Doi: 10.60072/book.icoge2026.025

AI-Enabled E-Portfolios for Authentic Assessment and Digital Credentialing in Language Learning

Ramesh Mani Suvedi

School of Education, Lincoln University College, Malaysia

Corresponding author: ***rameshmani@lincoln.edu.my***

Abstract

Traditional language assessment systems are increasingly inadequate for capturing the dynamic, process-oriented nature of second language acquisition in digital and online learning environments. Despite growing interest in e-portfolios and AI-assisted feedback, few empirical studies have examined their integration within a unified, tiered-access platform that also incorporates digital credentialing. This mixed-methods study evaluates an AI e-portfolio system for language learners, combining CEFR-aligned automated feedback with digital credentials. Tracking 120 users through platform analytics, surveys, and 18 interviews, the research demonstrates how automated validation processes can enhance global academic recognition.

Learners demonstrated significantly higher engagement with AI feedback than anticipated, with 78% rating it as helpful or very helpful. Portfolio submission rates were three times higher when AI feedback was available within 24 hours of submission. Expert comparison confirmed moderate-to-high AI-human agreement on grammatical and lexical dimensions, with lower agreement on task achievement, revealing a persisting limitation in automated discourse evaluation. The study contributes an original empirical account of AI-enabled e-portfolio deployment in language education, demonstrating its promise as an equitable, scalable assessment alternative while identifying equity concerns within tiered access models. Implications are drawn for instructional designers, language educators, assessment scholars, and EdTech developers.

Keywords: AI-Enabled E-Portfolio, Automated Feedback, Digital Credentialing, Language Assessment, Open Badges

Introduction

The current landscape of global English Language Teaching (ELT) in 2026 is defined by a profound "Trust Paradox" that sits at the intersection of unprecedented digital accessibility and a widening credibility gap. On one hand, the technological revolution has democratized elite knowledge to an extent previously unimaginable, allowing "hopeful" learners from the Global South, including regions like Nepal, Malaysia and India, to access high-level curricula from institutions, such as Harvard, Oxford, Cambridge, Coursera, Udemy, Future Learn, Google and so on without the prohibitive costs of physical relocation. This democratization is powered by a robust ecosystem of AI-enabled platforms, including Google for Education, SchoolAI (app.schoolai.com), and Fobizz.com, which offer affordable, life-changing educational experiences.

Furthermore, technical mastery is no longer the only way to acquire academic certifications by traditional universities. Online skills and certificates providers like FreeCodeCamp.com, W3schools.com, Cursor AI, Replit and cutting-edge models from DeepSeek and Anthropic offer specialised coding and language tracks that allow professionals to escalate their academic endeavors with satisfying results.

Assessing second language (L2) proficiency has long been recognised as both complex and controversial. As language learners become increasingly diverse in their linguistic backgrounds, learning pathways, educational settings, and access to support, the limitations of traditional standardised testing are becoming more evident. Conventional assessments tend to provide isolated snapshots of performance rather than capturing the ongoing, reflective, and developmental nature of language learning itself (Black & Wiliam, 1998).

At the same time, rapid digital transformation is changing how language education is delivered and experienced. Online learning platforms, asynchronous instruction, and AI-supported educational tools have expanded access to language learning while also increasing the demand for more flexible and innovative assessment practices. Many learners now study outside conventional classroom environments and often lack regular access to instructors who can provide timely and personalised feedback. This mismatch between learners' needs for continuous support and institutions' ability to provide it has become a major challenge in modern language education (Warschauer & Grimes, 2008).

Electronic portfolios, commonly known as e-portfolios, have frequently been proposed as a way to address this issue. By allowing learners to collect, reflect on, and present authentic samples of their language use over time, e-portfolios offer a broader and more meaningful representation of language development than single high-stakes examinations (Barrett, 2007; Yancey, 2009). Nevertheless, traditional e-portfolio systems face significant practical limitations. They require considerable instructor involvement, are time-consuming to assess, and often remain disconnected from modern digital credentialing systems (Beckett et al., 2013; Little, 2009).

This paper presents the design, implementation, and preliminary evaluation of one such platform: the AI-Enabled Language Portfolio (AILP). The platform was piloted with adult language learners enrolled in both free and paid course tiers, reflecting the varied access conditions commonly found in online language learning environments today. The study adopted a convergent parallel mixed-methods approach, combining platform analytics, learner surveys, semi-structured interviews, and expert rubric comparisons to examine system effectiveness, learner experiences, and issues of equitable access.

Review of Related Literature

The theoretical and empirical basis of this study draws from three closely connected areas: e-portfolio pedagogy and assessment, artificial intelligence in language education, and digital credentialing. This review brings together key scholarship from these fields in order to highlight the conceptual gap that the present study seeks to address.

E-Portfolios in Language Learning: Practical Implementation

Empirical Evidence from Language Education

Research on the use of e-portfolios in language education has generally produced positive findings, although outcomes often depend on the learning context. Klenowski et al.

(2006) reported that e-portfolios in higher education encouraged learner reflection and promoted a developmental understanding of language performance. Similarly, Cheng and Warren (2010) found that portfolio assessment in EFL writing courses improved both writing quality and learners' metacognitive awareness of the writing process, particularly among lower-proficiency learners who benefited from repeated cycles of feedback and revision.

Scalability and the Instructor Bottleneck

A recurring challenge within the e-portfolio literature is balancing pedagogical quality with operational scalability. Meaningful portfolio assessment requires substantial instructor involvement, including reviewing multiple drafts, providing individualised feedback, and engaging learners in reflective dialogue. In large-scale online learning environments, where instructors may be responsible for hundreds of asynchronous learners, maintaining this level of interaction becomes extremely difficult (Dörnyei & Ushioda, 2011; Locker & Jaques, 2018).

This scalability challenge has increased interest in AI-assisted feedback systems as a way to reduce dependence on instructor availability while still supporting formative assessment processes.

Artificial Intelligence in Language Assessment

Automated Essay Scoring: Origins and Development

The use of computational approaches in language assessment dates back more than sixty years. One of the earliest systems, Project Essay Grade, was introduced by Page (1966). This system relied on surface-level textual features such as word frequency, sentence length, and syntactic complexity as indicators of writing quality. Over time, more advanced automated scoring systems were developed, including e-rater (Burstein et al., 1998), Intelligent Essay Assessor (Landauer et al., 2003), and MY Access! (Vantage Learning, 2000).

Since 2017, the field has undergone major transformation due to the emergence of transformer-based large language models (LLMs). Beginning with models such as BERT (Devlin et al., 2019) and advancing rapidly through GPT-3 (Brown et al., 2020), GPT-4, and related systems, AI technologies have become significantly more capable of understanding and generating natural language.

CEFR-Aligned Automated Assessment

Automated Writing Assessment (AWE) systems are an early and widely used form of artificial intelligence in writing instructions (Shen, 2025). An important advancement in AI-supported language assessment has been the alignment of automated feedback systems with established language proficiency frameworks, particularly the Common European Framework of Reference for Languages (CEFR). The CEFR offers a six-level proficiency scale ranging from A1 to C2 and provides detailed "can-do" descriptors that outline observable communicative abilities at each level (Council of Europe, 2001). Because the framework is widely recognised by educational institutions, testing organisations, and language programmes worldwide, it provides a strong foundation for structuring AI-generated language feedback.

Learner Perceptions of AI Feedback

Research examining learners' attitudes toward AI-generated feedback has generally reported positive responses, especially regarding speed, consistency, and the reduced

social pressure associated with automated systems. Warschauer and Grimes (2008) conducted one of the earliest systematic investigations into learner perceptions of automated writing evaluation (AWE) in second-language contexts. Their study found that learners appreciated the immediacy and specificity of automated feedback, although they still preferred human instructors for addressing higher-order writing concerns.

Limitations and Validity Concerns

Despite growing enthusiasm for AI-based language assessment, important concerns remain regarding its validity and fairness. Bridgeman et al. (2012) found that automated essay scoring systems might perform unevenly across different learner populations. In particular, lower agreement rates were observed for essays produced by non-native English speakers from diverse linguistic backgrounds compared to native speaker writing. This raises concerns about algorithmic bias, especially when systems are trained predominantly on English-first-language corpora.

Digital Credentialing in Education

The Architecture of Open Badges

The concept of open badges was formally introduced by Mozilla in 2011 through the Open Badges Infrastructure (OBI), a framework designed for issuing, displaying, and verifying digital credentials. Unlike simple visual achievement icons, open badges are portable and verifiable digital records containing embedded metadata, such as the issuing organisation, award criteria, supporting evidence, and date of issuance (Mozilla Foundation, 2011).

Gibson et al. (2015) conducted one of the earliest comprehensive studies on the use of open badges in higher education and identified four major advantages: recognising informal and non-formal learning, motivating learners through visible achievement milestones, enabling credentials to be combined into larger qualification pathways, and allowing credentials to be shared across professional and social platforms.

Micro-Credentials and the Granularisation of Language Competence

Alongside the development of open badges, micro-credentials have emerged as another important form of digital certification. Micro-credentials are focused qualifications that certify achievement within a narrowly defined skill area (Oliver, 2019). In language education, they offer opportunities to recognise competencies that are often overlooked by traditional qualifications. Examples include writing formal business emails at a B2 level, demonstrating spoken fluency in hospitality-related communication, or summarising academic texts in a second language.

Motivation, Gamification, and Digital Badges

Research into the motivational impact of digital badges has been strongly influenced by self-determination theory (Deci & Ryan, 1985; Ryan & Deci, 2000) and gamification theory (Deterding et al., 2011; Hamari et al., 2014). From the perspective of self-determination theory, badges function as external motivators that can encourage learners to internalise motivation when the rewards are perceived as meaningful and personally relevant. Under these conditions, learners may begin to see achievement activities as aligned with their own goals and identities.

Abramovich et al. (2013) conducted a significant study on badge systems in online learning environments and found that the motivational effectiveness of badges depended heavily on their design.

Credibility, Trust, and Institutional Endorsement

One of the most significant challenges facing digital credentials is the issue of credibility. Although badges and certificates often motivate learners, confidence in their external recognition remains uncertain. Selingo (2018) found that employer acceptance of digital badges and micro-credentials was still relatively limited, especially in sectors where traditional university degrees continue to hold strong signalling value. Gallagher and Palmer (2020) argued that institutional partnerships and co-endorsement play a crucial role in increasing trust in alternative credentials, highlighting the importance of collaboration between educational providers and recognised institutions.

Research Gap

The literature reviewed above highlights an important gap in current research. Although substantial scholarship exists on e-portfolio pedagogy, AI-assisted language feedback, and digital credentialing individually, very few studies have explored the integration of all three within a single platform designed specifically for language learning. To date, no published empirical study has examined such integration within a tiered-access environment that combines AI-generated feedback, portfolio assessment, and digital credentialing in one unified system.

In addition, existing research has paid limited attention to the equity implications of tiered-access AI assessment systems. As digital learning technologies become increasingly common, concerns are growing that unequal access to advanced educational tools may reproduce or intensify existing educational inequalities. Despite this concern, little practical work has investigated how different access levels within AI-enabled assessment systems may shape learner experiences and outcomes. This study aims to address both of these gaps.

Objectives

This study aims to achieve the following objectives:

- To develop and describe an integrated AI-enabled e-portfolio system that combines automated CEFR-aligned feedback, multimodal artefact submission, and competency-based digital credentialing.
- To examine learners' engagement with AI-generated feedback and explore their perceptions of its usefulness across both free and paid course tiers.
- To evaluate the validity of AI-generated feedback by comparing it with expert human assessment across four language-related dimensions.
- To investigate learners' attitudes towards digital credentials and explore how these credentials influence motivation and portfolio completion.
- To analyse issues of equity within tiered access conditions and identify design considerations for creating more inclusive AI-enabled assessment systems.

Research Questions

The study is guided by the following research questions:

- RQ1: How do language learners perceive and interact with AI-generated feedback within an integrated e-portfolio environment, and are there differences between free-tier and paid-tier learners?

- RQ2: To what extent does AI-generated language feedback correspond with expert human assessment across four CEFR-aligned dimensions: grammatical accuracy, lexical range and appropriacy, discourse coherence, and task achievement?
- RQ3: What role do digital badges and certificates play in motivating learners to engage with and complete portfolio-based assessment activities?
- RQ4: What equity-related concerns arise from the implementation of a tiered-access AI-enabled e-portfolio system, and how might these concerns be addressed through platform design and educational policy?

Methodology

This study employed a convergent parallel mixed-methods design (Creswell & Plano Clark, 2018), in which quantitative and qualitative data were collected simultaneously, analysed separately, and then integrated during the interpretation stage. This approach was selected because it allows for a more comprehensive understanding of the research problem than either quantitative or qualitative methods alone. Quantitative data from platform analytics and learner surveys provided measurable insights into patterns of engagement and learner attitudes, while qualitative interview data offered deeper contextual explanations of these patterns and learner experiences.

Study Design and Paradigm

The research was situated within a pragmatist paradigm (Tashakkori & Teddlie, 2010), which prioritises the research questions themselves when determining methodological choices. Rather than adhering strictly to one methodological tradition, pragmatism values the complementary strengths of both quantitative and qualitative approaches.

System Description: The AI-Enabled Language Portfolio (AILP)

Platform Architecture

The AILP platform was developed as a modular, cloud-based web application consisting of three interconnected functional components. The first component was a portfolio management interface that allowed learners to upload multimodal artefacts and maintain reflective learning journals. The second component was an AI-powered feedback and assessment engine designed to provide CEFR-aligned analysis of written and spoken language submissions. The third component was a digital credentialing module capable of issuing open badges and verifiable PDF certificates once learners achieved specified competency benchmarks.

The platform was built using Moodle 4.1 as a headless learning management system, combined with customised API integrations linking the portfolio interface to the AI assessment engine and credentialing infrastructure. The credentialing system followed the Open Badges v2.0 specification and incorporated blockchain-based verification to ensure credential authenticity.

Tiered Access Structure

The platform operated using a dual-tier access model intended to accommodate learners with differing financial and institutional resources. The free and paid tiers provided different levels of access to platform features, while maintaining a shared core learning environment. summarises the distribution of features across both tiers

Table 1. Feature distribution across free and paid platform tiers

Feature	Free Tier	Paid Tier
Portfolio submission & storage	Yes (unlimited)	Yes (unlimited)
AI written feedback	Yes (up to 10/month)	Yes (unlimited)
Spoken language feedback	No	Yes
Learning analytics dashboard	Basic (submission count)	Full (progress tracking, trends)
Course completion certificate	Yes	Yes
Competency micro-badges	Core badges only (3)	Full suite (12 badges)
Instructor review request	No	Yes (2 per module)
Peer feedback tools	Yes	Yes

AI Feedback Engine

The AI feedback engine analyses submitted written artefacts across four dimensions derived from CEFR writing assessment criteria (Council of Europe, 2001): (a) grammatical accuracy and range, (b) lexical resource and appropriacy, (c) discourse coherence and cohesion, and (d) task achievement and communicative effectiveness. Each dimension is scored on a 1–5 Likert scale calibrated to CEFR levels (1 = A1–A2; 3 = B1–B2; 5 = C1–C2), and specific actionable feedback is generated for each dimension within the same response.

Participants

The pilot study involved 120 adult learners enrolled in online English language courses delivered through the AILP platform during an eight-week period. Participants were recruited through purposive sampling from the existing platform user base, stratified to achieve balanced representation across proficiency levels and course tiers. Table 2 presents the participant profile.

Table 2. Participant demographic and proficiency profile by course tier.

Characteristic	Free Tier (n = 65)	Paid Tier (n = 55)	Total (n = 120)
Age range (years)	19–52 (M = 31.4)	22–48 (M = 33.1)	19–52 (M = 32.2)
Gender (% female)	54%	60%	57%
CEFR proficiency (A2–B1)	48%	36%	43%
CEFR proficiency (B2–C1)	52%	64%	57%
L1 background (languages)	14 languages	11 languages	18 languages
Prior e-portfolio experience	31%	44%	37%
Country of residence (top 3)	Malaysia, Nepal, Sudan, Syria, Sri Lanka, Saudi Arabia	Malaysia, Saudi Arabia, Bangladesh	Malaysia, Nepal, Yemen

All participants were adults (over 18 years old) with voluntary enrollment in different online courses. The study excluded learners who did not complete at least three portfolio submissions during the pilot period, as minimum engagement was required for meaningful data generation. Eighteen participants from the broader cohort were selected for semi-structured interviews using maximum variation sampling to ensure representation of diverse proficiency levels, tier memberships, and L1 backgrounds.

Participant Profile by Course Tier

Grouped bar chart showing key demographic and proficiency characteristics of the 120 participants (free: n = 65; paid: n = 55). Variables: total per tier, CEFR proficiency (A2-B1 and B2-C1), gender (female), and prior e-portfolio experience. Paid-tier learners showed higher B2-C1 proficiency (64% vs. 52%) and prior e-portfolio experience (44% vs. 31%), both relevant to interpreting tier-based outcome differences.

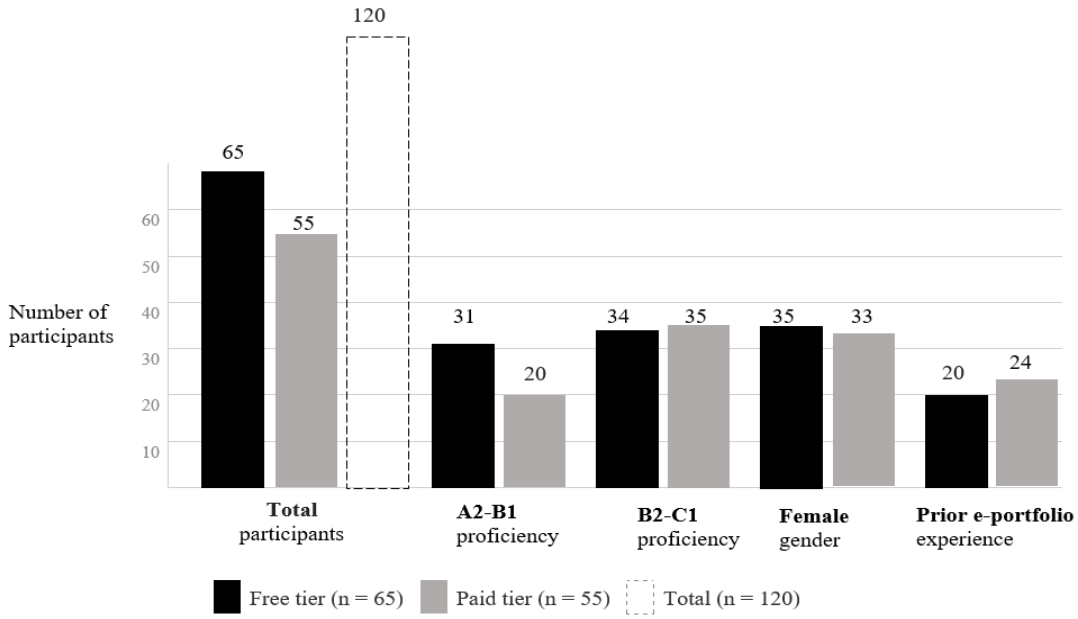

Figure 1. Participant profile by course tier

Participants recruited via purposive sampling; stratified for proficiency and tier balance

Figure 1. Participant demographic and proficiency profile by course tier (n = 120). Free tier (dark bars, n = 65); Paid tier (light bars, n = 55). Purposive sampling, stratified for proficiency and tier balance.

Note: 18 L1 backgrounds; top countries: Malaysia, Nepal, Sudan, Syria. Mean age: free M = 31.4 yrs, paid M = 33.1 yrs. Learners with fewer than 3 submissions were excluded.

AILP System Architecture: Three Functional Layers

Three interconnected layers within a Moodle 4.1 headless LMS: (1) portfolio management interface for multimodal artefact submission; (2) AI feedback engine, GPT-4 fine-tuned on CEFR-annotated samples, providing four-dimensional CEFR-calibrated scoring; (3) digital credentialing module issuing Open Badges v2.0 credentials with blockchain verification. The dual-tier access model operates horizontally across all three layers.

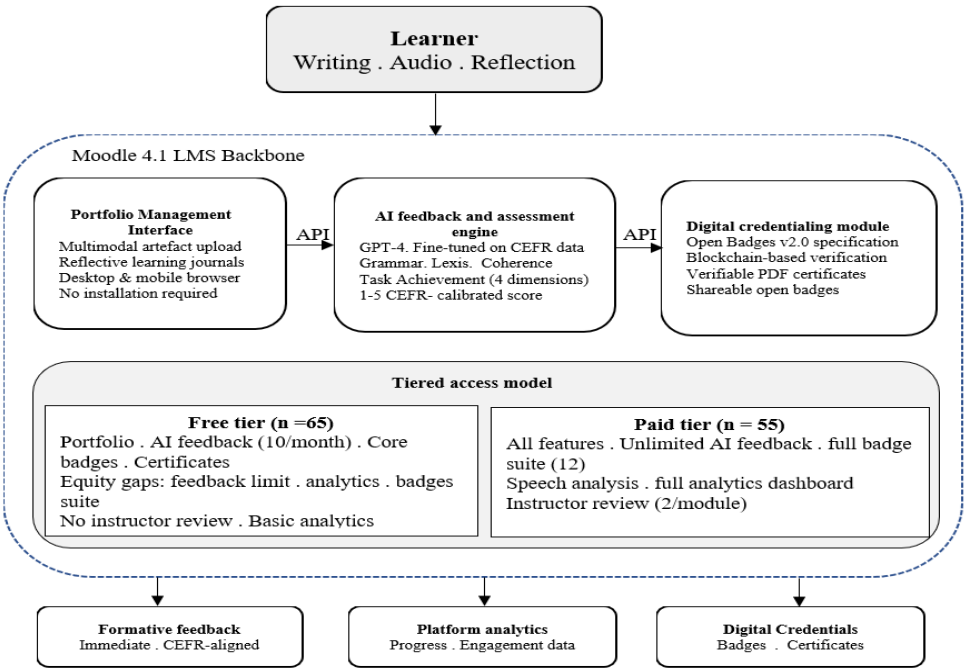

Figure 2. AILP platform architecture showing the three functional layers (portfolio management, AI feedback engine, digital credentialing module) within the Moodle 4.1 LMS backbone and the dual-tier access model.

Note: All features were accessible via desktop and mobile web browsers with no software installation required. API integrations connect the three layers; the credentialing module links to an external blockchain verification registry.

Data Collection Instruments

Platform Analytics

Anonymised platform analytics data were extracted from the LMS dashboard at the close of the pilot period. Metrics captured included: total submissions per learner; AI feedback interaction rate (submissions followed by revision within 72 hours); feedback ratings (learner-initiated helpfulness ratings on a 5-point Likert scale); time-on-task data; badge and certificate award rates; and cross-tier comparative engagement data. These metrics operationalised the quantitative aspects of RQ1, RQ3, and RQ4.

Learner Surveys

Two structured surveys were administered: a mid-pilot survey (Week 4) focusing on AI feedback perceptions and platform usability, and a post-pilot survey (Week 8) focusing on credential attitudes, overall system evaluation, and equity perceptions. Both surveys employed a combination of five-point Likert scale items, semantic differential scales, and open-ended response fields. The surveys were pilot-tested with five learners not included in the main study and revised for clarity before administration. Survey reliability was assessed using Cronbach's alpha, yielding acceptable internal consistency coefficients (mid-pilot: $\alpha = .82$; post-pilot: $\alpha = .79$).

Semi-Structured Interviews

Eighteen participants completed individual semi-structured interviews conducted via video conferencing platforms, lasting between 30 and 40 minutes. Interview protocols were developed deductively from the research questions and inductively from

preliminary survey themes, covering five topic areas: (1) prior experience with e-portfolios and AI tools; (2) perceptions of AI feedback quality and usefulness; (3) the role of digital credentials in motivation and identity; (4) experiences of the tier model and perceived equity; and (5) recommendations for system improvement. Interviews were audio-recorded, transcribed verbatim, and member-checked for accuracy.

Expert Rubric Comparison

To assess the construct validity of AI-generated feedback (RQ2), two experienced language assessment specialists, both holding postgraduate qualifications in applied linguistics with a minimum of ten years' assessment experience, independently evaluated a random stratified sample of 30 learner submissions (15 from each tier; 15 spanning A2–B1, 15 spanning B2–C1) using the same four-dimensional rubric employed by the AI system. AI scores and feedback on the same submissions were provided to the experts without identifying information, and inter-rater agreement between expert assessments and AI outputs was calculated using Cohen's kappa (κ) and intraclass correlation coefficients (ICC).

Analysis

Quantitative Analysis

Quantitative data from platform analytics and surveys were analysed using IBM SPSS Statistics 29. Descriptive statistics (means, standard deviations, frequencies, and percentages) were used to characterise engagement patterns and survey response distributions. Independent samples t-tests and Mann-Whitney U tests were used to compare quantitative outcomes between free and paid tier participants, with statistical significance set at $p < .05$. Effect sizes were reported using Cohen's d for parametric comparisons and r for non-parametric comparisons.

Qualitative Analysis

Interview transcripts and open-ended survey responses were analysed using reflexive thematic analysis following the six-phase framework of Braun and Clarke (2006, 2019). Familiarisation with data was followed by systematic coding in NVivo 14, enabling iterative refinement of an initial coding frame into thematic clusters. Two researchers independently coded a 20% subset of transcripts to establish coding reliability ($\kappa = .76$), and divergences were resolved through discussion. Analysis was conducted with explicit reflexive awareness of the researcher's dual role as system developer and study investigator.

Integration

Quantitative and qualitative findings were integrated at the interpretation stage using a joint display matrix (Fetters et al., 2013), in which quantitative patterns were placed alongside corresponding qualitative themes to identify points of convergence, complementarity, and divergence. Divergent findings, instances where quantitative data pointed in a different direction from qualitative accounts, were treated as analytically productive and subjected to additional interpretive scrutiny.

Ethical Considerations

All participants provided informed written consent prior to data collection. Anonymity was ensured through the replacement of all identifying information with participant codes in transcripts and data files. Learners were informed that non-participation would have

no effect on their course access or assessment outcomes. Data were stored on encrypted institutional servers accessible only to the research team, and will be retained for five years following publication in line with institutional data management policy.

Findings and Discussion

This section presents the integrated findings from the four data sources, organised by research question. Unless otherwise stated, percentages refer to survey respondents (n = 112 usable mid-pilot surveys; n = 108 usable post-pilot surveys) and analytics data refer to the full cohort of qualifying participants (n = 120).

RQ1: Learner Engagement with and Perceptions of AI Feedback Engagement Patterns

Overall, learners submitted a mean of 7.4 portfolio artefacts during the eight-week pilot period (SD = 2.9), compared to a pre-pilot institutional average of 2.3 submissions per learner per equivalent period—a statistically significant increase ($t(119) = 14.2$, $p < 0.001$, $d = 1.30$). Paid-tier learners submitted significantly more artefacts ($M = 9.1$, $SD = 2.4$) than free-tier learners ($M = 5.8$, $SD = 2.8$; $t(118) = 7.1$, $p < 0.001$, $d = 1.29$). However, among free-tier learners who reached the platform-imposed monthly feedback limit (n = 27), subsequent submission rates dropped markedly—an average of 2.1 submissions in the week following limit-exhaustion, compared to 4.8 in the preceding week—suggesting that feedback availability was a significant driver of submission behaviour.

Portfolio completion rates, defined as submitting at least one artefact per active week across the eight-week pilot, were 74% for paid-tier learners and 58% for free-tier learners, a statistically significant difference ($\chi^2(1) = 4.02$, $p = .045$). However, free-tier learners who had not exhausted their feedback allowance showed completion rates (71%) statistically comparable to paid-tier learners ($p = .67$), reinforcing the interpretation that feedback volume, rather than tier membership per se, was the primary engagement driver.

Perceived Feedback Quality

On the post-pilot survey, 78% of respondents rated AI feedback as “helpful” or “very helpful” overall, with 82% rating it positively for grammar and vocabulary and 71% for discourse organisation. Ratings were significantly lower for task achievement feedback (61%), consistent with the expert comparison findings reported in Section 4.2.

Interview data provided qualitative texture to these ratings. The most consistently appreciated dimension of AI feedback was immediacy: sixteen of eighteen interviewees mentioned receiving feedback quickly as a significant factor in their continued engagement. One participant (P7, B2, free tier) expressed this as follows:

> *"When I submit something at 11 o'clock at night, I don't want to wait three weeks for a teacher to read it. I get the feedback immediately. I read it, I understand it, I fix it. That is how I want to learn."*

Learners at lower proficiency levels (A2–B1) expressed more mixed evaluations. Several (n = 9) found the vocabulary used in AI feedback itself to be challenging, reporting difficulty understanding the diagnostic commentary even when they broadly understood the corrections suggested. This finding points to a design gap in proficiency-adaptive feedback language—a limitation discussed further in Section 5.

Tier Comparison

The cross-tier comparison yielded a nuanced pattern. While paid-tier learners reported higher overall feedback satisfaction (M = 4.12 vs. M = 3.78 on a 5-point scale; t(110) = 2.89, p = .005, d = 0.55), qualitative data suggested that this difference was largely attributable to the richer feature set of the paid tier—particularly the spoken language feedback and analytics dashboard—rather than the quality of written feedback per se, which learners across both tiers rated comparably (M = 3.91 vs. M = 3.83; p = .41). This supports the platform's design intention that written feedback quality should not vary by tier.

RQ2: Validity of AI Feedback — Expert Comparison

The expert rubric comparison yielded differentiated results across the four assessment dimensions, as summarised in Table 3.

Table 3. Inter-rater agreement between AI-generated and expert human assessment scores across four CEFR-aligned dimensions.

Assessment Dimension	Cohen's κ	ICC (95% CI)	Agreement Level
Grammatical accuracy & range	.72	.81 (.67–.90)	Substantial
Lexical resource & appropriacy	.68	.77 (.62–.87)	Substantial
Discourse coherence & cohesion	.54	.63 (.45–.77)	Moderate
Task achievement & communicative effectiveness	.38	.47 (.26–.63)	Fair
Overall composite score	.61	.71 (.55–.82)	Substantial

Substantial agreement on grammatical accuracy (κ = .72) and lexical resource (κ = .68) confirms that the AI feedback engine reliably operationalises these lower-order constructs in ways that align with expert judgement. Moderate agreement on discourse coherence (κ = .54) suggests that the engine captures some but not all of the features that expert assessors attend to in evaluating text organisation and cohesive flow—a finding consistent with Llosa and Malone (2019) and attributable in part to the inherently holistic and interpretive nature of coherence assessment. The fair agreement on task achievement (κ = .38) represents the most significant finding from the expert comparison and warrants extended discussion.

RQ3: Motivational Role of Digital Credentials

Quantitative Evidence of Credential Motivation

By the close of the pilot period, 89 of 120 participants (74%) had been awarded at least one digital badge, and 61 (51%) had been awarded a course segment completion certificate. Among those who had received a badge, 91% reported on the post-pilot survey that they intended to share it publicly—most commonly via LinkedIn (68%), followed by personal or academic portfolios (43%) and WhatsApp/social media (31%).

Survey items directly probing motivational effects showed strong positive responses: 84% of respondents agreed or strongly agreed that "earning a badge or certificate

motivated me to complete the required portfolio tasks," and 79% agreed that "digital credentials made my progress feel more real and meaningful." The motivational effect was significantly stronger for paid-tier learners ($M = 4.31$ vs. $M = 3.94$; $t(106) = 2.54$, $p = .013$, $d = 0.49$), potentially reflecting the broader range of badges available to paid learners and the associated richer sense of a credential pathway.

Qualitative Accounts of Credential Value

Interview data revealed three distinct dimensions of credential value. The first, and most frequently articulated, was identity recognition, the sense that a badge or certificate made visible a competence that had previously existed only in the learner's private experience.

User Type 1:

> *"I have been speaking English for years and everyone in my workplace knows I speak well. But I never had anything to show on paper. Now I have a badge. It says B2 Academic Writing. I put it on LinkedIn and three people commented. That matters."*

> Type 2: *"I have been learning English in a foreign land, but I have never been able to assess my actual English language finesse with AI, which is also considered to be a native English examiner. After assessing my English language ability using 'sharedpearls.com', I am confident that what I am perfect in and what I need to improve in order to attain my academic goals."*

The second dimension was goal crystallisation—the sense that the credential system gave learners a concrete and proximate target to work towards, transforming a diffuse aspiration (improving English) into a structured achievement pathway:

> *"Without the badge, I just knew I had to submit. With the badge, I knew exactly what I had to do and what I would get. It was like a map. I knew where I was going."*

The third dimension—credential credibility—was more ambivalent. While most learners were motivated by badges, several expressed concern about their external recognition:

> *"I like it. But I don't know if my employer will think it is real. It's from a computer. They don't know this company. Cambridge—they know. IELTS—they know. This? I'm not sure."*

This credibility concern was expressed by 12 of 18 interviewees in some form, making it the most prevalent concern in the qualitative dataset and aligning with wider literature on employer trust in non-traditional credentials (Selingo, 2018; Gallagher & Palmer, 2020).

AI Feedback as a Scalable Formative Tool

The engagement data from this study provide compelling evidence that AI-enabled feedback can function as a genuine driver of formative learning activity in language education contexts. This finding extends Warschauer and Grimes's (2008) earlier work on automated writing evaluation by demonstrating engagement effects in a more naturalistic, non-experimental, longitudinal online learning context.

The expert comparison results confirm a differentiated picture of AI validity that has significant implications for how AI feedback should be positioned pedagogically. For grammatical and lexical feedback, the AI engine demonstrates significant agreement with expert assessment and can therefore be considered a reliable source of formative feedback that learners can rely on.

Digital Credentials as Motivational Architecture

The credential motivation findings extend the empirical literature on badge-based motivation in several meaningful ways. The 84% agreement that credentials motivated task completion is among the highest reported in the literature for a non-gaming educational context (cf. Hamari et al., 2014; Abramovich et al., 2013), and the qualitative accounts of identity recognition and goal crystallisation suggest that the motivational mechanism operates at a deeper level than simple extrinsic reward—evidence that is consistent with the identified and integrated regulation modes of self-determination theory (Deci & Ryan, 1985; Ryan & Deci, 2000).

Equity and the Ethics of Tiered Assessment

The equity findings of this study represent perhaps its most significant contribution to the literature, and the most pressing implication for platform design and policy. The feedback cliff effect, the analytics information inequalities, and the credential access gap together constitute a pattern of structural disadvantage embedded in the tiered access model—one that systematically disadvantages precisely those learners who are most likely to need the platform's formative and credentialing affordances essentially.

Implications for Theory, Practice, and Policy

Theoretically, this study contributes to the ongoing development of a sociotechnical theory of AI-mediated formative assessment. The findings demonstrate that AI feedback is not simply a technical substitute for human feedback but a distinct pedagogical medium with its own affordances (immediacy, consistency, scalability, non-judgmental tone), limitations (task achievement validity, proficiency-adaptive language), and equity implications (differential access effects).

For practitioners, the study provides an empirically grounded model for integrating AI feedback within portfolio-based language assessment, demonstrating that this integration is both technically feasible and pedagogically effective at scale.

For policy, the equity findings constitute an argument for regulatory attention to the access conditions of AI-enabled educational tools. As AI assessment platforms increase in language education contexts globally, the risk of replicating and amplifying existing educational inequalities through tiered access models is real and significant.

Table 1. Independent Sample Size Effects.

		Standardizer[a]	Point Estimate	95% Confidence Interval	
				Lower	Upper
AI-based tools improve the overall quality of language teaching/learning.	Cohen's d	.996	0.019	-0.516	0.555
	Hedges' correction	1.010	0.019	-0.509	0.547
	Glass's delta	.793	0.024	-0.511	0.560

AI can help provide personalized learning experiences for students.	Cohen's d	1.042	-0.399	-0.938	0.143
	Hedges' correction	1.056	-0.394	-0.925	0.141
	Glass's delta	1.167	-0.356	-0.899	0.195
a. The denominator used in estimating the effect sizes. Cohen's d uses the pooled standard deviation. Hedges' correction uses the pooled standard deviation, plus a correction factor. Glass's delta uses the sample standard deviation of the control group.					

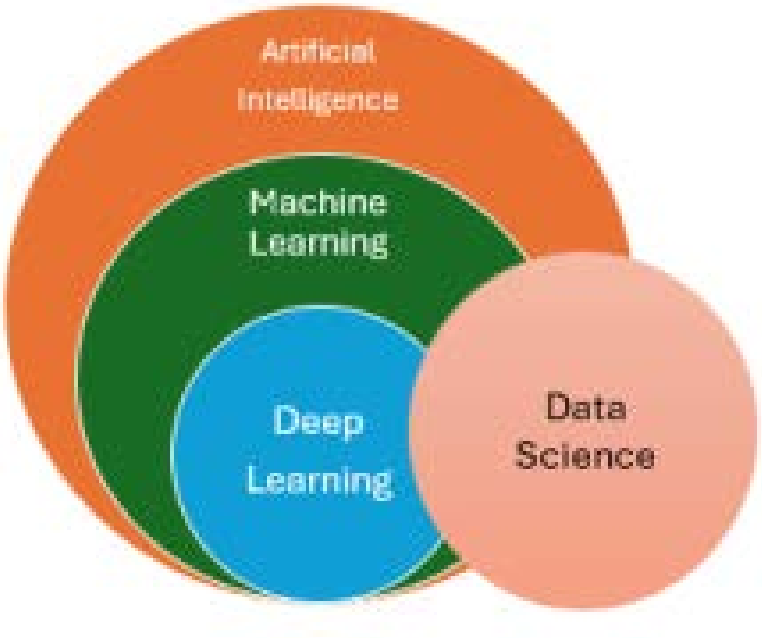

Figure 1: Subsets of AI

Conclusion

This study has reported on the design, implementation, and mixed-methods evaluation of the AI-Enabled Language Portfolio (AILP), an integrated platform combining AI-powered formative feedback, portfolio-based authentic assessment, and digital credentialing for adult language learners. The pilot study generated evidence across four research questions, producing findings that are both encouraging and cautionary in their implications for the field.

On the encouraging side: learner engagement with AI feedback substantially exceeded historical baselines; feedback immediacy emerged as a powerful driver of formative revision cycles; expert comparison confirmed substantial AI validity for grammatical and lexical assessment; and digital credentials demonstrated strong motivational effects consistent with self-determination theory's account of identified regulation. These findings collectively position AI-enabled e-portfolio systems as a promising and empirically grounded alternative to conventional language assessment in online and large-scale educational contexts.

The study's principal limitations include its single-platform, single-institution, eight-week pilot design, which limits generalisation of findings to other technological and institutional contexts. Future research should address these limitations through multi-site, longitudinal investigations; comparative studies of AI feedback validity across diverse L1 populations and proficiency levels; and empirical evaluation of the institutional endorsement strategies most effective in establishing digital credential credibility. Researchers should also attend to the long-term learning outcomes of AI-enabled portfolio assessment: a question this pilot study, by design, could not answer.

References

Abramovich, S., Schunn, C., & Higashi, R. M. (2013). Are badges useful in education? It depends upon the type of badge and expertise of learner. *Educational Technology Research and Development, 61*(2), 217–232. https://doi.org/10.1007/s11423-013-9289-2

Barrett, H. C. (2007). Researching electronic portfolios and learner engagement: The REFLECT initiative. *Journal of Adolescent & Adult Literacy, 50*(6), 436–449. https://doi.org/10.1598/JAAL.50.6.2

Beckett, G. H., Amaro-Jimenez, C., & Beckett, K. S. (2013). Students' use of asynchronous discussions for academic discourse socialisation. *Distance Education, 34*(1), 97–118.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health, 11*(4), 589–597.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27–40.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (pp. 206–210). ACL.

Cheng, W., & Warren, M. (2010). Making a difference: Using peers to assess individual students' contributions to a group project. *Teaching in Higher Education*, *5*(2), 243–255.

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

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification." In *Proceedings of the 15th International Academic MindTrek Conference* (pp. 9–15). ACM.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). ACL.

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs—Principles and practices. *Health Services Research*, *48*(6 Pt 2), 2134–2156.

Gallagher, S. R., & Palmer, J. (2020). The pandemic pushed universities online. The change was long overdue. *Harvard Business Review*.

Gibson, D., Ostashewski, N., Flintoff, K., Grant, S., & Knight, E. (2015). Digital badges in education. *Education and Information Technologies*, *20*(2), 403–410.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? A literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences* (pp. 3025–3034). IEEE.

Klenowski, V., Askew, S., & Carnell, E. (2006). Portfolios for learning, assessment and professional development in higher education. *Assessment & Evaluation in Higher Education, 31*(3), 267–286. https://doi.org/10.1080/02602930500352816

Landauer, T. K., Laham, D., & Foltz, P. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Lawrence Erlbaum Associates.

Little, D. (2009). Language learner autonomy and the European Language Portfolio: Two L2 English examples. *Language Teaching*, *42*(2), 222–233.

Llosa, L., & Malone, M. E. (2019). Alignment between CEFR levels and automated writing evaluation scores. *Language Testing*, *36*(3), 375–397.

Locker, L., & Jaques, P. (2018). Feedback and formative assessment in online learning environments. *Computers & Education*, *124*, 82–96.

Mozilla Foundation. (2011). *Open badges for lifelong learning*. Mozilla Foundation. https://wiki.mozilla.org/images/5/59/OpenBadges-Working-Paper_012312.pdf

Oliver, B. (2019). *Making micro-credentials work for learners, employers and providers*. Deakin University.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, *47*(5), 238–243.

Selingo, J. J. (2018). The false promises of worker retraining. *The Atlantic*.

Shen, F. (2025). A study of human-AI interaction patterns in artificial intelligence-assisted second language writing. *International Journal of English Literature and Social Sciences*, *10*(2).

Tashakkori, A., & Teddlie, C. (2010). *SAGE handbook of mixed methods in social and behavioral research* (2nd ed.). SAGE Publications.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3*(1), 22–36.

Yancey, K. B. (2009). Electronic portfolios a decade into the twenty-first century: What we know, what we need to know. *Peer Review*, *11*(1), 28–32.