education sciences

MDPI

Article

Chatbot-Supported Written Mediation and Pluricultural Competence in Adult EFL: An Exploratory Study in Official Language Schools

Esther Cores-Bilbao [1,*] and María-del-Carmen Méndez-García [2]

[1] Faculty of Humanities and Social Sciences, Universidad Isabel I, 09003 Burgos, Spain
[2] Department of English Philology, Universidad de Jaén, 23071 Jaén, Spain; cmendez@ujaen.es
* Correspondence: esther.cores@ui1.es

Abstract

This exploratory study examines whether chatbot-mediated written interaction supports adult B2 English learners' performance in online interaction, pluricultural competence, and mediation in Official Language Schools (OLS) in Spain. The intervention was built around a fictional-culture scenario in which learners had to resolve a cultural misunderstanding between a Spanish visitor and a host from an invented culture. In the experimental condition, students interacted with a chatbot previously configured with information about the fictional culture; in the control condition, students worked in pairs in a chatroom, with one peer acting as the cultural expert. Interaction texts were independently rated by two researchers using a Common European Framework of Reference for Languages (CEFR) Companion Volume-informed rubric. The dataset comprised 16 learners in the control group and 24 in the experimental group, each rated by two evaluators. Inter-rater reliability reached acceptable levels for all aggregated dimensions, with ICC(2,1) values above 0.70. Mann–Whitney U tests showed no significant between-group differences in online interaction or pluricultural competence, whereas the chatbot-supported condition, which included sustained-questioning scaffolding, was associated with significantly higher mediation scores. The findings suggest that chatbot use may be pedagogically promising for mediation-oriented writing tasks, although the evidence should be interpreted cautiously because the study is exploratory, the sample is small, and the scenario relied on a fictional cultural frame.

Keywords: adult EFL; chatbot-mediated interaction; CEFR Companion Volume; fictional culture; mediation; online interaction; Official Language Schools; pluricultural competence; exploratory study

Academic Editors: Erin R. Pletcher and Travis R. Pollen

Received: 24 March 2026
Revised: 14 May 2026
Accepted: 22 May 2026
Published: 27 May 2026

1. Introduction

Generative artificial intelligence (GenAI) and large language model (LLM) chatbots have rapidly gained prominence in language education. Reviews and meta-analytic work suggest that these tools can support language learning, but they also converge on a more cautious conclusion, namely that their pedagogical value is context-dependent rather than automatic. Systematic reviews of GenAI and ChatGPT in language education similarly show a rapidly expanding empirical base, while also emphasising variation in research designs, learner populations, target skills, and pedagogical implementations (Lee et al., 2025; Li et al., 2024, 2025). Earlier reviews of chatbot-supported language learning highlighted

technological, pedagogical, and social affordances, while also noting design limitations and the need to understand how human and non-human interlocutors shape the learning process differently (Huang et al., 2022; Ji et al., 2023).

This need for caution is similarly reflected in empirical studies. Tutton and Cohen (2025), for example, argue that AI requires careful pedagogical orchestration if it is to support language learning, since it cannot replicate the distinctively social interactivity of the language classroom. Other studies report positive outcomes for chatbot-assisted conversational English (Alenezi & Alenezi, 2025), teacher interest in AI-supported flipped English classrooms (Ling, 2025), and AI-mediated informal digital learning in ESL writing (Nguyen et al., 2025), yet these findings also reflect the heterogeneity of aims, contexts, and target constructs in the literature (Wiboolyasarin et al., 2025). Productive engagement with GenAI also depends on more than tool availability, extending to learner agency (Du & Alm, 2024), critical evaluation of AI output, and the ability to integrate AI strategically into communicative activity (Alm, 2024). The issue, then, is to identify the conditions under which AI may support specific dimensions of communicative performance.

This question gains importance in light of the CEFR Companion Volume (CEFR/CV), which places mediation, online interaction, and plurilingual/pluricultural competence at the centre of contemporary language education and assessment (Council of Europe, 2020). This emphasis reflects a broader reorientation of language education and assessment toward action-oriented, socially situated language use, with mediation, plurilingualism, and collaborative meaning-making occupying a central role (North, 2021; Piccardo et al., 2019). From this perspective, language performance cannot be reduced to the production of accurate forms alone, as it also involves the capacity to interpret, reframe, explain, and negotiate meaning with others.

Mediation is the construct that most directly connects this view of language performance with chatbot-based tasks, since it involves making meaning accessible for others through explaining, reframing, clarifying, and helping interlocutors reach workable understanding, often across linguistic or cultural boundaries (Council of Europe, 2020). In digitally supported pedagogy, these processes are meaningful when learners must produce language while also managing misunderstandings, interpreting unfamiliar information, and adapting meaning to the needs of others. A chatbot may provide an interactional partner for mediation-oriented tasks, allowing these processes to be elicited in a focused and observable task environment.

This view of the chatbot as an interactional partner also reflects a broader shift in scholarship on AI-mediated language use. Recent work has treated AI not simply as another digital resource, but as a distinct interactional configuration within language education. Along these lines, Muñoz-Basols and Fuertes Gutiérrez (2025) understand AI as a new dimension of interaction in language learning, with implications for language contact, personalised practice, and pedagogically structured engagement with linguistic and intercultural content. However, much of this discussion remains conceptual and pedagogical in orientation rather than empirical. Further evidence is therefore needed on how such AI-mediated interaction relates to specific CEFR/CV-informed constructs, especially when mediation, online interaction, and pluricultural competence are examined separately rather than subsumed under broad measures of language gain.

This issue is singularly relevant in Official Language Schools (OLS) in Spain, a context in which adult learners in public language education may engage with AI-mediated tasks under conditions, expectations, and motivations that differ substantially from those of university students or younger school populations, who dominate much of the existing literature. Research on digital language immersion and virtual learning environments has shown that digitally mediated exposure may support the development of sociolinguistic

and intercultural competence (Soler Montes & Juan-Lázaro, 2025; Yan & Lowell, 2025). However, this body of work has not generally examined chatbot-mediated written interaction as a mediation-oriented task format, nor has it typically analysed CEFR/CV-informed mediation, online interaction, and pluricultural competence as distinct dimensions.

To address this gap, the present study reports a small-scale intervention conducted with adult B2 learners in OLS, state-funded public institutions dedicated to adult language teaching within the non-university education system. The task was built around a fictional rather than a real national culture to control the input conditions, reduce the risk of reproducing essentialist claims about existing communities, and focus learners' attention on inferencing, mediation, and interpretation. Participants had to resolve a hypothetical cultural misunderstanding between a Spanish visitor and a host from the fictional culture of Luxuria. The scenario included a series of potentially face-threatening incidents involving physical contact, conversational style at a formal meal, dress expectations, and appropriate behaviour toward domestic staff.

A central methodological feature of the present design lies in the use of a fictional culture as the basis for the task. In research on language, culture, and intercultural communication, real national or ethnic cultures can easily become reified through simplified descriptions or stereotypical assumptions (Pöllmann, 2026), particularly when learners are asked to interpret unfamiliar behaviour under time-constrained classroom conditions (Baker, 2015). By contrast, a fictional culture creates a controlled interpretive space in which all participants work from the same information base and in which culturally framed behaviour can be examined without attributing essentialised traits to existing communities (Davis et al., 2019). This choice also aligns with simulation-based approaches to intercultural competence development, which use controlled intercultural encounters to support learners' knowledge, adaptability, confidence, and assessment in culturally complex situations (Johnson, 2010). The fictional culture was therefore used as a methodological device suited to the non-confirmatory aims of the study. Although this design reduces the ecological validity of the task, since real intercultural encounters are also shaped by lived experience, affect, and identity positioning (Méndez García, 2017), it allows the task to foreground inferencing, clarification, explanation, and repair as core processes of mediation, while reducing the likelihood that learners' performance will depend primarily on prior knowledge, personal experience, or preconceived representations of real-world groups (Holden et al., 2021). The design nevertheless preserves the cultural dimension of the task, since participants still need to interpret norms, intentions, and expectations that are presented as socially meaningful within a coherent cultural frame.

Against this background, the present study addresses two related gaps, the limited evidence on chatbot use with adult learners in OLS and the relative scarcity of research that examines mediation as a construct in its own right rather than as part of broader, undifferentiated measures of language gain. It differs from closely related empirical work in three main respects. First, whereas previous studies on chatbot-assisted conversational English, AI-supported flipped classrooms, and AI-mediated informal digital learning have generally examined language development through broader measures of performance, engagement, or perception, this study separates online interaction, pluricultural competence, and mediation as distinct assessment dimensions. Second, it focuses on adult learners in OLS, a language-education setting that remains underrepresented in AI-mediated language learning research (Rubio-Gragera et al., 2025). Third, it uses a controlled fictional-culture scenario designed to elicit mediation processes while reducing reliance on prior cultural knowledge or stereotypical representations of real-world groups. The contribution of the study therefore lies in bringing together chatbot-mediated written interaction, CEFR/CV-

informed dimensional assessment, adult OLS learners, and a fictional-culture task designed to foreground mediation.

The study examines whether chatbot-mediated written interaction, compared with peer interaction, is associated with differences in online interaction, pluricultural competence, and mediation in an adult B2 EFL setting.

Accordingly, the study is guided by the following research questions:

(RQ1) To what extent is written interaction with a chatbot, as opposed to peer-to-peer written interaction, associated with differences in adult B2 learners' performance in online interaction, pluricultural competence, and mediation?
(RQ2) Are any observed differences associated with chatbot-mediated interaction distributed across all assessed dimensions, or are they specifically concentrated in mediation?
(RQ3) What AI-specific interactional behaviours can be identified in the experimental group?

The following hypotheses were formulated as directional expectations appropriate to an exploratory design:

H1. *The experimental group will outperform the control group in mediation;*

H2. *Any effect associated with chatbot-mediated interaction will be dimension-specific rather than global.*

2. Materials and Methods

2.1. Research Design

This study adopted an exploratory comparative intervention design. Two instructional conditions were compared for a single mediation-oriented written task: chatbot-mediated interaction and peer-to-peer chatroom interaction. The study is preliminary in two senses. First, it investigates an under-researched configuration that combines adult language learners, CEFR/CV-based mediation assessment, and a fictional-culture scenario. Second, it aims to identify dimension-specific tendencies rather than to establish definitive causal claims. The present design was thus intended to test the pedagogical plausibility of a chatbot-supported task targeting mediation rather than to establish a stable causal effect across contexts. The study accordingly prioritises transparent task description, construct-sensitive assessment, and cautious interpretation over strong causal rhetoric.

2.2. Participants and Context

The participants were adult learners of English enrolled at B2 level in OLS. Participants were drawn from intact class groups in the OLS context, and no random assignment was carried out. Both groups were enrolled at the same CEFR level (B2), which provided a common curricular reference, although no formal baseline measure was collected for statistical equivalence testing. The control group comprised 16 learners and the experimental group 24 learners. This imbalance reflected the natural size of the intact class groups available at the participating OLS, together with differences in attendance and consented participation on the day of data collection. All learner texts were anonymised through participant codes before analysis. Because the sample was relatively small and the groups were unequal in size, the inferential analyses were interpreted cautiously and complemented with descriptive statistics and effect-size estimates. These features position the study as a small-scale exploratory intervention rather than a confirmatory trial, with reduced statistical power and limited scope for causal inference. At the same time, the participant profile remains educationally relevant because adult learners in OLS are still underrepresented in the emerging literature on AI-mediated language learning.

2.3. Task and Instructional Conditions

The pedagogical task required learners to address a simulated cultural misunderstanding involving a Spanish visitor and a host from a fictional culture. The use of a fictional culture was intentional. It enabled the researchers to design a common scenario with controlled cultural information, to avoid essentialising real-world communities, and to focus the task on mediation processes such as explanation, clarification, reformulation, and negotiation of culturally situated meaning. The fictional culture, named Luxuria, was associated with explicit norms concerning celebrations, fashion, dining etiquette, social hierarchy, gift-giving, and personal space. The task scenario presented learners with four specific incidents during the visitor's stay with their host. The full task prompts administered to the experimental and control groups are provided in Supplementary Material S1.

In the experimental condition, learners interacted in writing with a chatbot configured with information about the fictional culture and instructed to respond as a culturally knowledgeable interlocutor within the Luxuria scenario. Learners were asked to describe what had happened during their trip and to pose at least ten questions to the chatbot in order to identify the causes of the misunderstandings and consider how they could have been avoided. This requirement was introduced as procedural scaffolding, given learners' limited prior experience with chatbots and the risk that they might otherwise treat the tool as a single-response information source rather than as an interlocutor for sustained inquiry. It was intended to promote sustained engagement with the scenario, although it also introduced a degree of task asymmetry between conditions by potentially encouraging more iterative clarification, reformulation, and advice-seeking in the experimental group. In the control condition, learners completed the task through peer-to-peer written interaction in a chatroom, with one member of each pair assuming the role of cultural expert familiar with the fictional culture. The peer assigned the role of cultural expert was provided with the same background information about Luxuria that informed the chatbot configuration, so that both conditions were based on a comparable cultural knowledge base. Both conditions were allotted the same time for task completion (45 min).

The task materials used in both conditions were initially generated with the support of ChatGPT-4o and were subsequently reviewed, revised, and pedagogically validated by the authors. Specifically, ChatGPT was used to help create the fictional culture of Luxuria and to draft the scenario prompts administered to the experimental and control groups. The cultural profile of Luxuria and the supporting materials used to configure the chatbot are provided in Supplementary Material S2. For replicability, the chatbot was prompted to act as an English-teaching AI interlocutor for mediation practice, drawing on background materials about invented cultures and responding to learners as a culturally knowledgeable host in simulated intercultural offence scenarios. For consistency, all participants in the experimental condition interacted with the same initial chatbot configuration, which was instructed to respond in English, remain within the Luxuria scenario, and prioritise explanations of cultural norms and misunderstandings over general language correction. What varied across conditions was not the cultural problem itself, but the kind of dialogue through which cultural clarification and mediation were pursued, sustained inquiry with a configured chatbot or explanatory exchange with an informed peer. During task completion, researcher intervention was limited to procedural clarification and did not include guidance on content, interpretation, or strategy.

2.4. Data Sources and Instrument

The primary data source consisted of the written interaction texts generated in both conditions. The independent variable was instructional condition, operationalised in two levels: chatbot-mediated interaction and peer-to-peer chatroom interaction. The

main dependent variables were the aggregated scores for online interaction, pluricultural competence, mediation, and the overall shared rubric. The six AI-specific items were analysed descriptively within the experimental condition and were not treated as between-group dependent variables. The texts were independently rated by two researchers using a rubric derived from CEFR/CV descriptors. The common rubric contained 23 items organised into three dimensions: online interaction (Items 1–6), pluricultural competence (Items 7–15), and mediation (Items 16–23). The experimental dataset also included six AI-specific items (Items 24–29) designed to capture behaviours associated with interaction with a chatbot. These supplementary items were informed by the construct of artificial intelligence literacy as operationalised by Carolus et al. (2023) in the Meta AI Literacy Scale (MAILS), a self-report instrument that measures core AI literacy facets such as using and understanding AI, while also addressing related competences such as AI self-efficacy and AI self-management. In the present study, the added items were not intended to reproduce the full scale, but to capture selected AI-related dimensions of learners' engagement with the chatbot within the specific pedagogical task.

The rubric was designed to preserve the multidimensionality of the task rather than to reduce performance to a single outcome measure. Online interaction was assessed through learners' management of written digital exchange, while pluricultural competence concerned their interpretation of the situation in relation to culturally framed expectations. Mediation, by contrast, referred to learners' capacity to clarify, reformulate, and use the interaction to construct workable understanding. This distinction was important because pluricultural competence involved recognising and interpreting culturally situated assumptions, whereas mediation required learners to make those interpretations accessible to an interlocutor through explanation, reformulation, clarification, and advice. The AI-specific items were included to characterise behaviours unique to the experimental condition, such as asking culturally oriented questions or using the chatbot to explore appropriate behavioural and non-verbal alternatives. Given that agreement was stronger at the composite than at the individual-item level, especially for online interaction, the rubric should be interpreted as more robust at the level of composite scores than of individual descriptors.

Each item was scored on a three-point ordinal scale: 1 = Poorly, 2 = Partially, and 3 = Completely. Since the CEFR/CV provides descriptors but does not prescribe a specific procedure for measuring their degree of attainment, the three-point scale was adopted as a transparent and practicable way of distinguishing between limited, partial, and fuller realisation of each descriptor. This format was appropriate for the exploratory nature of the study because it avoided implying spurious precision while still allowing raters to identify meaningful differences in performance. At the same time, as with any concise ordinal scale, it may have reduced sensitivity to finer distinctions between performances, particularly in dimensions where textual evidence was less readily observable. The full assessment rubrics, including the CEFR/CV-informed descriptors and the AI interaction items used for the experimental condition, are provided in Supplementary Material S3.

In addition to the numerical ratings, the raters recorded significant textual fragments as qualitative evidence; however, the present article reports only the quantitative strand. This limitation is intentional rather than incidental. The current paper concentrates on whether chatbot mediation is associated with dimension-specific performance tendencies.

2.5. Rating Procedure and Data Preparation

The two researchers rated the texts independently. Scores were first analysed separately to estimate inter-rater reliability. The verbal categories were then recoded numerically. Student identifiers were normalised to ensure accurate matching across raters, and aggregated scores were calculated at student level by averaging the two raters' scores for each

item and then computing mean scores for each dimension and for the overall shared rubric. Because inter-rater reliability was acceptable at the level of aggregated dimensions, the analyses prioritised composite dimension scores over isolated item-level contrasts. This decision was both psychometric and interpretive, as in a non-confirmatory study based on authentic written interaction, broader dimension scores provided a more stable and meaningful basis for comparing the two conditions.

2.6. Data Analysis

The analyses were conducted in SPSS Statistics v31. Inter-rater reliability for aggregated dimension scores was estimated with ICC(2,1), a two-way random-effects model appropriate when ratings are provided by independent raters and the intended inference extends beyond the specific raters involved. Given the ordinal rubric data, small sample size, and limited basis for assuming normality, between-group comparisons were conducted using Mann–Whitney U tests. Means are reported descriptively to facilitate comparison across aggregated rubric dimensions, whereas inferential interpretation relies on non-parametric results. Rank-biserial correlations were calculated as effect-size estimates for the Mann–Whitney U tests using the formula rrb = 1 − 2U/(n1n2), with positive values indicating higher scores in the experimental group. Effect sizes were interpreted descriptively, in keeping with the exploratory design. The analyses focused on the three shared dimensions of the rubric and on the overall shared score. Although p values were not adjusted for multiple comparisons, the mediation result would remain statistically significant under a Bonferroni threshold for the four shared outcomes. The six AI-specific items in the experimental condition were analysed descriptively, and no preregistration was conducted. The analytic code is available from the corresponding author on reasonable request.

3. Results

3.1. Inter-Rater Reliability

Inter-rater reliability was estimated for the aggregated dimension scores using ICC(2,1). As shown in Table 1, all composite scores reached acceptable reliability levels, with ICC values above 0.70 in both groups. Agreement was notably strong for the total shared rubric in the control group and for mediation in the experimental group. These results support the use of dimension-level composite scores as the primary basis for the inferential analyses.

Table 1. Inter-rater reliability for aggregated dimension scores (ICC(2,1)).

Dimension	Control	Experimental	Interpretive Note
Online interaction	0.703	0.774	Acceptable agreement
Pluricultural competence	0.796	0.748	Acceptable agreement
Mediation	0.784	0.836	Acceptable to good agreement
Total shared rubric	0.837	0.765	Acceptable to good agreement
AI-specific interaction	—	0.743	Acceptable agreement

Note: ICC = intraclass correlation coefficient. Values are reported for aggregated dimension scores rather than for individual items because reliability was stronger at the composite level.

3.2. Descriptive and Inferential Comparison of the Two Groups

The descriptive statistics and inferential results reported in Table 2 suggest a broadly similar profile for the two groups in online interaction and pluricultural competence, whereas a clearer difference appeared in mediation. The experimental group also obtained

a higher mean score on the total shared rubric, although this difference did not reach statistical significance. Mann–Whitney U tests confirmed that the only statistically significant between-group difference was found in mediation, in favour of the experimental group, U = 83.0, $p = 0.003$, rrb = 0.568. No significant differences emerged for online interaction, pluricultural competence, or the total shared rubric score, although the latter showed a small-to-moderate descriptive tendency favouring the experimental group. This tendency should nevertheless be interpreted in relation to the structure of the task itself, since the experimental condition explicitly required sustained questioning.

Table 2. Descriptive and inferential comparison between the control and experimental groups.

Dimension	Control Mean	Experimental Mean	U	*p*	Rank-Biserial r
Online interaction	1.828	1.955	165.0	0.463	0.141
Pluricultural competence	1.635	1.637	192.0	1.000	0.000
Mediation	1.539	2.018	83.0	0.003	0.568
Total shared rubric	1.652	1.852	128.5	0.082	0.331

3.3. AI-Specific Interactional Behaviours in the Experimental Group

The six AI-specific items were analysed descriptively because no directly equivalent measures were available in the control condition. The pattern indicated in Table 3 suggests that the experimental group made greater use of the chatbot to explore culturally appropriate behaviours, diverse perspectives, sensitive topics, and non-verbal communication cues. By contrast, the lowest mean corresponded to the use of the AI's language-processing capabilities to refine language output. These findings imply that participants tended to use the chatbot primarily as a source of culturally relevant clarification rather than as a general-purpose language correction tool.

Table 3. Descriptive results for the AI-specific items in the experimental group.

AI-Specific Item	Mean	SD	Median
Uses effective questions to obtain relevant information from the chatbot	2.083	0.289	2.000
Explores diverse topics and perspectives through interaction with the AI	2.292	0.498	2.000
Uses the AI's language processing capabilities to refine language output	1.250	0.399	1.000
Uses the chatbot to understand non-verbal communication cues in cultural contexts	2.250	0.622	2.000
Uses the chatbot to explore appropriate behaviours in diverse cultural settings	2.292	0.542	2.250
Explores sensitive topics across societies using the chatbot as a neutral informant	2.208	0.450	2.000

3.4. Dimension-Level Comparison Between Groups

Figure 1 visually summarises the dimension-level comparison between the two conditions. The pattern reinforces the interpretation of a selective difference concentrated in mediation rather than a generalised advantage for the experimental group. The results form a consistent quantitative profile in which the chatbot condition did not yield a uniform

advantage across the rubric, although higher performance was observed in the dimension most closely tied to the explanatory and repair-oriented demands of the task.

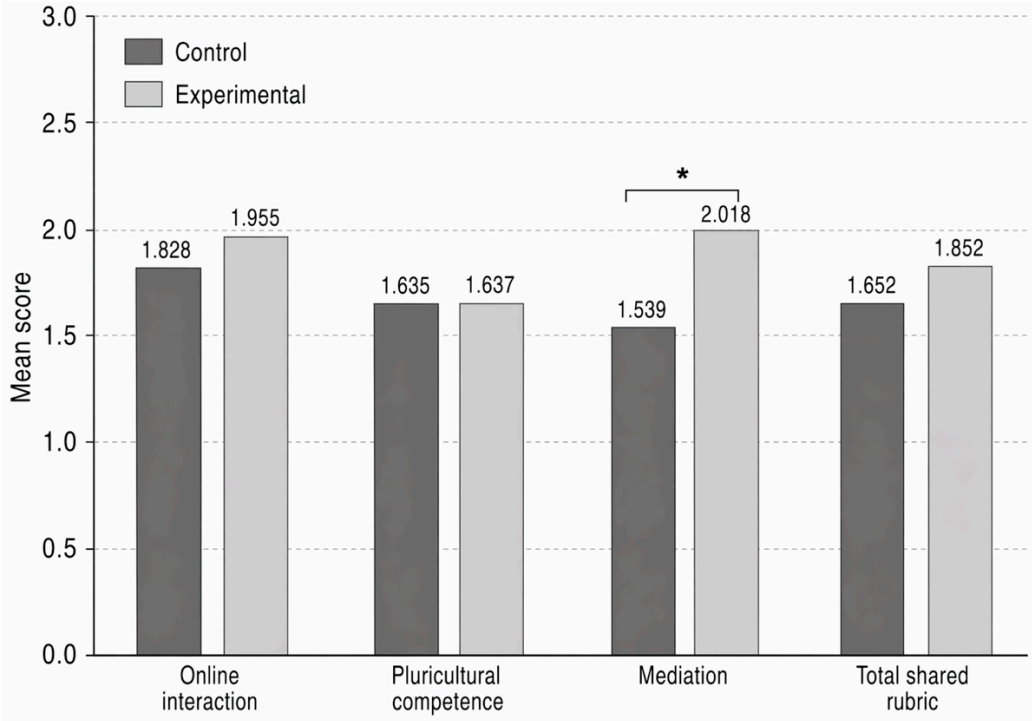

Figure 1. Mean scores for the control and experimental groups across the three shared dimensions and the overall shared rubric score. Note. Bars represent mean scores. * indicates a statistically significant between-group difference in mediation (U = 83.0, p = 0.003, rrb = 0.568).

4. Discussion

The findings point to a selective rather than generalised pattern associated with chatbot-mediated interaction. The experimental group did not outperform the control group in online interaction or pluricultural competence, nor did it show a statistically significant advantage on the total shared score. By contrast, the mediation dimension yielded the clearest difference, with a moderate-to-large effect size. This result is theoretically meaningful given the pedagogical focus of the task. Even so, this result should not be overinterpreted. One plausible interpretation is that the chatbot supported mediation-oriented discourse moves such as clarification, elaboration, reframing, and advice-giving. Another, more critical possibility is that the chatbot functioned as a more readily available and non-judgemental interlocutor than a peer, thereby making it easier for learners to sustain the questioning sequence required by the task. From this perspective, the observed advantage may reflect both the technological affordances of chatbot mediation and the interactional convenience of an interlocutor that remains consistently responsive throughout the task. The study thus supports a cautious, dimension-specific interpretation of chatbot affordances, but not a strong claim that AI improved overall communicative performance.

The AI-specific items further support this interpretation. Learners appeared to use the chatbot less as a tool for refining linguistic output than as a resource for obtaining culturally situated explanations, exploring behavioural alternatives, and clarifying non-verbal or sensitive aspects of the scenario. This pattern lends support to the view that the chatbot's contribution was mediation-oriented rather than broadly linguistic.

These results align with current calls to move beyond generic claims about AI effectiveness in language education. Recent reviews suggest that chatbot-supported language learning can be beneficial, but they also emphasise variability across learning outcomes,

task designs, and pedagogical conditions (Huang et al., 2022; Ji et al., 2023; Lyu et al., 2025). Our data fit this more cautious line of interpretation. The chatbot did not appear to enhance every dimension of performance; instead, any advantage seems to have been concentrated in the construct most directly activated by the task. In this respect, the present findings extend previous empirical work on chatbot-assisted conversational practice, AI-supported flipped classrooms, and AI-mediated informal digital learning by showing that chatbot-related gains may be more visible when assessment distinguishes mediation from broader measures of language performance, engagement, or perception.

In relation to the working hypotheses, H1 received tentative support, as the experimental group obtained higher mediation scores than the control group. H2 was also tentatively supported insofar as only mediation reached statistical significance, although the total shared score showed a non-significant tendency favouring the experimental group. Similarly, the reliability analyses showed that aggregated dimensions were more dependable than isolated rubric items. However, the mediation effect should be interpreted as emerging from the chatbot-supported task configuration as a whole. Although the questioning requirement was pedagogically justified as procedural scaffolding for learners with limited prior chatbot experience, it also limits the extent to which interlocutor type can be separated from the degree of prompted inquiry in the experimental condition.

The fictional-culture scenario is central to this interpretation. Because the task did not rely on an existing national culture, learners could not simply reproduce familiar stereotypes or rely on prior real-world cultural knowledge. Instead, they had to work within a controlled but unfamiliar frame and make sense of behavioural norms, expectations, and misunderstandings through interaction (Baker, 2015; Davis et al., 2019). In the experimental condition, the chatbot functioned as a stable cultural informant within that frame. In the control condition, that role was assumed by a peer acting as cultural expert. The fact that mediation, rather than pluricultural competence, emerged as the clearest differentiating dimension suggests that the main affordance of the chatbot may have been interactional support for meaning-making rather than broader intercultural sophistication per se (Holden et al., 2021).

The absence of a clear between-group difference in pluricultural competence also deserves fuller consideration. Although the scenario involved cultural misunderstanding, the task may have elicited practical problem solving more readily than deeper pluricultural reflection. At B2 level, learners may be better positioned to explain what went wrong and to formulate advice than to produce sustained evidence of decentring, perspective comparison, or critical reflection on cultural positioning. In addition, the fictional-culture design may have reduced reliance on stereotypes while also limiting some of the richer affective, experiential, and identity-related resources that often inform pluricultural engagement with real-world groups (Baker, 2015; Davis et al., 2019). The absence of a difference in this dimension is better understood as evidence that the task activated mediation and pluricultural competence to different degrees, despite the relationship between the two constructs. This reading accords with CEFR/CV B2 descriptors, which expect learners to do more than explain misunderstandings by recognising differing assumptions, reframing perspectives, and interpreting culturally situated behaviour on the basis of sufficient evidence. The results also accord with Johnson's (2010) view that simulated intercultural encounters can support learners' knowledge, adaptability, and confidence in culturally complex situations while allowing assessment to take place under controlled conditions.

The study has several limitations. Although inter-rater reliability reached acceptable levels for all aggregated dimensions, the stronger interpretability of composite scores than of isolated descriptors suggests that some CEFR/CV-informed categories may be difficult to apply consistently in short written mediation tasks. Such tasks may not always provide

sufficient evidence of interactional development, multiple perspectives, or sustained intercultural positioning. This is especially true of B2 descriptors such as restating and reframing positions, recognising that what is taken for granted in one context may not be shared in another, interpreting cultural cues appropriately, and explaining how communicative practices may generate misunderstanding (Council of Europe, 2020, p. 125). For this reason, inconsistent rating of some descriptors should be understood less as a weakness of the construct than as a limitation related to task brevity and the amount of observable evidence available.

In addition, the use of a fictional culture, while methodologically useful, necessarily limits ecological generalisation to real intercultural encounters. Task comparability across conditions also needs to be considered, since both groups were given the same amount of time to complete the task (45 min), whereas the experimental condition included a more explicit requirement for sustained questioning, which may itself have encouraged iterative clarification and reformulation. Moreover, because chatbot interaction may have been relatively novel for some learners, part of the observed engagement in the experimental condition could reflect a novelty effect rather than the affordances of chatbot-mediated mediation alone. The present article also reports only the quantitative strand of a broader dataset; more detailed insight into how learners positioned themselves culturally and how raters interpreted borderline cases requires qualitative analysis of the interaction logs.

Future research should test whether the same pattern holds in larger samples, in longitudinal designs, and in tasks that compare fictional and real cultural scenarios. Mixed-methods analyses of interaction logs would help clarify how learners formulate explanations, ask for clarification, and negotiate culturally shaped meanings when working with AI versus peers. More broadly, the findings support a research agenda that treats mediation as a particularly promising lens for investigating the pedagogical uses of chatbots in language education, while keeping pluricultural analysis analytically distinct and maintaining a critical view of whether observed gains reflect the technology itself or the interactional architecture built around it.

5. Conclusions

This exploratory study examined chatbot-mediated and peer-to-peer written interaction in a mediation-oriented EFL task carried out by adult B2 learners in OLS. Using a CEFR/CV-informed rubric, the study found no clear between-group differences in online interaction or pluricultural competence, but it identified a statistically significant advantage for the chatbot condition in mediation, which may be pedagogically promising in mediation-oriented task designs. Under the parameters of this task, the chatbot appeared to function less as a generic language helper than as a culturally configured interlocutor that supported learners' attempts to clarify and repair a sequence of misunderstandings.

The findings do not support a claim of broad AI superiority. Instead, they suggest that chatbot use may be well suited to tasks that require learners to clarify, reframe, and negotiate meaning within a culturally framed scenario. However, this conclusion must remain cautious, as the observed mediation advantage may reflect the affordances of the chatbot as an interlocutor alongside the patient and continuously available nature of the system, the possible novelty of chatbot interaction for some learners, and the specific interactional demands imposed on the experimental group. The study thus contributes to current debates by showing that the pedagogical value of chatbot interaction may be dimension-specific and tightly linked to task design.

For practice, the results suggest that chatbots may be most productively used as complementary tools in mediation-oriented writing tasks rather than as wholesale substitutes for peer interaction. Their pedagogical value may be greatest when tasks are explicitly designed

to elicit sustained questioning, clarification, reformulation, and culturally framed explanation, and when comparability between interactional conditions is carefully maintained.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/educsci16060844/s1, Supplementary Material S1 provides the task prompts administered to the experimental and control groups. Supplementary Material S2 provides the cultural profile of Luxuria and the supporting materials used to configure the chatbot. Supplementary Material S3 provides the full assessment rubrics used in the study, including the CEFR/CV-informed descriptors for online interaction, pluricultural competence and mediation, as well as the AI interaction items used for the experimental condition. The anonymised rating matrices can be made available by the authors upon reasonable request, subject to institutional and privacy restrictions.

Author Contributions: Conceptualisation, E.C.-B. and M.-d.-C.M.-G.; methodology, E.C.-B. and M.-d.-C.M.-G.; formal analysis, E.C.-B.; investigation, E.C.-B. and M.-d.-C.M.-G.; data curation, E.C.-B.; writing—original draft preparation, E.C.-B.; writing—review and editing, E.C.-B. and M.-d.-C.M.-G.; supervision, M.-d.-C.M.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Consejería de Desarrollo Educativo y Formación Profesional through the research project AMICIA: Aprendizaje Mediado e Intercultural con Inteligencia Artificial (Project No. PIV-016/25).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by Universidad de Jaén Ethics Committee, approval code 20240610/JUN.PRY, with approval granted on 10 June 2024.

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

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available because the interaction texts are linked to educational participants and could compromise anonymity if released in full.

Acknowledgments: The authors thank the participating learners and teaching staff at the Escuela Oficial de Idiomas Do Mundo Lume, in Ayamonte, for their collaboration in the study.

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

Abbreviations

The following abbreviations are used in this manuscript:

CEFR	Common European Framework of Reference for Languages
CEFR/CV	CEFR Companion Volume
OLS	Official Language Schools
GenAI	Generative artificial intelligence
ICC	Intraclass correlation coefficient
LLM	Large language model

References

Alenezi, A., & Alenezi, A. (2025). Evaluating the effectiveness of chatbot-assisted learning in enhancing English conversational skills among secondary school students. *Education Sciences*, *15*(9), 1136. [CrossRef]

Alm, A. (2024). Exploring autonomy in the AI wilderness: Learner challenges and choices. *Education Sciences*, *14*(12), 1369. [CrossRef]

Baker, W. (2015). Research into practice: Cultural and intercultural awareness. *Language Teaching*, *48*(1), 130–141. [CrossRef]

Carolus, A., Koch, M. J., Straka, S., Latoschik, M. E., & Wienrich, C. (2023). MAILS—Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100014. [CrossRef]

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment—Companion volume*. Council of Europe Publishing.

Davis, K. A., Taylor, A. R., Reeping, D., Murzi, H. G., & Knight, D. B. (2019). Experiencing cross-cultural communication on a home campus: Exploring student experiences in a cultural simulation activity. *Journal on Excellence in College Teaching*, *30*(4). Available online: https://celt.miamioh.edu/index.php/JECT/article/view/273 (accessed on 23 March 2026).

Du, J., & Alm, A. (2024). The impact of ChatGPT on English for Academic Purposes (EAP) students' language learning experience: A self-determination theory perspective. *Education Sciences*, *14*(7), 726. [CrossRef]

Holden, L. R., LaMar, M., & Bauer, M. (2021). Evidence for a cultural mindset: Combining process data, theory, and simulation. *Frontiers in Psychology*, *12*, 596246. [CrossRef] [PubMed]

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237–257. [CrossRef]

Ji, H., Han, I., & Ko, Y. (2023). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, *55*(1), 48–63. [CrossRef]

Johnson, W. L. (2010). Using immersive simulations to develop intercultural competence. In T. Ishida (Ed.), *Culture and computing. Lecture notes in computer science* (Vol. 6259, pp. 1–15). Springer. [CrossRef]

Lee, S., Choe, H., Zou, D., & Jeon, J. (2025). Generative AI (GenAI) in the language classroom: A systematic review. *Interactive Learning Environments*, *34*(1), 335–359. [CrossRef]

Li, B., Lowell, V. L., Wang, C., & Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education: Examining research on ChatGPT's use in language learning and teaching. *Computers and Education: Artificial Intelligence*, *7*, 100266. [CrossRef]

Li, B., Tan, Y. L., Wang, C., & Lowell, V. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. *Computers and Education: Artificial Intelligence*, *9*, 100445. [CrossRef]

Ling, Y. (2025). Voices from the flip: Teacher perspectives on integrating AI chatbots in flipped English classrooms. *Education Sciences*, *15*(9), 1219. [CrossRef]

Lyu, B., Lai, C., & Guo, J. (2025). Effectiveness of chatbots in improving language learning: A meta-analysis of comparative studies. *International Journal of Applied Linguistics*, *35*(1), 834–851. [CrossRef]

Méndez García, M. D. C. (2017). Intercultural reflection through the autobiography of intercultural encounters: Students' accounts of their images of alterity. *Language and Intercultural Communication*, *17*(2), 90–117. [CrossRef]

Muñoz-Basols, J., & Fuertes Gutiérrez, M. (2025). Opportunities for artificial intelligence (AI) in language teaching and learning. In J. Muñoz-Basols, M. Fuertes Gutiérrez, & L. Cerezo (Eds.), *Technology-mediated language teaching: From social justice to artificial intelligence* (pp. 417–450). Multilingual Matters.

Nguyen, L. T. H., Dinh, H., Dao, T. B. N., & Tran, N. G. (2025). Teachers' perceptions and students' strategies in using AI-mediated informal digital learning for career ESL writing. *Education Sciences*, *15*(10), 1414. [CrossRef]

North, B. (2021). The CEFR Companion Volume—What's new and what might it imply for teaching/learning and for assessment? *CEFR Journal—Practice and Research*, *4*, 5–24. [CrossRef]

Piccardo, E., North, B., & Goodier, T. (2019). Broadening the scope of language education: Mediation, plurilingualism, and collaborative learning: The CEFR Companion Volume. *Journal of e-Learning and Knowledge Society*, *15*(1), 17–36. [CrossRef]

Pöllmann, A. (2026). Three pitfalls in intercultural education: Cultural uniformity, pedagogical neutrality, and presumptions of competence. *Globalisation, Societies and Education*, *24*, 1–12. [CrossRef]

Rubio-Gragera, M., Palacios-Rodríguez, A., Cabero-Almenara, J., & Fernández Scagliusi, M. V. (2025). Digital teaching competence regarding foreign languages and learning modes at Official Language Schools in Andalusia (Spain). *Societies*, *15*(4), 99. [CrossRef]

Soler Montes, C., & Juan-Lázaro, O. (2025). Digital language immersion (DLI) and virtual exchanges. In J. Muñoz-Basols, M. Fuertes Gutiérrez, & L. Cerezo (Eds.), *Technology-mediated language teaching: From social justice to artificial intelligence* (pp. 312–350). Multilingual Matters.

Tutton, M., & Cohen, D. (2025). Reconceptualizing the role of the university language teacher in light of generative AI. *Education Sciences*, *15*(1), 56. [CrossRef]

Wiboolyasarin, W., Wiboolyasarin, K., Tiranant, P., Jinowat, N., & Boonyakitanont, P. (2025). AI-driven chatbots in second language education: A systematic review of their efficacy and pedagogical implications. *Ampersand*, *14*, 100224. [CrossRef]

Yan, W., & Lowell, V. L. (2025). Integrating artificial intelligence and extended reality in language education: A systematic literature review (2017–2024). *Education Sciences*, *15*(8), 1066. [CrossRef]