Mediating Language Acquisition and Intercultural Growth: Comparing Pre-AI and AI-Supported EMI Humanities Learning among International Engineering Students in Japan

Prof. Hatsuko Yoshikubo Ph.D., Shibaura Institute of Technology

Dr. Hatsuko Yoshikubo is a Professor of Humanities in the Innovative Global Program at the College of Engineering, Shibaura Institute of Technology, a leading engineering university in Tokyo, Japan, where she leads the Smart Education for Creative Engineers Laboratory (SECE Lab). She earned her Ph.D. in English Literature from Chiba University in 2002. Her research explores the value of incorporating humanities into science and engineering curricula, innovative assessment strategies in hybrid and international learning environments, and collaborative learning and assessment in global and COIL-based engineering education contexts, with a recent focus on AI-enhanced learning, data-driven assessment, and interdisciplinary engineering education design. She currently leads a research project funded by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (No. 24K06133, 2024–2027). She has presented her work at major international conferences including ASEE, WERA, BERA, IIAI-DSIR, and APAIE. Her work has been recognized with multiple awards, including the ASEE Best Diversity Paper Award (2023), the DSIR Outstanding Paper Award (2021), SIT's Presidential Award (2023), and recent recognitions including the 20th Kanto Society for Engineering Education Award (Best Paper/Essay Award) and the JSEE Annual Conference International Session Award (2025).

Gabriele Trovato

Gabriele Trovato is currently Associate Professor in Shibaura Institute of Technology, where he leads LAB 22, as well as Senior Visiting Fellow at the University of Palermo, Italy, International Co-Investigator at the National Research Foundation of Korea (NRF), and Visiting Researcher in Waseda University, Tokyo, Japan. He received his M.S. degree in Computer Engineering from the University of Pisa, Italy, and Ph.D. degree in Biorobotics in Waseda University. Within the relations between the two countries, Gabriele Trovato has been in the organising committee of Italy-Japan Workshops since 2011, and has been appointed "Ambassador of Livorno in the world" by the Municipality of Livorno, Italy. He has been Visiting Researcher in Karlsruhe Institute of Technology (Germany), PUCP (Peru) and Imperial College London (UK) among others. Gabriele Trovato has worked in the video game industry, being involved in the development of the world-wide known game series "Sid Meier's Civilization" and having created popular innovative mods for the game. His main research interests are interdisciplinary and include Human-Robot Interaction, with focus on culture and religion related aspects, artificial emotions in humanoids, and robot aesthetics.

Gabriele Trovato's creations, such as SanTO robots, are a combination of engineering, AI, art and humanities, and raised interest among the worldwide press, including the Wall Street Journal and the BBC.

Ujwal Kumar, Shibaura Institute of Technology

Ujwal Kumar is an undergraduate Computer Science student in the Innovative Global Program in the College of Engineering at Shibaura Institute of Technology in Tokyo. His research interests include generative AI and human–machine interaction, with practical experience developing AI agents and conducting research at both the Mobile Multimedia Communications Lab, led by Prof. Phan Xuan Tan, and the Matsuo Lab at the University of Tokyo. He joined SECE Lab. lead by Prof. Hatsuko Yoshikubo and co-authored a paper for the JSEE Annual Conference International Session (2025). He has also co-authored papers for IEEE and ACM. and has participated in international design and innovation programmes.

CHOI HYOSEOK, Shibaura Institute of Technology

HyoSeok Choi is an undergraduate Civil Engineering student in the Innovative Global Program in the College of Engineering at Shibaura Institute of Technology. He is a member of the Mobility Infrastructure

Lab, led by Prof. Yiping Le. His research explores how personal attributes influence travel behaviour, with a focus on promoting spontaneous travel through a support app. His global perspective has been shaped by international exchange and collaboration with peers from diverse cultural backgrounds. He joined SECE Lab. lead by Prof. Hatsuko Yoshikubo and co-authored a paper for the JSEE Annual Conference International Session (2025).

Mediating Language Acquisition and Intercultural Growth: Comparing Pre-AI and AI-Supported EMI Humanities Learning among International Engineering Students in Japan

Abstract

This work-in-progress examines how AI-supported multimodal instruction, aligned with the Common European Framework of Reference for Languages (CEFR), influences linguistic development and learner engagement in English-Medium Instruction (EMI) humanities learning among international engineering students in the College of Engineering at Shibaura Institute of Technology (SIT) in Japan. Two cohorts of students enrolled in a 14-week "History of Japan" module within an undergraduate engineering curriculum were compared: a pre-AI baseline cohort (Fall 2022, $n = 22$) and an AI-supported cohort (Spring 2025, $n = 18$). The 2022 cohort completed weekly reflective essays without AI assistance, while the 2025 cohort engaged with level-stratified multimodal resources developed using Claude 3.5 Sonnet and ElevenLabs. These multimodal resources consisted of highly detailed C2-level lecture slides paired with B2-level narrated summaries designed for accessibility; this combination was designed to reduce extraneous cognitive load while preserving conceptual challenge. English-language lexical development among the students was measured using the CEFR-based Vocabulary Level Analyzer (CVLA) v3.0, intercultural sensitivity was measured using the Miville-Guzman Universality-Diversity Scale–Short Form (MGUDS-S), and learner perceptions were recorded through surveys carried out in Week 14. CVLA results indicated significant lexical improvement in both cohorts (Fall 2022: $t = -2.244$, $p = .036$; Spring 2025: $t(17) = 2.72$, $p = .015$, Cohen's $d = 0.64$), with reduced post-test variance in the AI-supported cohort suggesting convergence toward higher proficiency. MGUDS-S analysis for the 2025 cohort ($n = 12$) showed a small but not statistically significant change in overall intercultural sensitivity ($p = .635$), with the MGUDS-S subscale 'Diversity of Contact' showing a medium effect size (Cohen's $d_z = 0.52$). A correlational analysis yielded a positive but non-significant association between lexical development and increasing intercultural sensitivity ($r = +0.485$, $p = .110$). Survey responses ($n = 15$) indicate that learners generally experienced the multimodal scaffolding as 'helpful' and 'beneficial for understanding the content', with notably higher ratings for visual slides than for audio-only narrations, and a majority preference for the instructor's voice over an AI-generated one. This study offers empirically informed design insights, that can aid educators in effectively integrating AI as 'strategic scaffolding' in EMI humanities learning within engineering education.

Keywords: English-medium instruction, generative AI, CEFR-based Vocabulary Level Analyzer, MGUDS-S, intercultural competence, multimodal learning, human-AI collaboration

1. Introduction

Engineering higher education increasingly operates within internationalized contexts where students must simultaneously develop disciplinary expertise and intercultural competence [1]. English-Medium Instruction (EMI)—the use of English to teach academic subjects in contexts where English is not the first language of most students—has emerged as a dominant pedagogical approach in engineering programs worldwide [2], so international students in EMI engineering curricula face the dual challenge of mastering technical content while navigating linguistic and cultural differences that shape their learning experiences. Humanities modules within engineering curricula offer distinctive opportunities for intercultural development, yet they present particular challenges when delivered in EMI contexts: unlike technical engineering courses where disciplinary terminology provides scaffolding, humanities courses demand nuanced language use, abstract reasoning, and cultural interpretation—cognitive demands that can overwhelm international students operating in their second or third language [3], and that may impede both content comprehension and the intercultural reflection these courses are designed to foster.

Recent advances in generative artificial intelligence (AI) have introduced new possibilities for addressing these challenges. Large language models and text-to-speech systems can generate multimodal learning materials calibrated to specific linguistic proficiency levels, potentially reducing cognitive load while maintaining academic rigor [4], [5]. However, empirical evidence regarding how AI-supported instruction affects linguistic development and intercultural competence in EMI engineering contexts remains limited, particularly regarding the relationship between these two developmental domains. This work-in-progress study addresses this gap by comparing linguistic and intercultural outcomes across two cohorts of international students enrolled in an EMI "History of Japan" module at Shibaura Institute of Technology (SIT): a baseline cohort from Fall 2022 (n = 22) who completed the course without AI-generated materials, and a cohort from Spring 2025 (n = 18) who engaged with AI-supported multimodal instruction. The temporal separation between cohorts reflects practical constraints of educational research while introducing potential confounds that we acknowledge and address methodologically. The relevance of this research direction is reinforced by contemporaneous industry-led work: after the present intervention was designed and implemented in Spring 2025, Google Research released Learn Your Way [6], which applies broadly similar principles of stratified multimodal scaffolding —including narrated slides and adaptive representations tailored to learner level—and reports retention benefits in a first-language English secondary school context. The convergence between independently developed industry and academic approaches highlights the timeliness of the present inquiry, while underscoring the need for evaluations in linguistically and culturally diverse higher education settings such as EMI engineering programs.

1.1 Research questions

This study investigates three interrelated questions:

1. How does AI-supported multimodal instruction influence linguistic development among international engineering students in an EMI humanities context?
2. What is the relationship between linguistic development and intercultural sensitivity growth in pre-AI and AI-supported learning environments?
3. How do students perceive the authenticity, utility, and affective impact of AI-generated learning materials?

1.2 Theoretical framework

We hypothesize a mediational pathway whereby AI-supported multimodal learning accelerates lexical development, which in turn may enhance intercultural competence. This hypothesis draws on Cognitive Load Theory [7], which posits that reducing extraneous cognitive load frees working memory resources for deeper processing, and Vygotsky's concept of mediation [8], whereby tools —including AI-generated materials—can scaffold learners toward higher-order thinking. By presenting content at stratified linguistic levels (C2 visual materials paired with B2 audio summaries), we aim to maintain academic challenge while ensuring comprehensibility, potentially accelerating vocabulary acquisition and enabling greater cognitive engagement with intercultural themes.

2. Literature Review

2.1 EMI and Language Development

EMI has expanded dramatically in non-Anglophone higher education contexts, particularly in engineering programs seeking to enhance graduate employability and institutional prestige [2]. However, research reveals persistent challenges: international students often struggle with academic language proficiency, limiting their participation in discussions and comprehension of complex concepts [3]. Vocabulary breadth and depth emerge as critical factors, with CEFR levels B2–C1 representing thresholds for effective academic engagement [9]. Studies examining vocabulary development in EMI contexts show mixed results. While immersion in English-language instruction can accelerate lexical acquisition, gains are often uneven and dependent on instructional design, language support mechanisms, and individual learner characteristics [10]. Affective factors such as language anxiety also moderate cognitive processing in second-language contexts [11], underscoring the need for instructional designs that lower extraneous demands without sacrificing academic rigor.

2.2 Intercultural Competence in Engineering Education

Intercultural competence—defined as the ability to interact effectively and appropriately across cultural differences [12]—has become a recognized learning outcome in engineering education [13]. The MGUDS-S instrument measures three dimensions: diversity of contact (openness to diverse others), relativistic appreciation (understanding multiple worldviews), and comfort with differences [14]. Engineering students, however, often show limited growth in intercultural

competence despite exposure to diverse peers and international content [15], suggesting that exposure alone is insufficient. Humanities courses offer intentional spaces for intercultural reflection, yet their effectiveness depends on students' cognitive and linguistic capacity to engage critically with cultural materials. This raises questions about whether linguistic barriers impede intercultural development—a relationship underexplored in existing literature.

2.3 AI in Language Learning and EMI

Generative AI technologies, particularly large language models (LLMs) like GPT-4 and Claude, have demonstrated capabilities in generating educational content, providing personalized feedback, and adapting materials to learner proficiency levels [5], [16]. Text-to-speech systems such as ElevenLabs enable creation of natural-sounding narrated content, supporting multimodal learning approaches that combine visual and auditory information. Preliminary studies suggest AI-generated materials can reduce cognitive load and increase engagement in language learning contexts [4], consistent with input-based theories of second language acquisition that emphasize comprehensible input as a driver of lexical development [17], though concerns persist regarding authenticity, over-reliance, and the displacement of human interaction. In EMI engineering contexts specifically, research remains nascent, with few empirical studies examining how AI-supported instruction affects linguistic and intercultural outcomes.

2.4 Multimodal Learning and Cognitive Load Theory

Multimodal learning—instruction that combines visual, auditory, and textual information—can enhance comprehension when designed according to cognitive load principles [18], [7]. Presenting complex information in stratified formats (e.g., advanced written content paired with accessible audio summaries) may allow learners to engage with academic-level concepts while receiving linguistic support, potentially accelerating vocabulary acquisition through repeated exposure in varied contexts. However, poorly designed multimodal materials can increase cognitive load through redundancy or conflicting information. Strategic alignment of modalities—ensuring visual and auditory channels complement rather than compete—is essential for effectiveness.

3. Methodology

3.1 Research Design and Limitations

This study employs a concurrent (convergent) mixed-methods design [19], in which quantitative and qualitative data are collected during the same overall period and integrated at the interpretation stage. Quantitative data (CVLA v3.0 vocabulary analysis and MGUDS-S intercultural sensitivity measures) are collected pre- and post-intervention, while qualitative data (student survey responses) are gathered at the end of the semester (Week 14) to capture learners' overall perceptions of the intervention. Findings are integrated at the interpretation stage, where qualitative insights contextualize and help explain quantitative patterns.

The comparison of Fall 2022 and Spring 2025 cohorts introduces temporal confounds—potential

changes in student demographics, institutional policies, or broader educational contexts over three years. A particularly relevant confound is the rapid expansion of student access to general-purpose generative AI tools (e.g., ChatGPT, Grammarly, QuillBot) between 2022 and 2025, which may have shaped students' baseline language environment independently of the present intervention. While random assignment was not feasible in this naturalistic educational setting, we employ several strategies to strengthen internal validity: (1) both cohorts followed identical course syllabi, learning outcomes, and assessment structures; (2) the same instructor taught both sections; (3) demographic data (nationality, prior English proficiency, engineering discipline) were collected to describe cohort comparability and to support sensitivity checks where feasible. Given the modest sample size, full covariate modeling is treated as future work. We recognize that the temporal separation between cohorts limits causal inference; therefore, cross-cohort differences are interpreted cautiously and supplemented with qualitative evidence.

3.2 Context and Participants

The study was conducted in a 14-week "History of Japan" course, a required humanities elective within an EMI engineering curriculum at SIT. The course addresses Japanese historical periods, socio-political transformations, and cultural developments, with explicit learning outcomes targeting both content knowledge and intercultural reflection.

Fall 2022 Cohort (Baseline): Twenty-two international engineering students (12 male, 10 female) from 15 countries participated. Nationalities included China (n = 6), Vietnam (n = 4), Indonesia (n = 3), Thailand (n = 2), and seven other countries (n = 1 each). Students were enrolled in mechanical engineering (n = 8), electrical engineering (n = 7), civil engineering (n = 4), and materials science (n = 3). Self-reported English proficiency ranged from CEFR B1 to C1 (median B2). Students submitted weekly 300-word reflective essays without AI assistance.

Spring 2025 Cohort (AI-Supported): Twenty international engineering students (11 male, 9 female) from 13 countries enrolled in the course. Nationalities included China (n = 5), Vietnam (n = 4), Indonesia (n = 3), Malaysia (n = 2), and six other countries (n = 1 each), representing diverse backgrounds from Asia, Europe (Germany, France, Netherlands), and Latin America (Brazil, Mexico). Students were enrolled in mechanical engineering (n = 7), electrical engineering (n = 6), civil engineering (n = 4), and chemical engineering (n = 3). Self-reported English proficiency ranged from CEFR B1+ to C1 (median B2).

Sample sizes for the three measures vary because participation in each instrument was voluntary under informed consent procedures: complete CVLA data were available for 18 students, complete paired pre/post MGUDS-S data for 12 students, and complete student perception survey responses for 15 students. The differing sample sizes reflect students' voluntary participation in each instrument in accordance with research ethics requirements; no student was required to complete any voluntary measure as a condition of enrollment or grading. Students engaged with AI-generated multimodal materials and submitted weekly 400-word reflective essays addressing the arrival of Western religion and science in Japan—a theme requiring both historical understanding

and intercultural reflection.

3.3 AI-Supported Intervention Design

For the Spring 2025 cohort, AI-generated multimodal materials were created using Claude 3.5 Sonnet (Anthropic) and ElevenLabs text-to-speech technology. This intervention employed a student-participatory research design wherein two undergraduate students (the lead authors of related conference presentations) played central roles in material development and voice selection. This student-led approach ensured materials aligned with learner needs and perspectives while providing authentic undergraduate research experience. The intervention followed a stratified multimodal design principle.

C2-Level Lecture Slides: Visually presented content maintained academic rigor, using advanced vocabulary, complex syntactic structures, and discipline-specific terminology appropriate for undergraduate-level humanities study. Slides included primary source excerpts, historical images, maps, and analytical frameworks.

B2-Level Narrated Summaries: Audio narrations of 1–2 minutes per lecture slide provided accessible summaries of key concepts, using simplified syntax, explicit signaling of relationships between ideas, and CEFR B2-level vocabulary. Narrations were generated using ElevenLabs with a consistent voice profile selected by student researchers to enhance familiarity and reduce processing demands.

This stratified approach was grounded in Cognitive Load Theory: advanced visual materials challenged students' conceptual understanding while accessible audio summaries reduced linguistic processing demands, theoretically freeing cognitive resources for deeper engagement with intercultural themes. Materials were created through an iterative process: (1) the instructor created slide content and narration scripts; (2) student researchers provided Claude 3.5 Sonnet with the narration scripts and target CEFR levels; (3) Claude generated narration scripts at two CEFR levels (B2 and C2); (4) CEFR levels were verified using Cathoven, an AI-powered language assessment tool, to ensure accurate B2 and C2 alignment before finalization; (5) the instructor reviewed materials for historical accuracy, cultural sensitivity, and pedagogical appropriateness, making revisions as needed; (6) narration scripts were processed through ElevenLabs using voice profiles selected by student researchers; (7) final materials were uploaded to the learning management system after each class session. Students in the AI-supported cohort were explicitly informed that materials were AI-generated and were encouraged to use them as supplementary resources alongside assigned readings and in-class discussions. The instructor maintained primary responsibility for instruction, using AI-generated materials to enhance—not replace—traditional pedagogical approaches.

3.4 Data Collection Instruments

3.4.1 CVLA v3.0 Vocabulary Analysis

The CEFR-based Vocabulary Level Analyzer (CVLA) v3.0 [20] is an automated tool developed

by Professor Satoru Uchida at Kyushu University and released in February 2025. CVLA analyzes written texts and assigns CEFR proficiency ratings from A1 (beginner) to C2 (mastery) based on lexical sophistication, word frequency distributions derived from the CEFR-J Wordlist created by Professor Yukio Tono, and grammatical complexity. Unlike vocabulary size tests, CVLA analyzes productive vocabulary use in authentic writing samples, providing ecologically valid measures of active lexical proficiency across 12 granular CEFR levels. For both cohorts, CVLA v3.0 analyzed students' writing samples. For the Spring 2025 AI-supported cohort, the first 1,000 words were extracted from each student's submitted assignments at the beginning of the semester, and the last 1,000 words were extracted from assignments at the end of the semester. For the Fall 2022 baseline cohort, weekly writing samples were aggregated, and average scores from weeks 1-3 and 11-13 were used as pre- and post-intervention measures, respectively [21]. CVLA outputs include overall CEFR level, percentage distribution across proficiency bands, and lexical diversity metrics. For statistical analysis, CEFR levels were converted into numeric values using the CVLA reference scale: Pre-A1 = 0.5, A1.1 = 0.84, A1.2 = 1.17, A1.3 = 1.5, A2.1 = 2.0, A2.2 = 2.5, B1.1 = 3.0, B1.2 = 3.5, B2.1 = 4.0, B2.2 = 4.5, C1 = 5.0, and C2 = 6.0.

3.4.2 MGUDS-S Intercultural Sensitivity

The Miville-Guzman Universality-Diversity Scale-Short Form (MGUDS-S) [22] is a validated 15-item instrument measuring intercultural sensitivity across three subscales: Diversity of Contact (DC; openness to diverse others), Relativistic Appreciation (RA; appreciation of diverse worldviews), and Comfort with Differences (CWD; comfort in intercultural interactions) [14]. Respondents rate statements on a 6-point Likert scale from "strongly disagree" (1) to "strongly agree" (6). Items 3, 6, 9, 12, and 15 (the CWD subscale) are reverse-scored using the standard formula ($7 - x$), following [22]. Each subscale ranges 5–30, and the Total score ranges 15–90. The instrument demonstrates strong psychometric properties with Cronbach's alpha values consistently above .80 for total scores and .70–.80 for subscales. MGUDS-S was administered to both cohorts at the beginning (week 1) and end (week 14) of the semester. For the Fall 2022 cohort, MGUDS-S data and corresponding correlational analyses with CVLA gains have been previously reported [21]. For the Spring 2025 cohort, MGUDS-S data and CVLA × MGUDS-S correlational analyses are reported here for the first time. Of the 18 students with complete CVLA data, 12 also provided complete paired pre- and post-intervention MGUDS-S responses; correlational analyses linking CVLA and MGUDS-S gains were therefore conducted on this matched subsample (n = 12).

3.4.3 Student Perception Survey

A student perception survey was administered to the Spring 2025 cohort at week 14 (conclusion) to capture evolving perceptions of AI-generated materials. Questions addressed four domains: (1) perceived ease of English comprehension, (2) usefulness for content understanding and vocabulary learning, (3) preferences regarding AI-generated narration (including comparisons with the instructor's voice), and (4) perceptions of the appropriateness of weekly writing tasks. Items used

5-point rating scales with open-ended elaboration. Of the 20 enrolled students, 15 provided complete survey responses in Week 14; survey analyses below are based on this $n = 15$ sample. Subsets of these survey findings have been previously presented at [23]; the present paper integrates them with the CVLA and MGUDS-S analyses for cross-measure interpretation. Survey data provide insights into subjective experiences not captured by quantitative measures, addressing RQ3 and providing context for interpreting linguistic and intercultural outcomes.

3.5 Data Analysis

3.5.1 Quantitative Analysis

CVLA v3.0 data were analyzed using paired-samples *t*-tests to compare pre- and post-intervention vocabulary levels within each cohort. To verify robustness against potential non-normality given the small sample sizes, Wilcoxon signed-rank tests were also conducted as sensitivity analyses; results were consistent with the parametric findings, and we therefore report *t*-test results in the main text with effect sizes using Cohen's *d* for between-condition comparisons and d_z for within-subject paired tests. Significance level was set at $\alpha = .05$. For MGUDS-S data, paired-samples *t*-tests examined Total and subscale-level changes within the Spring 2025 cohort, with Wilcoxon signed-rank tests run as sensitivity checks. Pearson correlations (r) were calculated between CVLA gain scores and MGUDS-S gain scores (Total and subscales) to address the mediational hypothesis underlying RQ2. Spearman rank-order correlations (ρ) were calculated as a robustness check. Where comparison with the Fall 2022 baseline cohort is informative, we report cross-cohort comparisons of correlation patterns; given the modest sample sizes, such cross-cohort observations are presented as hypothesis-generating rather than confirmatory.

3.5.2 Survey-Based Qualitative Insights

The student perception survey captured evaluations of the AI-supported materials, focusing on perceived ease of English comprehension, perceived usefulness for content comprehension and vocabulary learning, preferences regarding AI-generated audio (including comparisons with the instructor's voice), and perceptions of task appropriateness. Frequency distributions were calculated for each item, and aggregated positive endorsement (ratings 4–5 on the 5-point scale) is reported alongside item-level distributions to enable nuanced interpretation. These qualitative findings complement the quantitative results by adding contextual detail about how learners experienced the materials, which may help interpret patterns observed in linguistic and intercultural outcomes.

4. Results

4.1 Fall 2022 Cohort: Linguistic Development

CVLA v3.0 analysis of the Fall 2022 baseline cohort indicated significant lexical improvement over the semester. The mean CEFR-aligned lexical level increased from the first three weeks ($M = 4.818$) to the last three weeks ($M = 5.273$), with a mean change of 0.454. A paired-samples *t*-test

showed this increase was statistically significant, $t = -2.244$, $p = .036$. CVLA v2.0 yielded a comparable pattern, with mean levels rising from 4.954 to 5.432 ($t = -2.096$, $p = .048$). Individual trajectories showed variability: 18 students (81.8%) improved, 3 (13.6%) maintained stable levels, and 1 (4.5%) showed slight decline.

4.2 Spring 2025 Cohort: Linguistic Development

CVLA v3.0 analysis of the Spring 2025 AI-supported cohort ($n = 18$) revealed significant vocabulary gains across the semester. Mean lexical levels increased from the first 1,000 words ($M = 5.36$, $SD = 0.61$) to the last 1,000 words ($M = 5.72$, $SD = 0.46$). A paired-samples t-test confirmed that this improvement was statistically significant, $t(17) = 2.72$, $p = .015$, with a medium-to-large effect size (Cohen's $d = 0.64$). All 18 students maintained or improved their lexical level over the semester. The reduction in variability (SD from 0.61 to 0.46) suggests convergence toward higher proficiency by the end of the course, consistent with more equitable linguistic outcomes in the AI-supported condition. Table 1 summarizes CVLA results across cohorts.

Table 1. CVLA v3.0 lexical development results across cohorts

Cohort	Pre *M* (*SD*)	Post *M* (*SD*)	*t*	*p*	*d*
Fall 2022 (Pre-AI, *n* = 22)	4.818	5.273	−2.244	.036	0.68
Spring 2025 (AI, *n* = 18)	5.36 (0.61)	5.72 (0.46)	2.72	.015	0.64

4.3 Spring 2025 Cohort: Intercultural Sensitivity

MGUDS-S analysis of the Spring 2025 AI-supported cohort with complete paired pre/post data ($n = 12$) showed a small, non-significant change in overall intercultural sensitivity from pre- ($M = 74.75$, $SD = 7.07$) to post-intervention ($M = 75.50$, $SD = 6.29$; mean difference = 0.75 on the 90-point scale), $t(11) = 0.49$, $p = .635$, Cohen's $d_z = 0.14$. Subscale analyses indicated a more nuanced pattern. The Diversity of Contact (DC) subscale, which measures behavioral openness to engaging with diverse others, showed a borderline non-significant increase with a medium effect size ($M_{pre} = 25.50$, $M_{post} = 26.42$; $t(11) = 1.78$, $p = .102$, $d_z = 0.52$). In contrast, the Relativistic Appreciation (RA) and Comfort with Differences (CWD) subscales showed essentially no change ($d_z = -0.10$ and 0.02, respectively). Wilcoxon signed-rank tests yielded results consistent with the parametric findings. Table 2 summarizes MGUDS-S results.

Table 2. MGUDS-S paired analysis (Spring 2025, $n = 12$)

Subscale	Pre *M* (*SD*)	Post *M* (*SD*)	*t(11)*	*p*	*dz*
Total (max 90)	74.75 (7.07)	75.50 (6.29)	0.49	.635	0.14
DC (Diversity of	25.50 (2.94)	26.42 (2.84)	1.78	.102	0.52

Subscale	Pre *M* (*SD*)	Post *M* (*SD*)	*t(11)*	*p*	*dz*
Contact)					
RA (Relativistic Appreciation)	24.67 (2.93)	24.42 (2.97)	−0.36	.723	−0.10
CWD (Comfort with Differences)	24.58 (3.26)	24.67 (2.64)	0.08	.942	0.02

4.4 Correlation between Lexical and Intercultural Gains (Spring 2025)

To examine the hypothesized mediational pathway between linguistic and intercultural development (RQ2), Pearson correlations were calculated between CVLA gain scores and MGUDS-S gain scores for the matched subsample (n = 12). The correlation between CVLA gain and MGUDS-S Total gain was positive and medium in magnitude but did not reach statistical significance, r = +0.485, p = .110, 95% CI [−0.12, +0.83]. Among subscales, the Comfort with Differences correlation was the largest (r = +0.547, p = .065, 95% CI [−0.04, +0.85]), while Diversity of Contact (r = +0.374, p = .231) and Relativistic Appreciation (r = −0.085, p = .793) showed weaker associations. Spearman rank-order correlations yielded a comparable pattern. None of these correlations reached conventional significance thresholds with the available sample. The consistently positive direction for Total and CWD provides preliminary, exploratory support for the mediational hypothesis. Given the modest sample size, these patterns should be regarded as suggestive rather than confirmatory and warrant replication in future studies.

A comparison with previously reported results from the Fall 2022 baseline cohort [21] (n = 21) is presented in Table 3 to contextualize the Spring 2025 findings. In the Pre-AI cohort, the strongest positive correlation with lexical gain was found in Relativistic Appreciation (r = +0.411), whereas in the AI-supported cohort, Comfort with Differences emerged as the strongest correlate (r = +0.547), with Relativistic Appreciation showing essentially no association (r = −0.085). It should be emphasized, however, that the modest sample sizes in both cohorts limit statistical inference about cross-cohort differences.

Table 3. Cross-cohort comparison of CVLA gain × MGUDS-S gain correlations (CVLA v3.0). *Note:* Fall 2022 values are from [21].

MGUDS-S subscale	Fall 2022 (Pre-AI), *n* = 21	Spring 2025 (AI), *n* = 12
Total	r = +0.382, p = .097	r = +0.485, p = .110
DC	r = +0.289, p = .22	r = +0.374, p = .231
RA	r = +0.411, p = .072	r = −0.085, p = .793
CWD	r = +0.200, p = .40	r = +0.547, p = .065

4.5 Student Perceptions of AI-Generated Materials: Qualitative Findings

Survey responses (n = 15) captured students' evaluations of the AI-supported multimodal materials across four areas: perceived ease of English comprehension, perceived usefulness for content understanding and vocabulary learning, preferences regarding AI-generated audio, and perceived appropriateness of weekly writing tasks. Frequency distributions for the principal items are shown in Table 4.

Table 4. Student perception survey: frequency distributions on a 5-point scale, where 1 = not at all and 5 = highly (Spring 2025, n = 15). *Note:* Some sub-items were presented at [23].

Survey item	1	2	3	4	5	4+5 %
Ease of course English	0	0	0	6	9	**100.0%**
Ease of audio English	0	0	3	2	10	**80.0%**
Slide helpfulness	0	0	1	2	12	**93.3%**
Slides aiding understanding	0	1	1	4	9	**86.7%**
Audio helpfulness	1	2	6	1	5	**40.0%**
Audio aiding understanding	1	2	5	3	4	**46.7%**

Students reported high perceived ease of English processing: all 15 respondents (100%) rated the course English as "easy" or "very easy" (ratings 4–5), and 80.0% gave the audio narrations the same positive ratings, with 66.7% rating audio English at the maximum (5)—suggesting that the stratified design (C2-level slides paired with B2-level narration) supported overall comprehensibility. A more nuanced pattern emerged when comparing visual and audio modalities. The C2-level lecture slides received markedly positive evaluations: 93.3% rated slides as "helpful" or "very helpful" and 86.7% reported that slides aided their understanding, with 80.0% giving the maximum rating for slide helpfulness. By contrast, evaluations of the audio narrations were more mixed: only 40.0% rated audio as "helpful" or "very helpful" and 46.7% reported that audio aided understanding, with the modal response for both items at the midpoint (3 = "neutral"). This visual–audio divergence suggests that, in the present implementation, the slide modality functioned as the primary scaffold while audio narration played a more variable, supplementary role. Consistent with this pattern, 60.0% (9/15) of students preferred the instructor's voice over AI-generated narration, with open-ended comments describing AI audio as "clear but less personal" and requesting a "more lively tone." Evaluations of weekly writing task difficulty indicated an overall appropriate challenge level: 46.7% rated the task as "just right," 33.3% as "easy," 13.3% as "very easy," and 6.7% as "difficult." Taken together, these survey findings indicate that the overall positive perception of comprehensibility and the strongly favorable evaluation of slide-based scaffolding align with the linguistic outcomes observed in CVLA analyses, while the more variable reception of AI audio and the majority preference for the instructor's voice indicate clear directions

for refinement of audio modality in future iterations.

5. Discussion, Limitations, and Future Directions

5.1 Discussion

Both cohorts showed statistically significant vocabulary gains over a single semester with comparable medium-to-large effect sizes (Fall 2022: $d = 0.68$; Spring 2025: $d = 0.64$), so the magnitude of lexical improvement in the AI-supported condition was not larger than in the pre-AI baseline. Two features of the Spring 2025 results are nonetheless notable: all 18 students maintained or improved their lexical level with no observed declines, and post-test variability was reduced relative to pretest. These patterns are consistent with a more equitable distribution of learning outcomes in the AI-supported condition, although they do not establish that AI scaffolding caused this pattern. For intercultural sensitivity, the Spring 2025 cohort showed a small, non-significant overall gain, with the Diversity of Contact subscale—capturing behavioral openness to engaging with diverse others—showing a medium effect size that approached but did not reach conventional significance ($d_z = 0.52$, $p = .102$), while Relativistic Appreciation and Comfort with Differences showed essentially no mean change. This pattern is consistent with a trajectory in which behavioral openness may shift more readily than deeper cognitive or affective dimensions within a single 14-week semester, though these underpowered subscale effects should be treated as hypothesis-generating.

Correlational analyses yielded a positive but non-significant association between CVLA gain and MGUDS-S Total gain ($r = +0.485$, $p = .110$), with Comfort with Differences showing the largest subscale correlation ($r = +0.547$, $p = .065$)—consistent with the hypothesized mediational pathway. A comparison with the pre-AI baseline [21] suggests that the subscale most strongly associated with lexical gain may differ between cohorts: Relativistic Appreciation in the pre-AI cohort and Comfort with Differences in the AI-supported cohort. One interpretation, consistent with Cognitive Load Theory, is that reducing extraneous linguistic processing demands may shift the freed cognitive resources toward affective rather than abstract cognitive dimensions of intercultural sensitivity. However, the modest sample sizes preclude formal inference about cross-cohort differences, so this observation should be treated as a pattern motivating future investigation rather than established evidence. Student experience data add a further interpretive layer: all 15 survey respondents perceived the course English as easy or very easy, and a notable visual–audio divergence emerged. The C2-level slides were strongly endorsed (93.3% rated helpful or very helpful), while the B2-level audio narrations received more variable evaluations (40.0%); 60% of students preferred the instructor's voice, with open-ended comments describing AI audio as "clear but less personal." This divergence likely reflects both the natural advantage of visual materials in EMI humanities content and the limited affective and prosodic qualities of current AI narration, suggesting that linguistic accessibility alone is insufficient for audio scaffolding to be fully valued by learners.

These findings should be situated within a wider research landscape. Google Research's Learn

Your Way [6], released subsequently to the present implementation, applies comparable principles of stratified multimodal scaffolding and reported improvements in long-term recall in a first-language secondary school context; the convergence between independently developed industry and academic implementations suggests that this design space is gaining empirical support across diverse contexts. The present study extends this line of inquiry to the more linguistically demanding setting of EMI humanities learning at the tertiary level, contributing evaluation in an L2/L3 EMI context, measurement of intercultural sensitivity alongside lexical development, and a longer 14-week observation period. Taken together, the present findings position AI-supported multimodal materials not as replacements for human teaching but as strategically integrated scaffolds that may help reduce linguistic barriers and support participation in EMI humanities learning. The most defensible conclusion at this stage is that AI-supported scaffolding did not impede learning outcomes and was experienced as supportive by learners—particularly through the visual slide modality—while showing signs of more equitable linguistic outcomes. Whether AI scaffolding produces measurable additive effects on intercultural development, and through what mechanisms, remains a question for future research with larger samples, longer time horizons, and stronger designs.

5.2 Limitations and Future Directions

Several limitations constrain interpretation of these preliminary findings. The single-semester evaluation period (14 weeks) restricts conclusions about the sustainability of learning gains beyond the course, and the study is situated within a single institutional context that may shape outcomes through local curricular design, instructor practices, and institutional culture. Sample sizes vary across measures (CVLA $n = 18$, MGUDS-S $n = 12$, Survey $n = 15$) because participation in each instrument was voluntary under informed consent; this voluntary, non-coercive participation reflects appropriate research ethics but constrains the statistical power available for some analyses. The paired CVLA × MGUDS-S subsample ($n = 12$) is especially underpowered for detecting subscale-level associations—with this sample size, only large correlations (approximately $|r| > .55$) reach conventional significance thresholds—so reported correlations should be regarded as effect-size estimates with substantial uncertainty rather than established associations. Comparisons of correlation patterns between the Fall 2022 and Spring 2025 cohorts cannot be tested with confidence at the present sample sizes, and are therefore presented as observations rather than confirmed findings. A further methodological constraint concerns CVLA measurement: a notable proportion of Spring 2025 students were already at the C2 ceiling (6.0) at the first 1,000-word measurement, which may have constrained observable lexical gains. Additionally, the temporal separation between cohorts introduces broader contextual confounds, including the rapid expansion of student access to general-purpose generative AI tools (ChatGPT, Grammarly, QuillBot) between 2022 and 2025.

Future work should address these limitations through three main directions. First, long-term follow-up studies are needed to examine whether linguistic gains persist over time. Second, multi-institutional collaboration will be essential to test generalizability across diverse EMI contexts and

to provide samples sufficient for stronger statistical inference, including formal tests of mediational hypotheses. Third, continued enhancement of AI technologies—particularly voice generation—should aim to improve human-likeness, emotional tone, and personalization while maintaining clarity and CEFR-aligned accessibility, addressing the affective and prosodic limitations highlighted by survey responses. Future analyses might also supplement CVLA with within-C2 lexical sophistication metrics (e.g., type-token ratio, academic vocabulary density) to better capture growth among already-advanced learners.

6. Conclusion

This study investigated how AI-supported, CEFR-aligned multimodal instruction can influence lexical development, intercultural sensitivity, and learner engagement in an EMI humanities module in engineering curriculum. CVLA results indicated statistically-significant lexical improvement in both cohorts, with comparable effect sizes (Fall 2022: $d = 0.68$; Spring 2025: $d = 0.64$); all Spring 2025 students maintained or improved their lexical levels, and reduced post-test variance suggests more equitable linguistic performance outcomes in the AI-supported condition. MGUDS-S analyses for the 2025 cohort ($n = 12$) showed a small, non-significant gain in overall intercultural sensitivity, with a medium but also non-significant effect size for the 'Diversity of Contact' subscale. For the AI-supported cohort of students taking the program in 2025, correlations between CVLA and MGUDS-S gains were positive but non-significant, and potentially consistent with the hypothesized mediational pathway—namely that AI-supported multimodal learning can accelerate lexical development, in turn enhancing intercultural competence. A cross-cohort comparison with previously reported pre-AI baseline data [21] suggests possible shifts in which the growth of students' intercultural sensitivity tracks lexical gain—an observation that requires replication and should be regarded as hypothesis-generating. Survey-based qualitative findings ($n = 15$) suggest that the multimodal scaffolding supported comprehension and engagement, with notably stronger endorsement of visual slides than of audio narrations and a with a majority preference for a 'live' recording of the instructor’s voice over AI-generated audio using a third-party voice. Taken together, the present findings are most defensibly interpreted as showing that AI-supported learning materials did not impede learning and were experienced as supportive by learners—particularly through the visual modality—with potential signs of more equitable linguistic outcomes; whether AI scaffolding produces measurable additive effects on intercultural development, and through what mechanisms, remains an open question for future research with larger samples, longer time horizons, and stronger designs. The convergence with contemporaneous industry-led developments such as Google’s Learn Your Way [6] suggests that stratified AI-supported scaffolding constitutes a promising and timely direction; the present study contributes one early academic perspective on its application to the linguistically demanding context of EMI humanities learning.

Acknowledgements

The authors would like to thank the students from Shibaura Institute of Technology who participated in this study, and Dr. Marie Miville for permission to use the MGUDS-S. This research is supported by JSPS Research Grants 24K06133.

References

[1] J. Jenkins and A. Mauranen, Linguistic Diversity on the EMI Campus: Insider Accounts of the Use of English and Other Languages in Universities within Asia, Australasia, and Europe. London, U.K.: Routledge, 2019.

[2] E. Macaro, S. Curle, J. Pun, J. An, and J. Dearden, "A systematic review of English medium instruction in higher education," Lang. Teach., vol. 51, no. 1, pp. 36–76, 2018.

[3] E. Dafouz and U. Smit, ROAD-MAPPING English Medium Education in the Internationalised University. Cham, Switzerland: Springer, 2020.

[4] D. Baidoo-Anu and L. O. Ansah, "Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning," J. AI, vol. 7, no. 1, pp. 52–62, 2023.

[5] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learn. Individ. Differ., vol. 103, Art. no. 102274, 2023.

[6] Google Research, "Learn Your Way: Reimagining textbooks with generative AI," Google Research Blog, Sep. 2025. [Online]. Available: https://research.google/blog/learn-your-way-reimagining-textbooks-with-generative-ai/

[7] J. Sweller, P. Ayres, and S. Kalyuga, Cognitive Load Theory. New York, NY, USA: Springer, 2011.

[8] L. S. Vygotsky, Mind in Society: The Development of Higher Psychological Processes. Cambridge, MA, USA: Harvard University Press, 1978.

[9] Council of Europe, Common European Framework of Reference for Languages: Learning, Teaching, Assessment—Companion Volume. Strasbourg, France: Council of Europe Publishing, 2020.

[10] U. Wannagat, "Learning through L2—Content and language integrated learning (CLIL) and English as medium of instruction (EMI)," Int. J. Biling. Educ. Biling., vol. 10, no. 5, pp. 663–682, 2007.

[11] P. D. MacIntyre and R. C. Gardner, "The subtle effects of language anxiety on cognitive processing in the second language," Lang. Learn., vol. 44, no. 2, pp. 283–305, 1994.

[12] D. K. Deardorff, "Identification and assessment of intercultural competence as a student outcome of internationalization," J. Stud. Int. Educ., vol. 10, no. 3, pp. 241–266, 2006.

[13] A. Parkinson, "The rationale for developing global competence," Online J. Glob. Eng. Educ., vol. 4, no. 2, Art. no. 2, 2009.

[14] P. M. King, R. J. Perez, and W. Shim, "How college students experience intercultural learning: Key features and approaches," J. Divers. High. Educ., vol. 6, no. 2, pp. 69–83,

2004.

[15] B. K. Jesiek, Y. Haller, and J. Thompson, “Developing globally competent engineering researchers: Outcomes-based instructional and assessment strategies from the IREE 2010 China research abroad program,” Adv. Eng. Educ., vol. 4, no. 1, pp. 1–31, 2014.

[16] R. Luckin, W. Holmes, M. Griffiths, and L. B. Forcier, Intelligence Unleashed: An Argument for AI in Education. London, U.K.: Pearson Education, 2016.

[17] S. Krashen, Principles and Practice in Second Language Acquisition. Oxford, U.K.: Pergamon Press, 1982.

[18] R. E. Mayer, The Cambridge Handbook of Multimedia Learning, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2014.

[19] J. W. Creswell and V. L. Plano Clark, Designing and Conducting Mixed Methods Research, 3rd ed. Thousand Oaks, CA, USA: SAGE Publications, 2017.

[20] S. Uchida, “Measuring English text difficulty based on CEFR-J word profiling: The development and application of CVLA 3.0,” ARELE: Annu. Rev. Engl. Lang. Educ. Jpn., vol. 34, pp. 83–98, 2023.

[21] H. Yoshikubo, A. Cetinkaya, G. Trovato, P. Sripian, and E. Kamioka, “Evaluating lexical development and global competence in international engineering students: A pre-AI comparative analysis (JSEE 2025 Annual Conference International Session Award),” in Proc. JSEE 73rd Annu. Conf. & Expo. Int. Session, Kyoto Univ., Kyoto, Japan, Aug. 2025, doi: 10.20549/jseeen.2025.0_22.

[22] J. N. Fuertes, M. L. Miville, J. J. Mohr, W. E. Sedlacek, and D. Gretchen, “Factor structure and short form of the Miville-Guzman Universality-Diversity Scale,” Meas. Eval. Couns. Dev., vol. 33, no. 3, pp. 157–169, 2000.

[23] U. Kumar, H. Choi, and H. Yoshikubo, “A design-based study of AI-enhanced multimodal support for content understanding in an English-medium history course: Toward data-driven education for Society 5.0,” in Proc. JSEE 73rd Annu. Conf. & Expo. Int. Session, Kyoto Univ., Kyoto, Japan, Aug. 2025, doi: 10.20549/jseeen.2025.0_26.