REVISTA DE EDUCAÇÃO A DISTÂNCIA E ELEARNING

ISSN 2182-4967

Critically engaging with AI: A scaffolding model for L2 Spanish writing

Engajamento crítico com a IA: Um modelo de andaime para a escrita em espanhol como L2

Ángela Luque-Giráldez[1]

Abstract

The rise of Generative Artificial Intelligence presents a challenge in language education: while offering immediate linguistic support, its uncritical use often bypasses the essential cognitive stages of the writing process. This paper presents a classroom-based pedagogical proposal and practitioner reflection developed at university level with advanced Spanish C1 learners, aimed at reimagining written assessments. The proposal is structured into three scaffolded phases: First, using artificial intelligence tools to map opposing arguments on a controversial topic; second, formulating a position paper through independent research; and third, engaging in metacognitive reflection on the limitations of technological support. This design positions artificial intelligence as a pre-writing scaffold within the Zone of Proximal Development, requiring the student to assume final responsibility for independent drafting and discourse. This work explores how this approach fosters digital literacy and critical thinking, offering a practical guide and instructional framework for language educators seeking to integrate technology ethically and effectively. It suggests that shifting artificial intelligence towards the initial planning stages preserves the iterative and personal nature of the writing process in the L2 classroom.

Keywords: Spanish as a second language; SLA; artificial intelligence; digital literacy; writing skills

Resumo

O surgimento da Inteligência Artificial Generativa apresenta um desafio na educação de línguas: embora ofereça suporte linguístico imediato, o seu uso acrítico frequentemente ignora as etapas cognitivas essenciais do processo de escrita. Este artigo apresenta uma proposta pedagógica baseada em sala de aula e uma reflexão sobre a prática docente, desenvolvidas a nível universitário com alunos de espanhol de nível avançado C1, visando reimaginar as avaliações escritas. A proposta está estruturada em três fases em andaime: primeiro, o uso de ferramentas de inteligência artificial para mapear argumentos opostos sobre um tema controverso; segundo, a formulação de uma tomada de posição através de investigação independente; e terceiro, uma reflexão metacognitiva sobre as limitações do suporte tecnológico. Este modelo posiciona a inteligência artificial como um andaime de pré-escrita dentro da Zona de Desenvolvimento Proximal, exigindo que o aluno assuma a responsabilidade final pela redação e pelo discurso independentes. Este trabalho explora como esta abordagem promove a literacia digital e o pensamento crítico, oferecendo um guia prático e um enquadramento pedagógico para educadores de línguas que procuram integrar a tecnologia de forma ética e eficaz. Isso sugere que a utilização da inteligência artificial para as fases iniciais de planeamento preserva a natureza iterativa e pessoal do processo de escrita na sala de aula de L2.

[1] Oxford Brookes University, aluque-giraldez@brookes.ac.uk, https://orcid.org/0000-0001-5464-1653

Palavras-chave: Espanhol como segunda língua; SLA; inteligência artificial; literacia digital; competências de escrita.

1. Introduction

The rapid integration of Generative Artificial Intelligence (GenAI) in higher education has posed significant challenges as well as opportunities for second language (L2) instruction. Although current studies praise its ability to expand vocabulary, reduce learner anxiety, and improve overall production (Barrot, 2023; Du & Daniel, 2024; Huang & Cassany, 2025; Luan, 2024; Warschauer et al., 2023), classroom implementations frequently reveal a tendency toward uncritical, task-oriented usage characterised by minimal metacognitive reflection (Huang & Cassany, 2025). Tempted by the speed of these tools, students bypass the process of writing all together, overlooking its different stages.

This paper originates from a pedagogical challenge encountered in advanced (CEFR C1) undergraduate Spanish modules. It was observed that students were increasingly relying on AI, not as a supportive tool, but as a complete replacement of their efforts. In high-level L2 contexts, where developing an individual voice and complex argumentation is key, the shift toward AI-generated drafts was turning assessments into a hollow exercise. It reached a point where it was no longer clear what was actually being assessed —the student's linguistic competence or their ability to prompt—, posing a challenge to assessment validity in higher education (Corbin, Dawson, & Liu, 2025).

Therefore, this study presents a praxis-oriented instructional design and practitioner reflection centred on re-humanising the writing process. Drawing on cognitive process writing theory (Flower & Hayes, 1981) and Vygotskian sociocultural scaffolding (Vygotsky, 1978; van de Pol et al., 2010), this intervention constrains AI interaction strictly to the pre-writing phase across three structured stages, leaving final drafting (translating) and revision to independent, unassisted assessment. To guide this proposal and reflection, this work addresses two research questions:

- RQ1: How can an instructional design model structure GenAI as a pre-writing scaffold to promote critical engagement and argumentative depth in advanced L2 Spanish learners?
- RQ2: What evidence of student authorship appears after restricting AI access strictly to the pre-writing phase?

2. Theoretical framework

2.1. Writing as an iterative process

To understand how AI challenges traditional learning, we must define writing not as a final product, but as a cognitive journey. According to the classic model by Flower and Hayes (1981), writing is a recursive and non-linear process involving three main operations: planning, translating (turning ideas into text), and revising.

GenAI presents a distinct risk to this process-oriented approach. By providing a polished text instantly, the technology encourages students to bypass the translating stage. This bypass effect is a form of cognitive offloading (Risko & Gilbert, 2016) that eliminates the moment where students struggle with the language, which is precisely where deep learning occurs. Crucially, unguarded and continuous access to automated generation can leave learners worse off once technological scaffolding is removed, as fundamental capabilities struggle to consolidate when cognitive tasks are outsourced (Bastani et al., 2025). Pedagogical interventions must therefore ensure that these iterative stages are safeguarded rather than replaced.

This imperative also explains why GenAI is deliberately excluded from the translating and revision phases in the present proposal. While existing literature demonstrates that GenAI models can support every stage of writing —from planning to collaborative text generation and feedback (Luan, 2024)— allowing AI access during revision introduces significant pedagogical drawbacks. As Alharbi (2023) notes in reviewing automated feedback studies, L2 learners frequently treat automated revision suggestions as passive proofreading cues, implementing corrections “with no or little cognitive processing”. Restricting GenAI strictly to the pre-writing stage addresses this limitation directly: it leverages AI for initial ideation and conceptual framing, helping students overcome the “blank page” problem, while reserving translating, critical evaluation, and final revision unassisted, thereby protecting the validity of independent L2 language production.

2.2. AI as scaffolding within the Zone of Proximal Development

Scaffolding refers to “the temporary support provided for the completion of a task that learners otherwise might not be able to complete” (van de Pol et al., 2010). This study builds on Vygotsky’s (1978) concept of the Zone of Proximal Development (ZPD): the distance between what a learner can do independently and what they can achieve with the guidance of a “more knowledgeable other” (MKO), traditionally, a teacher or a peer (Vygotsky, 1978). Following van de Pol et al. (2010), effective scaffolding can be assessed against three core pedagogical criteria: contingency (tailoring support to learner needs), fading (gradually withdrawing assistance), and transfer of responsibility (enabling independent learner mastery). In this study, fading is operationalised as a full withdrawal of GenAI support before composition rather than a gradual reduction (see Section 2.1).

GenAI can contribute to this process because its conversational capacity allows it to produce coherent texts, lists of topics and outlines (Barrot, 2023), which learners can then critically evaluate, cross-examine, and refute. However, GenAI cannot function as the sole MKO. Stojanov (2023) autoethnographic account reports that while ChatGPT can simulate expert discourse, its outputs are frequently superficial, logically inconsistent or contradictory. Furthermore, one can frequently overestimate their own comprehension when interacting with polished AI responses, making unmediated tool reliance counterproductive (Stojanov, 2023). Because GenAI operates on probabilistic pattern recognition as opposed to genuine intentionality, its capacity to navigate the sociocultural and pragmatic dimensions of language is fundamentally limited (Godwin-Jones, 2024). Therefore, agency cannot be handed over to the machine. It must remain distributed across the triadic interaction of learner, teacher and tool, with advanced L2 learners being equipped to critically evaluate machine output under instructor supervision (Godwin-Jones, 2024; Stojanov, 2023).

For this reason, the scaffolding in this study is supplied by the instructional design as a whole (further explained in Section 3.3.), not by GenAI alone:

1. Preparatory support (Workshop and feedback sessions): A guided instructor workshop (during week 11) and targeted feedback (during week 12) help students diagnose gaps and logical inconsistencies in the AI output. As these are group sessions on a shared topic, contingency is limited.

2. Initial scaffolding (Phase 1): GenAI supplies a preliminary ideational map, establishing a baseline for analysis.

3. Withdrawal and transfer of responsibility (Phases 2 & 3): As students move from AI interrogation to independent research and unassisted translating, technological support is withdrawn.

By withdrawing technological assistance prior to composition, the design aims to promote final authorial agency and linguistic execution.

2.3. The UNESCO AI competency framework

Finally, the goal of integrating AI into the curriculum is to move beyond mere functional use toward AI literacy. The UNESCO AI Competency Framework for Students (Miao & Shiohira, 2024) provides a roadmap for this progression, emphasising that AI competency involves not only technical skill, but also understanding the ethical implications and limitations of the technology.

The framework describes three progressive levels of mastery: Understand, Apply and Create. Although it was originally developed to guide school curricula, it serves in this study as a conceptual benchmark for mapping higher-order digital competencies not as an empirically validated assessment rubric. For C1 Spanish students, the objective is to reach the Apply level, where learners bring a human-centred and ethical perspective to

the critical assessment and practical use of AI tools: interrogating AI output, identifying biases or hallucinations, and cross-referencing information with authentic documents. By framing the pedagogical design around this competency, the study aims to move learners from passive users to critical authors who can navigate a technology-mediated linguistic landscape.

2.4. Existing research

Recent empirical work on GenAI in L2 writing reflects both transformative potential and notable pedagogical friction. A technology review highlights ChatGPT's potential to offer timely feedback, topic generation and outlining support to L2 writers, while warning of fabricated references, overreliance and a limited capacity to reproduce individual voice (Barrot, 2023). Warschauer et al. (2023) add that its value for L2 writers is shaped by contradictions between imitation and plagiarism risk, unequal access, and the foundational writing skills needed to use the tools well. These tensions are particularly acute in hybrid workflows where human and machine text are co-constructed, complicating traditional definitions of revision, ownership and authorship (Alharbi, 2023). Within Spanish L2 research, Luan (2024) compared a teacher-guided ChatGPT workflow across planning, translating and revision with traditional instruction among Chinese B2-level university students. The ChatGPT group scored higher in grammar, lexical and originality, with more moderate gains in coherence, although the study also reported direct copying of AI output and uneven use of its suggestions. Because ChatGPT was used throughout the writing process, the study cannot show what each stage contributes. Instead, the present proposal restricts GenAI to the pre-writing phase, positioning the model as an ideational partner rather than a co-writer.

This restriction responds to emerging evidence of the cognitive costs of unrestricted AI use. In a randomised field experiment with nearly 1,000 secondary school mathematics students, unrestricted AI access improved practice performance by 48% but was followed by 17% lower scores on a later unassisted exam than the control group. A version designed to give hints rather than answers largely removed this negative effect, although it did not improve exam performance either (Bastani et al., 2025). Qualitative evidence from L2 Spanish points in a similar direction, although it concerns reflection rather than performance: Huang and Cassany (2025) analysed 370 prompts from ten Chinese postgraduate learners using ChatGPT for a week without teacher guidance and found mostly single, task-oriented interactions (303 simple interactions against 31 interaction cycles), which they link to limited reflection.

However, there is a gap in L2 research on phase-restricted pedagogical interventions, particularly in advanced undergraduate Spanish composition. The literature largely examines AI deployed across the writing process and, to our knowledge, little work has studied pre-writing-exclusive designs. Barrot (2023) describes a classroom sequence in which ChatGPT supports brainstorming and outlining but is not allowed during drafting, yet he reintroduces it for feedback and editing and reports no data. This proposal

addresses that gap by confining GenAI to conceptual brainstorming and structural scaffolding before translating.

3. Instructional design

The intervention took place within the context of a Final Year (Level 6) Spanish module at undergraduate level, which extends over 12 weeks. While the module covers a broad range of socio-cultural topics through traditional seminars, this AI-mediated intervention was introduced as a core assessment component to address the growing challenge of academic integrity in L2 writing.

A characteristic challenge of advanced language modules at this level is the fluctuating and often small student cohort. This trend is primarily due to the Year Abroad structure, where many students study abroad for an academic year, resulting in small-group environments for those remaining or returning.

This framework was developed over two consecutive academic years, across two cohorts totalling six students. Participants were advanced L2 Spanish learners at CEFR C1 level: four L1 English speakers and two L1 French speakers. While a sample size of six participants is small for quantitative testing, it reflects the reality of advanced language instruction, offering an opportunity for micro-pedagogical analysis.

3.1. Instructional framework and ethical use

Before engaging with the task, all students completed the university's mandatory Academic Integrity online module, which established a common ethical baseline by defining academic integrity as doing your own work and giving credit to others. Students were also instructed on three pillars of the university's AI policy: engagement of thinking (developing the confidence to draw personal conclusions), transparency (specifying the tools and prompts used) and accountability (submitting AI-generated work constitutes an academic integrity breach).

The intervention set an explicit rule: GenAI was permitted exclusively for pre-writing brainstorming and prohibited during translating and revision, so that the submission would reflect the learner's own L2 production. Such rules function primarily as discursive changes, that is, rules and guidelines whose success depends on student compliance (Corbin, Dawson, & Liu, 2025). Because discursive changes alone cannot ensure that a take-home submission reflects the learner's own production, the design complemented them with process-oriented measures that make AI use more visible: students pasted raw AI outputs into task appendices, submitted a detailed AI declaration, took part in a guided instructor workshop, and completed a graded metacognitive reflection evaluating the tool's limitations. In Corbin, Dawson e Liu (2025) terms, these measures remain largely discursive, since declarations and documentation still rely on student honesty, although the emphasis on process over product is consistent with their recommendation to reorient assessment from output to process. As the final assignment was completed

as take-home coursework, they reduce, but cannot eliminate, reliance on student compliance.

3.2. Formative workshop

After the mandatory training, students worked through a two-stage formative process:

- Week 11, guided workshop: Students participated in an in-class 2-hour guided workshop to analyse the limitations of GenAI using a standardised worksheet and a shared sample topic (Social media censorship). Working from a suggested prompt (*Presenta los principales argumentos a favor y en contra de la censura en redes sociales desde una perspectiva ética, legal y social*), they practised pasting raw AI outputs, evaluating argument gaps, integrating Harvard-referenced academic sources, drafting an unassisted introduction, and completing a trial reflection.

- Week 12, formative feedback: After submitting their workshop worksheets, they received instructor feedback in the following session, which clarified expectations regarding source integration, critical analysis and linguistic register.

3.3. Architecture of the assignment

Following the formative workshop and feedback cycle, students completed their final summative assignment independently as take-home coursework during the examination period. Replicating the framework practiced during the workshop, it comprised the same three distinct components:

- Phase 1: Argumentative mapping with AI consultation. Students chose a complex socio-cultural topic (e.g., the ethics of euthanasia or animal experimentation) and used a tool such as Copilot or Gemini to generate a bulleted list of pros and cons, helping them overcome the “blank page” anxiety. They pasted this raw AI output into their submission, excluded from the word count.
- Phase 2: Critical analysis and argumentation (800-1,000 words). Students wrote a position paper analysing, refuting or expanding upon the AI-generated arguments, synthesising information and building a coherent argument in Spanish in an academic style and register. Every claim or counter-argument had to be grounded in academic sources (books, journals, empirical studies) cited in-text and listed in a Harvard style reference list.
- Phase 3: Metacognitive reflection (400-700 words). The final component of the assessment is a reflective piece based on three specific sections adapted from Gibbs’ Reflective Cycle (1988):
    - Learning obtained: What did you discover when contrasting the AI-generated viewpoints with academic literature?

- Benefits and limitations: What advantages did AI offer in this task, and what risks or flaws emerged when relying on it?
- Future application: How can these insights be applied to future academic research and writing?

This reflection supports progression toward higher levels of UNESCO's AI competency framework (Miao & Shiohira, 2024), and encourages students to reflect on their own learning process and the ethics of technology.

3.4. Evaluation and marking criteria

To align the assignment with a process-oriented design, the marking criteria were redesigned around five categories: Understanding of the task (20%), Research and resources (25%), Reflection (15%), Language and presentation (15%) and Opinion and argumentation (25%).

The weighting aims to counteract the superficiality of AI-generated content. Allocating half of the grade to Research and resources and Opinion and argumentation encourages students to test every AI-generated claim against academic evidence, shifting cognitive effort from sentence construction to information synthesis and fact-checking. Reducing the weight of Language and presentation to 15% is not a dismissal of grammar, but a deliberate decision to lower the incentive to outsource linguistic polishing to AI and to prioritise complex argumentation over error-free output. Reflection carries a lower weighting (15%) because it is the shortest component. These are the five pillars as follows:

- Understanding of the task (20%): Assesses whether students complete all components and follow the instructions, including the permitted use of AI, while their own voice remains evident. Here, "use of AI tools" refers to the permitted brainstorming phase. At the highest band (A), success is defined by effective use of AI tools to support the work without displacing the student's voice; in lower bands, limited AI use or a contribution that is difficult to detect is penalised.
- Opinion and argumentation (25%): Assesses a well-founded and clear opinion maintained consistently, with the student driving the logical development of the argument.
- Reflection (15%): Evaluates AI literacy through the application of Gibbs' reflective framework. Students must move beyond functional use to offer meaningful insights into their interaction with the tool, demonstrating critical awareness of its limitations and its impact on their creative process.
- Language and presentation (15%): Assesses the ability to maintain a consistent academic register while navigating the complexities of advanced Spanish.
- Research and resources (25%): Rewards critical engagement with a wide range of relevant, external academic sources that inform the student's arguments.

Beyond grading, this rubric serves as the primary qualitative framework for analysing student outcomes in this study. The evidence in Section 4 is structured directly around

these pillars to examine how the design shaped student authorship, research behaviour, and metacognitive engagement. Table 1 presents an excerpt of the descriptors for high (Grade A), medium (Grade C), and low (Grade RC) bands.

Table 1.
High, medium and low bands from the marking criteria.

	Understanding of the task (20%)	**Opinion and argumentation (25%)**	**Reflection (15%)**	**Language and presentation (15%)**	**Research and resources (25%)**
A 70-74%	The project is well-executed, with all key parts completed and a clear understanding of the task. Objectives are met effectively. Strong use of AI tools to support the work, though individual voice remains evident. Instructions are followed.	Well-founded and clear opinion. The argument is coherent and logically developed, with a high level of intellectual rigor. The position is evident and maintained consistently.	Strong and thoughtful reflections on the topic, showing a good level of critical thinking. The reflective framework is well understood and applied. The reflection offers meaningful insights.	High levels of fluency, style and register. The writing demonstrates a strong individual voice. Great use of grammar, punctuation, vocabulary and spelling. Rich and varied vocabulary. The content is very well-organised, with a clear and logical structure.	Evidence of extensive research into a wide range of relevant sources, informing arguments. There is clear critical engagement with the sources. The research is detailed and covers all key aspects of the topic. Most sources are cited and referenced accurately.
C 50-54%	The project shows limited understanding of the task, with key parts incomplete or missing. The objectives are only partially met, and the use of AI tools is limited and it is difficult to detect an individual voice.	Opinion is somewhat clear but lacks substantial support. The argument is weak or lacks coherence in parts. The position is stated but not maintained.	Basic reflections with limited critical thinking. The reflective framework is acknowledged but not fully applied. The reflection lacks depth and could benefit from better analysis.	Basic fluency and range of style and register is acceptable. Some difficulty in expressing complex ideas but the intention can be perceived. Standards of grammar, vocabulary, spelling and punctuation may be uncertain. Some errors interfere with communication. The structure is inconsistent.	Evidence of research into a basic range of relevant sources which are not well-integrated into the discussion. The research covers only some aspects of the topic and lacks depth. Some sources are cited and referenced with errors or are missing.
RC 30-39%	The project shows almost no understanding of the task. Objectives are not met, and the work is largely incomplete. No effective use of AI tools and no evidence of individual contribution. Instructions are not followed.	Minimal opinion present, with little or no argument. The position is unclear or absent, and the argument lacks coherence. The content is disorganised and lacks structure.	Reflections are very minimal and show little to no understanding of the topic or the reflective process. There is no clear application of the reflective framework, and the reflection lacks depth and coherence.	Very poor fluency, style and register clearly creates difficulties in expressing ideas. Standards of grammar, vocabulary, spelling and punctuation are erratic, and errors make communication near impossible. The content has almost no structure. Language may appear erratic or mismatched with the student's previous work.	Almost no use of resources. There is no critical evaluation of sources. The research is extremely superficial or irrelevant, missing nearly all relevant aspects of the topic. References are mostly incorrect or absent

Source: Author's original work.

3.5. Data and analysis procedure

Data were collected from the complete submission of the six advanced L2 Spanish students. These were standard coursework artefacts gathered as part of routine curricular evaluation and anonymised before analysis. The qualitative thematic analysis focused on student-AI interaction logs, independent drafts, and reflective entries across three stages:

1. Extraction and anonymisation: Texts were extracted from student submissions and catalogued with anonymised identifiers (Student 1 through Student 6).

2. Coding: Texts were reviewed and coded according to the five assessment pillars in Section 3.4, with particular attention to discrepancies between the raw AI outputs in appendices and students' unassisted L2 production (e.g., lexical choices, grammatical errors, and argumentative deviations).

3. Illustrative selection: Excerpts in Section 4 were selected through purposive sampling to demonstrate representative patterns of student agency, critical evaluation of AI outputs, and persistent L2 linguistic features across performance bands.

Because this is a practitioner reflection on routine coursework, coding was carried out by the author in their dual role as module instructor and marker, therefore the findings are exploratory and not generalisable.

4. Results

To evaluate the pedagogical outcomes of this three-phase model, the six students' submissions were analysed against the five pillars of the marking rubric (Section 3.4). Across two cohorts, students selected socio-political topics, including animal experimentation, euthanasia, illegal immigration and cultural appropriation. Given the qualitative nature of a practitioner reflection, the results offer a micro-pedagogical examination of student production rather than a statistical generalisation. Counts (e.g., 2/6) indicate how many of the six submissions contained at least one instance of a coded pattern. Quoted examples were selected purposively to illustrate each pattern (Section 3.5) and are not exhaustive, so fewer students may be named in the examples than in the count.

4.1. Understanding of the task (20%)

All participants (6/6) successfully completed all three phases of the assignment: the raw AI argument map (Phase 1), an independent research-backed position paper (Phase 2), and a reflective commentary based on Gibbs' cycle (Phase 3). In their productions, students generally treated the AI output as a starting point rather than as a template for their final structure, consultation and independent authorial voice. However, compliance

with the pre-writing restriction cannot be fully verified from the submissions alone (see the unverified reference in 4.3).

4.2. Opinion and argumentation (25%)

This band required a coherent and well-founded thesis that challenged or refined the initial AI-generated points:

- Critical interrogation of AI logic (2/6 students): In analysing the AI-generated map on illegal immigration, Student 3 identified an internal contradiction: the tool initially labelled irregular immigration as an unmitigated "crisis", but later acknowledged that crisis framings are often media constructions. Pointing out this inconsistency enabled Student 3 to build an independent argument around political discourse. Similarly, Student 5 refuted the AI's oversimplified claim that no alternatives to animal testing exist by researching and integrating the scientific framework of the "3 Rs" (Reduction, Refinement, Replacement).

- Navigating nuance (1/6 students): Student 1 compared Belgian legislative controls on euthanasia with French healthcare classifications (ALD), but could not locate the Spanish equivalent.

- Argumentative density (1/6 students): In the same submission, Student 3's argumentation reached only the C band: the paper accumulated theoretical concepts and citations (e.g., "capitalist self-determination", attributed to Adamson, Triadafilopoulos and Zolberg) without explaining or analysing them. Performance therefore varied across components within a single submission.

4.3. Research and resources (25%)

All six students integrated external academic references into Phase 2. An audit of their bibliographies revealed variations in source verification and citation mechanics:

- Unverified reference (1/6 students): Student 6 cited Schrecker (1998) to support a case study on censorship, but the source could not be located in the university's physical holdings or digital databases. Unavailability does not establish that the reference is fabricated, but it raises two possibilities: an unchecked reference carried over from the Phase 1 AI output or use of AI beyond the permitted pre-writing phase. This is consistent with the risk described by Stojanov (2023), whose autoethnographic account of learning with ChatGPT reports missing contradictory answers and overestimating the author's own understanding, and with the documented tendency of ChatGPT to produce non-existent references (Barrot, 2023).

- Empirical claims without in-text citations (1/6 students): Student 1 incorporated specific figures (e.g., 3.6% of deaths in Belgium; 780€ average annual patient costs in France) without in-text citation. The submission does not show whether

these were found independently or supplied by the AI, which suggests a difficulty in distinguishing background knowledge from claims requiring attribution.

- Manual citation errors (3/6 students): Students included authors' first names or initials in in-text citations ("John Keown, 2002" [Student 1]; "Velázquez Cuj, O. N. G. y Castillo Burelo, C. A., 2025" [Student 2]). Student 2 also listed the director of a dissertation as a co-author (Carrasco Barranco and Lafuente Caballero, 2019). Student 5 included a in-text citation in English (Diaz and Rabadán, 2011). These errors are consistent with manual reference assembly rather than automated tools.

4.4. Language and presentation (15%)

Restricting AI to pre-writing means students' texts may retain linguistic errors. Five of the six submissions (Students 1-5) contained errors displaying authentic C1 interlanguage features, which is consistent with students producing independent prose:

- Morphosyntactic and agreement errors (4/6 students): gender disagreement ("a un grupo que a otra" [Student 2]), plural agreement ("personas son omnipresente" [Student 2]); verb agreement ("métodos diferentes que puedes reemplazar" [Student 5]); and mood ("fomenta que usamos" [Student 5]).
- Prepositions and L1 transfer (3/6 students): Student 1 (L1 French) produced "están pensado a legalizar" and "triunfan a establecer" which mirror French *penser à* and *réussir à*. The student also omitted a subordinate conjunction ("creo [que] decidir..."). Student 4 (L1 English) struggled with relative connectors ("por el cual derecho") and preposition choice ("derecho fundamental a tomar decisiones").
- Lexical choice and register: Students 2 and 3 made questionable lexical or register choices, such as "recursos… ricos" instead of "recursos de calidad" (Student 2) and "un poquito" instead of "un poco" (Student 3).

4.5. Reflection (15%)

Student reflections (Phase 3, Gibbs' cycle) show emerging critical awareness of AI's usefulness and limits, around four themes:

- Efficiency and depth (5/6 students): Students 1, 3, 4 and 6 reported that GenAI helped with initial ideation and provide a framework, but others found the output superficial [Student 4] or too general [Student 5 and 6].
- AI as a complementary tool (4/6 students) Student 2 wrote "Es una falacia asumir que [la IA] compite con fuentes académicas, es mejor usar motores de búsqueda tradicionales". Student 3 stressed the need for caution: "Es necesario ser prudente y mantener cierta preocupación por la posibilidad de recibir información errónea". This emerging student awareness is consistent with Godwin-Jones (2024), who argues that because GenAI operates on probabilistic modelling rather than genuine domain understanding, it should serve as a secondary

scaffolding tool within a distributed agency shared among student, instructor and tool.

- Cognitive offloading (2/6 students): Student 3 noted that delegating definitions and reformulations to AI freed time for analysis, and highlighted the importance of prompting in a way that keeps the AI as support: “Es importante saber cómo formular preguntas que no exijan una reflexión completa por parte de la IA, para que solo sirva como apoyo”. Student 2 agreed that AI should be used as an additional source of information: “No es cuestión de usar la ia como fuente única, sino como una adicional”.
- The cultural boundaries (1/6 students): Student 2 observed a “falta de experiencia cultural necesaria para entender estos temas”.

5. Discussion

Regarding RQ1, the design operated as a teacher-mediated scaffold: students used AI as an initial scaffold rather than as a standalone MKO. In line with the scaffolding criteria discussed in Section 2.2 (van de Pol et al., 2010), GenAI provided the preliminary ideational mapping that helped students approach multifaceted debates. This support was then withdrawn before translating, and responsibility for the final text was transferred back to the learner. Minor surface-level slips in final submissions (such as English conjunctions in bibliographies or formatting oversights) are consistent with manual work after this withdrawal, though they do not prove it. Conversely, the unverifiable reference in Student 6’s paper, whatever its origin, is a reminder that AI can produce non-existent references (Barrot, 2023; Warschauer et al., 2023) and shows why students need to corroborate sources. This suggests that the instructor’s role may need to shift from grammar correction towards verifying sources and auditing citations.

Regarding RQ2, the findings of this practitioner reflection suggest that the presence of authentic human error is consistent with independent writing rather than uncritical automated text generation. The surface-level grammatical and syntactic errors recorded in Section 4.4. are not simple failures of knowledge; they are compatible with the iterative and non-linear nature of writing (Flower and Hayes, 1981). Debating complex topics in an L2 at C1 level is cognitively demanding, so when students moved away from the AI-generated outlines to express their own researched opinions, their attention may have shifted from form towards content synthesis and argumentative development. This is an interpretation rather than an observed finding, since the artefacts do not show students’ composing processes, but it could help explain the uneven profile observed in some submissions, where sophisticated, specialised vocabulary appears alongside basic gender agreement or prepositional slips. Such errors can be understood as a form of authorial friction: visible traces of an L2 learner composing independently. However, these errors cannot serve as a method of detecting AI use. AI detectors frequently misclassify non-native English writing (Liang et al., 2023), although this evidence comes from English and has not been tested for Spanish. Warschauer et al. (2023) also note that L2 writers risk accusations of dishonesty both when the gap between their own writing and AI text is visible and when their writing seems unusually good, and Luan

(2024) likewise reports, in a Spanish L2 study, that markers found some texts hard to grade because passages seemed to exceed students' usual level. Any concern is referred to the institution's academic integrity procedures, where the student can explain their process.

Also, relating to RQ1, the reflections and some of the position papers suggest an emerging critical AI literacy, corresponding to the Apply level of the UNESCO AI Competency Framework for Students (Miao & Shiohira, 2024). The initial AI output offered a general map of arguments, and students' own research provided the evidence required to argue a position. When learners identified contradictions within the AI's response (Student 3), challenged its claims with their own research (Student 5) or compared national policies (Student 1), they were exercising the kind of critical evaluation this competency implies, which corresponds to the "corroborate" element of the AI literacy framework of Warschauer et al. (2023). Some students also recognised the limits of AI models: as Student 2 noted, AI lacks cultural experience necessary to understand certain topics. These observations are tentative, but they suggest that designs which require students to evaluate and move beyond machine-generated outlines may support the development of a more individual authorial voice.

Several limitations must be acknowledged:

- Sample size and design scope: The intervention involved a small cohort across two module runs without a control or comparison group. Consequently, the findings are qualitative, descriptive, and exploratory, not statistically generalisable.
- Unmonitored assessment environment: The task relied on take-home coursework. In the terms of Corbin, Dawson e Liu (2025), the restriction on AI use is largely discursive: students were told when AI was permitted, but nothing structurally prevented its use at later stages, so the submissions cannot confirm how strictly each student kept to the permitted phase. Moreover, where legitimate brainstorming ends and outsourced thinking begins is itself contested among students and staff (Corbin, Dawson, Nicola-Richmond, et al., 2025).
- Phase-restricted focus: GenAI was confined to pre-writing. The study does not evaluate AI use during translating or revision, which other research has examined (e.g., Alharbi, 2023, for automated writing tools; Luan, 2024, for ChatGPT in Spanish). Nor does it teach students to write with AI at the composition stage, which Warschauer et al. (2023) consider part of AI literacy.
- Abrupt withdrawal: Withdrawing AI support in a single step, rather than gradually, may be harder for some learners.
- Dual instructor-researcher role: The module instructor also coded and marked the data, which may introduce subjectivity (see Section 3.5).

Future research could compare take-home submissions with invigilated, in-class writing to examine how controlled conditions affect argumentative depth, L2 accuracy and

critical engagement with the tool. It could also test phase-specific variants, comparing AI feedback during revision with pre-writing-only designs in advanced L2 classrooms.

6. Conclusions

This instructional design and practitioner reflection set out to examine how GenAI can be structured as a pre-writing scaffold to promote critical engagement and argumentative depth in advanced L2 Spanish learners (RQ1), and what evidence of student authorship appears when AI access is restricted to that phase (RQ2).

In response to RQ1, a three-phase model that confines GenAI to argument mapping, withdraws it before translating, and closes with a metacognitive reflection offers a workable structure. In this small cohort it was associated with some critical engagement with AI output, although argumentative depth was uneven. In response to RQ2, students' unassisted texts carried traces of independent L2 production, but under take-home conditions this evidence is indicative rather than conclusive.

These answers point to a broader implication. In the age of GenAI, academic integrity in the C1 classroom may depend more on transparency and process than on the polish of the final product, and a polished essay is no longer sufficient evidence of a learner's own competence. L2 assessment in higher education could therefore continue moving towards a process-oriented model (Corbin, Dawson, & Liu, 2025), bearing in mind that such a shift changes what is being assessed. Ultimately, the goal of language instruction is to guide students towards using AI as a starting point, and to value the critical and linguistic growth that this independent work makes possible.

Note on the use of AI

GenAI tools were used during the preparation of this manuscript for assisting with translating the title and abstract into Portuguese and for general grammatical editing and language refinement of the author's original prose. No AI tools were used to generate research content, analyse data or draw conclusions.

Bibliography

Alharbi, W. (2023). AI in the foreign language classroom: A pedagogical overview of automated writing assistance tools. *Education Research International, 2023*, Article 4253331. https://doi.org/10.1155/2023/4253331

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, Article 100745. https://doi.org/10.1016/j.asw.2023.100745

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: Why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education, 50*(7), 1087-1097. https://doi.org/10.1080/02602938.2025.2503964

Corbin, T., Dawson, P., Nicola-Richmond, K., & Partridge, H. (2025). 'Where's the line? It's an absurd line': Towards a framework for acceptable uses of AI in assessment. *Assessment & Evaluation in Higher Education* 1–13. https://doi.org/10.1080/02602938.2025.2456207

Du, J., & Daniel, B. K. (2024). Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Computers and Education: Artificial Intelligence*, 6, Article 100230. https://doi.org/10.1016/j.caeai.2024.100230

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365-387. https://doi.org/10.2307/356600

Gibbs, G. (1988). *Learning by doing: A guide to teaching and learning methods*. Further Education Unit, Oxford Polytechnic.

Godwin-Jones, R. (2024). Distributed agency in language learning and teaching through generative AI. *Language Learning & Technology, 28*(2), 5-30. https://doi.org/10.64152/10125/73570

Huang, S., & Cassany, D. (2025). Spanish language learning in the AI era: AI as a scaffolding tool. *Journal of China Computer-Assisted Language Learning.* https://doi.org/10.1515/jccall-2024-0026

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

Luan, Y. (2024). Colaboración humano-máquina en el proceso de enseñanza-aprendizaje de la escritura en español: El impacto de ChatGPT en el contexto educativo de universitarios chinos. *RLA. Revista de Lingüística Teórica y Aplicada, 62*(2), 13-35. https://doi.org/10.29393/RLA62-1CHYL10001

Miao, F., & Shiohira, K. (2024). *AI competency framework for students*. UNESCO. https://doi.org/10.54675/JKJB9835

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676-688. https://doi.org/10.1016/j.tics.2016.07.002

Stojanov, A. (2023). Learning with ChatGPT 3.5 as a more knowledgeable other: An autoethnographic study. *International Journal of Educational Technology in Higher Education, 20*, Article 35. https://doi.org/10.1186/s41239-023-00404-7

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271-296. https://doi.org/10.1007/s10648-010-9127-6

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of English as a second language. *Journal of Second Language Writing*, 62, Article 101071. https://doi.org/10.1016/j.jslw.2023.101071

Submitted 15/02/2026
Approved 12/10/2026
Published 10/11/2026

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.