary learning. *Language Learning & Technology*, *22*(3), 1–19. https://doi.org/10.64152/10125/44651

Gonsalves, C. (2024). Generative AI's impact on critical thinking: Revisiting Bloom's Taxonomy. *Journal of Educational Technology Systems*, *52*(1), 1–16. https://doi.org/10.1177/02734753241305980

Harvard Business Publishing Corporate Learning. (2025). *Gen AI fluency at work: How organizations unlock the full potential of an AI-proficient workforce*. Harvard Business Publishing. https://www.harvardbusiness.org/wp-content/uploads/2025/04/CRE6080_CL_Perspective_Gen-AI-Fluency_April2025.pdf

He, L., Zhang, X., & Huang, Y. (2025). Integration of artificial intelligence in English for specific purposes education: Strategies and effectiveness. In R. Koul, S. Roslan, M. N. Bin Abdul Rahman, & C. E. C. Lee (Eds.), *Proceedings of the 2nd International Conference on Educational Development and Social Sciences (EDSS 2025)* (Vol. 924). Advances in Social Science, Education and Humanities Research. https://doi.org/10.2991/978-2-38476-400-6_68

Hong, W. C. H. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research. *Journal of Emerging Trends in Innovation*, *5*(1). https://doi.org/10.61414/jeti.v5i1.103

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430. https://doi.org/10.64152/10125/66973

Hyland, K. (2022). English for specific purposes: What is it and where is it taking us? *ESP Today*, *10*(2), 202–220. https://doi.org/10.18485/esptoday.2022.10.2.1

Jeon, J., & Lee, S. (2024). Can learners benefit from chatbots instead of humans? A systematic review of human-chatbot comparison research in language education. *Education and Information Technologies*, *29*, 1–26. https://doi.org/10.1007/s10639-024-12725-9

Johnsen, M. M. W., Sjølie, E., & Johansen, V. (2023). Learning to collaborate in a project-based graduate course: A multilevel study of student outcomes. *Research in Higher Education*, *65*, 439–462. https://doi.org/10.1007/s11162-023-09754-7

Kostka, I., & Toncelli, R. (2023). Exploring applications of ChatGPT to English language teaching: Opportunities, challenges, and recommendations. *TESL-EJ*, *27*(3). https://doi.org/10.55593/ej.27107int

Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26. https://doi.org/10.1093/applin/22.1.1

Lee, H. P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 1, 2025, Yokohama, Japan*. https://doi.org/10.1145/3706598.3713778

Li, Y. (2023). How critical is critical thinking skill for the future of work: Developing curriculum with an AI perspective. *Scholedge International Journal of Management & Development*, *10*(9). https://doi.org/10.19085/sijmd100901

Liu, W., & Wang, Y. (2024). The effects of using AI tools on critical thinking in English literature classes among EFL learners: An intervention study. *European Journal of Education*, *59*, Article e12804. https://doi.org/10.1111/ejed.12804

Lytovchenko, I., Lavrysh, Y., Synekop, O., Lukianenko, V., Chugai, O., & Shastko, I. (2025). The use of ChatGPT in task-based ESP learning at university: Does it make a difference? *International Journal of Interactive Mobile Technologies*, *19*(2), 4–22. https://doi.org/10.3991/ijim.v19i02.51115

Macedonia, M. (2015). Learning styles and vocabulary acquisition in second language: How the brain learns. *Frontiers in Psychology*, *6*, 1800. https://doi.org/10.3389/fpsyg.2015.01800

Maley, A., & Peachey, N. (Eds.). (2015). *Creativity in the English language classroom*. British Council. https://www.teachingenglish.org.uk/sites/teacheng/files/pub_F004_ELT_Creativity_FINAL_v2%20WEB.pdf

Ministerio de Educación, Formación Profesional y Deportes. (n.d.). *Lengua extranjera – Competencias específicas (Bachillerato) (Foreign languages – Specific competences (Baccalaureate))*. Gobierno de España. https://educagob.educacionfpydeportes.gob.es/curriculo/curriculo-lomloe/menu-curriculos-basicos/bachillerato/materias/lengua-extranjera/competencias-especificas.html

Mishan, F. (2022). The global ELT coursebook: A case of Cinderella's slipper? *Language Teaching*, *55*(4), 490–505. https://doi.org/10.1017/S0261444820000646

Mohamed, S. S. A., & Alian, E. M. I. (2023). Students' attitudes toward using chatbot in EFL learning. *Arab World English Journal*, *14*(3), 15–27. https://doi.org/10.24093/awej/vol14no3.2

Muchibbuddin, M., & Zakiyyah, A. M. (2024). Artificial intelligence assistance in English for communication competence of English for specific purposes. *Insania: Jurnal Pemikiran Alternatif Kependidikan*, *29*(2), 236–256. https://doi.org/10.24090/insania.v29i2.11140

Murawski, L. M. (2014). Critical thinking in the classroom...and beyond. *Journal of Learning in Higher Education*, *10*(1), 25–30.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Ngo, T. N., & Hastie, D. (2025). Artificial intelligence for academic purposes (AIAP): Integrating AI literacy into an EAP module. *English for Specific Purposes*, *77*, 20–23. https://doi.org/10.1016/j.esp.2024.09.001

OECD - Organisation for Economic Co-operation and Development. (2023). *OECD Digital Education Outlook 2023: Towards an effective digital education ecosystem*. OECD Publishing.

Peláez-Sánchez, I. C., Velarde-Camaqui, D., & Glasserman-Morales, L. D. (2024). The impact of large language models on higher education: Exploring the connection between AI and Education 4.0. *Frontiers in Education*, *9*, 1392091. https://doi.org/10.3389/feduc.2024.1392091

Qasem, F., Ghaleb, M., Mahdi, H. S., Al Khateeb, A., & Al Fadda, H. (2023). Dialog chatbot as an interactive online tool in enhancing ESP vocabulary learning. *Saudi Journal of Language Studies*, *3*(2), 76–86. https://doi.org/10.1108/SJLS-10-2022-0072

Romadhon. (2024). ChatGPT for business emails: Insights from English for specific purposes students. *Lingua: Journal of Language, Literature, and Teaching*, *20*(2), 210–224. https://doi.org/10.34005/lingua.v20i2.4151

Sardi, J., Darmansyah, C., Candra, O., Yuliana, D. F., Habibullah, Y., Yanto, D. T. P., & Eliza, F. (2025). How generative AI influences students' self-regulated learning and critical thinking skills: A systematic review. *International Journal of Engineering Pedagogy*, *15*(1), 94–108. https://doi.org/10.3991/ijep.v15i1.53379

Shahrokni, S. A. (2009). Second language incidental vocabulary learning: The effect of online textual, pictorial, and textual pictorial glosses. *The Electronic Journal for English as a Second Language*, *13*(3), 1–17.

Silitonga, L., Wiyaka, W., & Prastikawati, E. F. (2024). Boosting students' ESP vocabulary by utilizing AI chatbot. *ETERNAL (English Teaching Journal)*, *15*(2), 275–283. https://doi.org/10.26877/eternal.v15i2.605

Sinkus, T., & Ozola, I. (2024). Engineering student perceptions of AI technology implementation in ESP. In *Proceedings of the International Scientific Conference Engineering for Rural Development 2024* (pp. 381–389). Latvia University of Life Sciences and Technologies. https://doi.org/10.22616/ERDev.2024.23.TF073

Smith, J. (2025). *Enhancing ESP student critical thinking skills and vocabulary acquisition through a Genai-based project*. OFS Home. https://doi.org/10.56297/vaca6841/ZDDD4208/HNVJ7828

Sye, D., & Statton Thompson, D. (2023). Tools, tests, and checklists: The evolution and future of source evaluation frameworks. *Journal of New Librarianship*, *8*(1). https://doi.org/10.33011/newlibs/13/9

UNESCO. (2023). *Guidance for generative AI in education and research*. United Nations Educational, Scientific and Cultural Organization. https://doi.org/10.54675/EWZM9535

Universidad Europea. (n.d.). *Metodología (Methodology)*. https://universidadeuropea.com/conocenos/facultad-medicina-salud-deporte-madrid/metodologia/

Wang, D., Liu, Y., Jing, X., Liu, Q., & Lu, Q. (2024). Catalyst for future education: An empirical study on the impact of artificial intelligence-generated content on college students' innovation ability and autonomous learning. *Education and Information Technologies*, *29*(1), 1–26. https://doi.org/10.1007/s10639-024-13209-6

Waziana, W., Andewi, W., Hastomo, T., & Hasbi, M. (2024). Students' perceptions of the impact of AI chatbots on vocabulary and grammar in EFL writing. *Register Journal*, *17*(2), 352–382. https://doi.org/10.18326/rgt.v17i2.352-382

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, *20*(2), 232–245. https://doi.org/10.64152/10125/66826

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal*, *104*(4), 646–667. https://doi.org/10.1111/modl.12671

World Economic Forum. (2024). *Shaping the future of learning: The role of AI in Education 4.0* (Insight Report). World Economic Forum.

Yu, L., & Zin, Z. M. (2023). The critical thinking-oriented adaptations of problem-based learning models: A systematic review. *Frontiers in Education*, *8*, Article 1139987. https://doi.org/10.3389/feduc.2023.1139987

Zhang, Z., & Huang, X. (2024). The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon*, *10*(1), Article e25370. https://doi.org/10.1016/j.heliyon.2024.e25370

Appendices

Appendix 1. Chatbot data evaluation table 1 (stage 2)

1. Which DESIGN ASPECT (ideas + authors; layout; organization of ideas, key language, style of language) are you searching for with the Chatbot?	2. Did the Chatbot give you information you can use for the design of your text? If 'Yes', continue to 3. If 'no', how did you decide on the criteria for the design aspect (e.g. used a search engine/ your own knowledge)? Please describe. Then go to 5. ii)	3. Name of the website/ source given by Chatbot with that design criteria information (copy + paste). Then, go to 4. If NO SOURCE is associated to the design criteria information, type 'NO SOURCE GIVEN' below. Then go to TABLE 2	4. Is it clear that the specific criterion you can use to design your multimodal text comes from this website/ source? If 'Yes', continue to 5. i) If 'no', go to TABLE 2	5. i) What information about the design criterion from the website/ source are you going to use? (Briefly describe or copy + paste) Then, go to 6. 5. ii) What specific design criterion have you decided to use if it has NOT BEEN GIVEN by the Chatbot (Briefly describe or copy + paste)	6. CRAAP Test (Currency, Relevance, Authority, Accuracy, Purpose) observations about the website/ source. (If you have decided your own criteria without using a source/ website, do not complete this information for the specific design criterion)

Appendix 2. Chatbot data evaluation table 2 (stage 2)

1. Which design aspect (ideas + authors; layout; organization of ideas, key language, style of language) is the information for?	Why are you here? 1. Because NO SOURCE is given by the chatbot for a specific design criterion we want to use. 2. Because it is NOT CLEAR that the specific design criterion we want to use comes from that source the chatbot has given us. (Type '1' or '2')	3. What is the design criterion you are going to use to create your text?

Appendix 3. Comparative analysis checklist (stage 4)

Design Aspect	Question	Yes	No
Ideas + Author	1. Are all the key ideas and/or authors in your stage 3 text also included in the stage 4 AI-revised text?		
	2. If 'no' for question 1, how many key pieces of information or authors are now missing? Please describe.		
	3. Are there any new ideas or authors included in the stage 4 AI-revised text which were not in your stage 3 text?		
	4. If 'yes' for question 3, are you happy with these suggestions of ideas or authors by the chatbot?		
	5. Have the ideas changed in meaning in any way in the stage 4 text?		
	6. If 'yes', please describe how they have changed		
Organization of Ideas	7. Has the organization of ideas changed in any way in the stage 4 text?		
	8. If 'yes', please describe how the organization of ideas has changed		
Key Language	9. Is there any topic-related vocabulary in the stage 4 text that is new and that you would also include in your finalized version of the text? (if **yes**, this new language must be included in the section '**Key language upgrades**' below this table.		
Style of Language	10. Is the style of language used in this stage 4 text as you desired. In other words, does it meet your specific design criteria		
	11. If 'no', describe what is wrong with the style of language		

Appendix 4. Instructions for Presentation 1

In the first presentation, you must **discuss/justify**:

- The type of multimodal text the group has chosen and the idea(s) to communicate in it.
- What you knew about the text type and sports science-related idea(s) you want to communicate BEFORE using AI (i.e., about the idea(s), layout, language, style of language, etc.)
- Your experience searching for key ideas and design/text content criteria and your experience with the search prompts you used. (Did you have to change prompts to get better ideas? Did you have to abandon AI and use just a normal Google search, etc.)
- The evaluation of the sources provided by the AI output or Google (if applicable) using the CRAAP test.
- The final text criteria regarding layout, idea organization, key language, style of language that you will use to create your text.

Appendix 5. Instructions for Presentation 2

In the second presentation, you must **discuss/justify using examples**:

- (Reminder of) the design criteria you based your text creation on.
- How the text was created. Did you use AI to write it? What AI tools did you use, and why? Did you have to change any stage 2 design criteria?
- Summary of the (comparative) analysis of the AI revised text:
    - Are the idea(s) between the texts the same? Are they expressed with the exact same meaning?
    - Are the desired text-creation criteria related to textual content (e.g. order of ideas, language style, etc.) of the text adhered to?
    - What have you learned about language use? What new important vocabulary related to the topic have you learned? What new language has been incorporated into your finalized text and why?

- Did the texts produced at stages 3 and 4 influence the finalized multimodal text (stage 5)? How? Justify your decisions.

Appendix 6. Quantitative question – CT skills 1

14. En general, creo que es importante aplicar el pensamiento crítico (utilizando habilidades analíticas y evaluativas) al resultado creado por la inteligencia artificial generativa, como un chatbot de IA. // In general, I feel it is important to think critically (using analytical and evaluative skills) about the output created by generative artificial intelligence, such as an AI chatbot.

- Estoy muy de acuerdo // I strongly agree
- Estoy de acuerdo // I agree
- Estoy en desacuerdo // I disagree
- Estoy muy en desacuerdo // I strongly disagree

Appendix 7. Quantitative question – CT skills 2

15. Gracias al *AI Project*, he aprendido que es importante utilizar habilidades analíticas y evaluativas (por ejemplo, decidir la veracidad de una fuente de información, tomar decisiones de diseño basadas en datos fiables/creíbles, contrastar diferentes tipos de vocabulario utilizado, etc.) cuando se utiliza inteligencia artificial generativa, como un chatbot // Because of doing the AI Project, I've learned it's important to use analytical and evaluative skills (e.g deciding the veracity of a source of information, making design decisions based on reliable/credible data, contrasting different types of vocabulary used etc.) when using generative artificial intelligence, such as a chatbot.

- Estoy muy de acuerdo // I strongly agree
- Estoy de acuerdo // I agree
- Estoy en desacuerdo // I disagree
- Estoy muy en desacuerdo // I strongly disagree

Appendix 8. Quantitative question – Vocabulary

10. Siento que participando en el *AI Project *he aprendido lenguaje nuevo y útil en inglés relacionado con el tema de mi texto multimodal* // * I feel I learned useful new language in English related to the topic of my multimodal text doing the AI Project.

- Estoy muy de acuerdo // I strongly agree
- Estoy de acuerdo // I agree

- Estoy en desacuerdo // I disagree
- Estoy muy en desacuerdo // I strongly disagree

Appendix 9. Qualitative question – CT skills 2

11. ¿Puede explicar la razón por la que has elegido la opción de la pregunta 10? // Can you explain the reason why you choose the option you did in question 10?

Appendix 10. Qualitative question – Vocabulary

16. ¿Puede explicar la razón por la que has elegido la opción de la pregunta 15? // Can you explain the reason why you choose the option you did in question 15?

Appendix 11. Qualitative question – Most valuable lesson learned

19. ¿Cuál crees que ha sido la habilidad o lección más valiosa que has aprendido al hacer el AI Project? // What do you feel was the most valuable skill or lesson you learned from doing the AI Project?