FinLingPro: An AI-Gamified Learning Model for English for Financial Purposes among Accounting and Finance Students

1st Omer Mohamed Toha Alkanan
Lion Emirates Global-Hq Earth
New Zealand
alkanan1995@gmail.com

2nd Nelia Fariani,S
Management, Economics And Business
Swadaya Gunung Jati University
Cirebon, Indonesia
Neliafariani.2707@gmail.com

3rd Muhammad Irvan
Management, Economics And Business
Swadaya Gunung Jati University
Cirebon, Indonesia
irvanm249@gmail.com

4thM. Septian Arya Permana
Accounting, Economic And Business
Swadaya Gunung Jati University
Cirebon, Indonesia
mseptianaryapermana35@gmail.com

5thUly Aminarty
Accounting, Economic and Business
Swadaya Gunung Jati University
Cirebon, Indonesia
ulyaminartysubagja26@gmail.com

Abstract

This study presents *FinLingPro*, an AI- and gamification-integrated instructional model developed to enhance English for Financial Purposes (EFP) competencies among accounting and finance students. Grounded in the principles of English for Specific Purposes (ESP), the model incorporates adaptive Natural Language Processing (NLP) feedback and progressive gamified learning challenges. A quasi-experimental design involving an experimental group (n = 50) and a control group (n = 50) was conducted over an 8-week period. The results showed that students using FinLingPro achieved significantly higher gains in reading, writing, listening, and speaking within financial contexts, as well as greater accuracy in using financial terminology. The experimental group also exhibited increased motivation and sustained engagement, as indicated by platform analytics and motivation survey data. These findings suggest that FinLingPro effectively addresses pedagogical gaps in EFP instruction and offers a scalable solution for domain-specific language learning in higher education. The study concludes with recommendations for future model optimization and wider implementation across related disciplines.

Keywords *English for Financial Purposes, Artificial Intelligence, Gamification, ESP, Adaptive Learning*

I. INTRODUCTION

In contemporary finance education, English for Financial Purposes (EFP) is indispensable for tasks that range from parsing earnings releases and MD&A sections to interpreting analyst notes, drafting valuation memos, and briefing stakeholders. These genre-specific activities demand not only general proficiency but also command of numeracy-laden discourse (e.g., hedging language, risk qualifiers, guidance wording) and accurate deployment of discipline-specific terminology aligned with IFRS/GAAP conventions. As finance programs internationalize and workplace communication becomes increasingly cross-border, the cost of linguistic imprecision risesfrom misunderstanding disclosures to eroding the credibility of graduate job candidates in analyst, audit, and corporate finance roles.

However, conventional EFP instruction often falls short on three fronts. First, feedback is frequently generic and delayed, offering limited guidance on terminology accuracy, register, and pragmatics that are unique to financial genres. Second, practice opportunities can be fragmented across skills: reading and writing may be emphasized, while listening (e.g., earnings calls) and speaking (e.g., investor briefings) receive less systematic, authentic treatment. Third, sustaining motivation and persistence across multi-skill trajectories is challenging; learners may struggle to maintain deliberate practice without timely, granular signals of progress.

These shortcomings have tangible consequences. Students may misread tone in forward-looking statements, misuse key terms (e.g., impairment, materiality, liquidity), or overstate certainty in ways that conflict with risk-aware financial communication. In oral contexts, limited rehearsal with realistic constraints (time pressure, numeracy, clarifying questions) can depress fluency and confidence. At scale, programs face difficulty monitoring engagement and diagnosing where learning breaks downwhether at the level of genre structure, terminology selection, or discourse strategy.

Accordingly, there is a pressing need for an instructional approach that simultaneously delivers (i) domain-specific, timely feedback on language form and financial terminology, (ii) authentic, integrated practice across reading, writing, listening, and speaking, and (iii) mechanisms that sustain motivation and track progress with actionable learning analytics. This paper responds to that need by proposing and evaluating a structured model tailored to EFP, to be detailed in the subsequent sections.

Recent work in technology-enhanced language learning shows that Artificial Intelligenceespecially Natural Language Processing (NLP)can provide adaptive, fine-grained feedback on accuracy, usage, register, and discourse moves. In parallel, gamification has been linked to greater persistence and depth of practice through mechanics that cultivate competence, autonomy, and relatedness (as theorized in SDT) and support focused engagement (as described by Flow). Together, these strands suggest a plausible route to address two enduring challenges in English for Specific Purposes (ESP): delivering real-time, domain-sensitive feedback and sustaining motivation across multi-skill learning trajectories.

Within English for Financial Purposes (EFP) specifically, however, the empirical base remains fragmented. Studies often (i) target a single or limited set of skills (e.g., reading/writing) rather than an integrated four-skill approach; (ii) assess general language proficiency while overlooking discipline-specific terminology accuracy and pragmatic conventions of financial genres; and/or (iii) rely on self-report motivation without triangulating with learning analytics (e.g., time-on-task, level completion, retention). Moreover, rigorous causal estimation is uncommon: pretest–posttest designs may not test pre-existing equivalence or report effect sizes and assumption checks, limiting external validity and instructional transfer.

This study addresses those gaps with FinLingPro, an AI- and gamification-integrated model tailored to EFP. The novelty lies in: (1) a closed-loop NLP feedback pipeline that targets finance-specific lexis, register, and discourse features; (2) progressive mastery paths (quests/levels/badges) aligned to EFP outcomes, grounded in SDT/Flow principles; (3) comprehensive outcome modeling that spans four skills and financial terminology accuracy; (4) motivation and engagement measured through validated survey scales and platform analytics; and (5) a quasi-experimental evaluation with pretest equivalence checks, mixed ANOVA/ANCOVA, and effect sizes to support credible inferences for curriculum design. Collectively, these elements provide a replicable blueprint for domain-specific language instruction in finance while extending the evidence base on how AI and gamification interact to drive learning in ESP contexts.

This study evaluates whether FinLingProan AI- and gamification-integrated instructional modelimproves English for Financial Purposes (EFP) outcomes relative to conventional instruction. Primary endpoints are four-skill EFP performance (reading, writing, listening, speaking) and financial terminology accuracy; secondary endpoints are motivation and sustained engagement as captured by validated surveys and platform analytics (e.g., time-on-task, level completion, retention).

Research Questions (RQs).

- RQ1. Does FinLingPro yield greater gains in EFP reading, writing, listening, and speaking than conventional instruction?

- RQ2. Does FinLingPro increase accuracy and appropriate register of finance-specific terminology in student products?

- RQ3. Does FinLingPro improve motivation and sustained engagement compared with control?

Directional Hypotheses.

- H1. The experimental group outperforms the control on four-skill EFP posttest scores (controlling for pretest), with at least medium effect sizes.

- H2. The experimental group demonstrates higher financial terminology accuracy and genre-appropriate register.

- H3. The experimental group reports higher motivation and exhibits stronger engagement (greater time-on-task, higher level-completion, better retention).

Contributions.

- Theoretical. Extends ESP/EFP literature with causal evidence on an integrated AI + gamification model grounded in SDT/Flow for domain-specific language learning.

- Methodological. Offers a measurement blueprint combining four-skill performance, terminology-accuracy rubrics, motivation scales, and learning analytics, analyzed with transparent effect-size reporting.

- Practical. Delivers a replicable curriculum design (tasks, mastery paths, feedback loops) and actionable analytics indicators for instructors/program designers in finance education.

- Scope clarification. Findings speak to higher-education EFP contexts over an 8-week implementation; generalization to other ESP domains is discussed as future work.

II. METHOD

This study employed a quasi-experimental pretest–posttest design with a control group over eight weeks in an undergraduate accounting/finance program at [Institution, City, Country. Two intact course sections were assigned to

the experimental (FinLingPro) and control (business-as-usual) conditions, respectively. Instructional fidelity in the experimental section was monitored through platform logs (task availability, feedback events) and a weekly checklist completed by the instructor. Both sections covered identical EFP learning outcomes and core content (financial genres and tasks), differing only in the delivery model (FinLingPro versus conventional instruction).

Participants were [N = 100] second- or third-year students (experimental n = 50; control n = 50) enrolled in EFP coursework. Inclusion criteria were current enrollment and consent to participate; no exclusion was applied beyond incomplete pre/post data. Prior to the study, we conducted an a priori power analysis (repeated-measures, between–within interaction, f = 0.25, α = .05, 1−β = .80) indicating a minimum N ≈ 98 for adequate power. Participation was voluntary with written informed consent; data were anonymized and reported in aggregate. No monetary incentives were offered beyond course-embedded learning value.

Instruments

1. EFP Skill Assessments (four-skill battery).

- Reading. Timed comprehension of authentic financial texts (earnings releases/MD&A excerpts; 600–900 words) with multiple-choice and short constructed responses targeting gist, inference, numeracy-linked interpretation, and discourse signals.

- Listening. Audio clips from simulated/edited earnings calls and analyst briefings (2–3 minutes each) with items assessing gist, detail, speaker stance/hedging, and number processing.

- Writing. A 250–300-word earnings brief / valuation memo responding to a prompt (company snapshot + event). Scored with a rubric (0–4) on genre organization, argument clarity, terminology accuracy/precision, and linguistic control (grammar, register).

- Speaking. A 2–3 minute analyst-style oral briefing + 1 minute Q&A. Scored with a rubric (0–4) on message structure, fluency, stance/hedging appropriateness, terminology accuracy, and audience alignment.
  Two trained raters scored writing/speaking independently; inter-rater reliability was estimated via ICC(2,k). Disagreements ≥1 point were reconciled by discussion or a third rater.

2. Financial Terminology Accuracy Rubric.
   A task-agnostic rubric targeting term selection, form, and register in finance (e.g., *impairment, materiality, liquidity, guidance, YoY/ QoQ phrasing, non-GAAP adjustments*). Levels: 0 = incorrect/inappropriate, 1 = partly correct (form/register off), 2 = correct but inconsistent, 3 = consistently correct and context-appropriate. Applied to writing/speaking artifacts and sampled reading/listening responses where applicable.

3. Motivation Survey.
   A validated motivation scale adapted to the EFP context (e.g., Intrinsic Motivation Inventory subscales: interest/enjoyment, perceived competence, effort/importance; or MSLQ subsets). Items on a 5–7 point Likert scale; internal consistency reported (α/ω ≥ .70).

4. Engagement & Learning Analytics.
   Platform-derived indicators: time-on-task (minutes/week), level completion (% mastery goals achieved), retention (active weeks of use), feedback uptake (accepted/revised suggestions), and session frequency (log-ins/week). All analytics were pre-specified and computed with identical windows for both groups (the control section had access only to baseline LMS logs; analytics aligned on comparable proxies such as submission timestamps).

Procedure

- Week 0 (Orientation & Pretest). Consent, demographic/background survey, and pretests for the four skills plus the terminology diagnostic and motivation survey. Rater calibration (anchor scripts) conducted before scoring.

- Weeks 1–7 (Intervention).

  o Experimental (FinLingPro). Students completed weekly quests aligned to EFP outcomes (e.g., reading analyst notes → write earnings brief → deliver oral briefing). The platform delivered NLP-driven feedback on lexis/register/structure and unlocked progressive mastery paths (points, levels, badges). In-class facilitation emphasized deliberate practice and targeted mini-lessons from cohort error analytics.

o Control (Conventional). Same outcomes and topics via lectures, guided practice, and instructor feedback on drafts without AI-driven adaptivity or gamified mastery paths.
Fidelity checks: instructor checklist, platform usage snapshots, and a mid-course observation in each section.

- Week 8 (Posttest). Equivalent-form posttests for the four skills, terminology rubric scoring on final artifacts, and post motivation survey. Debriefing followed data collection.

Data Analysis

- Data Preparation. Screening for outliers and missingness; inclusion followed intention-to-treat principles. If item-level missingness ≤5%, we used mean-of-nearest neighbors within subscale; if >5%, we applied multiple imputation (m = 20) with group, pretest, and demographics as predictors. Normality and homogeneity were assessed (Shapiro–Wilk; Levene).

- Equivalence Checks. Pretest comparisons (t-tests/Mann–Whitney as needed) on skills, terminology, and motivation; baseline demographics compared via χ^2/Fisher.

- Primary Outcomes.

  o Four-skill EFP & Terminology. Mixed ANOVA with Time (pre, post) × Group (exp, ctrl); where pretest imbalance existed, ANCOVA models were estimated with pretest as covariate. Effects reported as partial η^2 and Cohen's d for gain scores; 95% CIs included.

  o Motivation & Engagement. Between-group comparisons on posttest (with pretest as covariate) and longitudinal models for analytics (weekly time-on-task, level completion) using linear mixed-effects with random intercepts for participants.

- Multiple Comparisons & Robustness. Familywise control via Holm adjustment. If assumptions were violated, we report robust alternatives (e.g., Yuen's trimmed-mean tests) and bootstrapped CIs (10,000 resamples).

- Reliability & Rater Agreement. Internal consistency (α/ω) for surveys; ICC(2,k) for writing/speaking; CIs for all reliability estimates.

- Sensitivity Analyses. Models with propensity-score covariates (e.g., GPA, prior English exposure) and dosage (platform minutes) to probe robustness in a quasi-experimental context.

III. RESULTS AND DISCUSSION

The two groups were comparable at baseline across all outcomes (reading, listening, writing, speaking, terminology, and motivation), with no significant pretest differences. Pre-intervention means clustered in the mid-range (≈55–62 on a 0–100 scale for EFP skills; ≈4.8–5.0 on a 1–7 scale for motivation), and dispersions were moderate (SD ≈ 8–10), indicating adequate variability for detecting change.

Instrument quality met accepted standards. The six-item motivation scale showed acceptable-to-good internal consistency (Cronbach's α ≥ .80), supporting its use for between-group comparisons. Performance ratings exhibited excellent inter-rater agreement on post-intervention writing and speaking, with ICC(2,k) ≥ .80, indicating stable scoring across raters. Together, these diagnostics support the validity of subsequent inferential analyses and suggest that observed post-intervention differences are unlikely to be artifacts of baseline imbalance or measurement unreliability.

After controlling for pre-test scores, ANCOVA showed consistent superiority for the experimental group (FinLingPro) on all outcomes. The post-test score differences were statistically significant for reading, listening, writing, speaking, and terminology accuracy (Holm-adjusted p<.05). Estimated effect sizes based on gain scores (Cohen's d) were overall in the moderate to large range, with the largest effects observed in writing and terminology accuracy, followed by reading, listening, and speaking.

The average gain profile (Post−Pre) showed a similar pattern (Figure 1): the FinLingPro group showed higher and more consistent improvements in all five outcomes compared to the control group. The partial effect size (partial η^2) of the ANCOVA model was moderate and increased to large for writing and terminology, indicating a substantial practical contribution, not just statistical significance.

Substantively, two direct implications can be drawn. First, the integration of adaptive NLP feedback appears to have the strongest impact on terminology accuracy and writing performance in the financial genrean area that indeed demands precise lexical selection and register. Second, improvements in listening and speaking suggest that

integrated exercises (e.g., earnings call simulations and analyst briefings) are supported by clear signals of progress, thereby also driving the transfer of oral skills. Overall, these results answer RQ1–RQ2 with consistent quantitative evidence, reinforcing FinLingPro's claim of effectiveness in multi-skill EFP competencies.

Adjusting for baseline motivation, ANCOVA showed a significant advantage for the experimental group, with a positive adjusted mean difference and moderate partial η^2, indicating practically meaningful improvements in students' motivational state. Learning-analytics corroborate these attitudinal gains: the weekly time-on-task slope and the level-completion slope were both significantly steeper for FinLingPro than for control, evidencing stronger persistence and mastery progression across the 8-week sequence. In addition, retention was higher in the experimental section (active weeks closer to 8/8), suggesting that the combined signals of progress and adaptive feedback helped sustain participation throughout the intervention.

Substantively, these results address RQ3. The patternhigher adjusted motivation, increasing practice time, faster completion of mastery targets, and stronger week-to-week retentionaligns with the intended design of FinLingPro: adaptive feedback that clarifies "what to fix next" and gamified mastery paths that make progress visible and rewarding. In short, the engagement infrastructure appears to convert motivational gains into more practice, which is consistent with the learning improvements reported in Sections III.B.

The largest effect patterns on writing and terminology are consistent with AI/automated writing evaluation (AWE) literature, which shows that automated feedback improves writing quality through adaptive suggestions specific to form, lexicon, and discourse organization; positive effects are also reported on writing accuracy and fluency in L2 contexts. The integration of NLP that checks financial terms and hedging registers directly targets the most difficult areas in EFPhence its strongest effect. The findings of moderate increases in listening/speaking are consistent with evidence that AI tools for language practice and oral assessment can improve performance and confidence, although the magnitude of the effect depends on the task design.

Increases in motivation, time-on-task, level-completion, and retention are consistent with SDT/Flow-based gamification mechanisms: elements such as points–levels–badges and progress indicators tend to strengthen feelings of competence, autonomy, and relatedness, which in turn promote persistence and learning outcomes. A recent meta-analysis in education found positive (though variable) effects on motivation and achievement when gamification design aligns with learning goals, rather than being merely cosmetic. The steeper weekly trajectory in the FinLingPro group demonstrates how this mechanism works.

Comparison with Prior Work

Compared to AI/AWE studies that generally focus on general writing (EAP/ESL), this paper expands the scope to financial terminology and the four skills in an integrated manner while tracking learning analytics; this combination is relatively rare in previous reports. In terms of effect size, our results (moderate–large) are consistent with recent AWE/AI reviews and meta-analyses reporting improvements in writing quality and accuracy; our additional contribution is simultaneous evidence on listening/speaking and domain-specific terminology. On gamification, our results are consistent with meta-analyses showing increases in motivation/engagement and, in many cases, achievementwith the caveat that success is highly dependent on design appropriateness.

Validity, Limitations, and Boundary Conditions

The quasi-experimental design opens up the potential for selection bias and teacher effect; we mitigated this through pretest controls (ANCOVA), high rater reliability (ICC(2,k)), and fidelity checks. The 8-week duration was sufficient to detect medium-term learning gains but did not test long-term retention. Generalization is primarily to higher education EFP; transfer to other ESP domains requires adaptation of terminology corpus and genre rubrics. The gamification literature also highlights risks of misaligned designs (e.g., context misfit, overjustification), making cross-course/institution replication important to test the limits of findings.

Practical Implications & Future Work

Implementation. Prioritize NLP feedback loops on terminology/register; design weekly mastery paths with visible progress indicators; establish simple analytics cut-offs (e.g., minutes/week, milestone levels) as early intervention signals.

Further evaluation. Conduct cross-class/campus RCTs, dosage analysis (minutes & feedback uptake), and ablation (AI-only vs gamification-only) to isolate component contributions.

Technical-pedagogical. Strengthen domain adaptation (IFRS/MD&A corpus, hedging phrases) and explore AI for scenario-based earnings call/analyst briefing oral practice with discourse cue-based automatic assessment. These

findings align with recent AI/GenAI research directions in language learning emphasizing task design appropriateness and evidence-based measurement.

IV. CONCLUSIONS

This study evaluated FinLingPro, an AI- and gamification-integrated instructional model for English for Financial Purposes (EFP). Across an eight-week quasi-experimental implementation, the experimental group outperformed the control on all targeted outcomes after controlling for baseline: reading, listening, writing, speaking, and financial terminology accuracy. Effects were largest for writing and terminology, and moderate for the other skills, indicating benefits that are both statistically and pedagogically meaningful.

Motivational and behavioral indicators supported these performance gains. The experimental group showed higher adjusted motivation, steeper weekly time-on-task and level-completion trajectories, and stronger retention across weeks. Together, these patterns align with the model's design logic: adaptive NLP feedback clarifies "what to fix next" at the level of domain-specific language and register, while gamified mastery paths render progress visible and rewardingpromoting persistence and deeper practice.

We draw three implications. First, coupling domain-aware NLP feedback with structured mastery paths is a scalable blueprint for EFP courses aiming to improve precision in financial discourse. Second, embedding simple learning-analytics cutoffs (e.g., minutes per week, level milestones) enables timely instructional responses for learners at risk of falling behind. Third, integrated four-skill task sequences (e.g., earnings-call → brief → oral briefing) appear to translate motivational gains into measurable performance improvements.

Limitations include the quasi-experimental design (with residual selection/teacher effects possible), an eight-week duration that does not test long-term retention, and a single higher-education context that may constrain generalization. Future work should pursue multi-site randomized trials, dosage and ablation studies (AI-only vs gamification-only), and domain adaptation of the NLP components using finance-specific corpora (e.g., IFRS/MD&A) to further strengthen both validity and transferability.

REFERENCES

Li, M., Liu, Q., & Gong, S. (2023). Examining the effectiveness of gamification as a tool for promoting teaching and learning: A meta-analysis. *Frontiers in Psychology, 14,* 1253549. https://doi.org/10.3389/fpsyg.2023.1253549

Zeng, J., Zhang, H., & Gao, M. (2024). Exploring the impact of gamification on students' academic performance: A meta-analysis. *British Journal of Educational Technology, 55*(5), 2263–2287. https://doi.org/10.1111/bjet.13471

Ryan, R. M., & Deci, E. L. (2020). *Self-determination theory: Basic psychological needs in motivation, development, and wellness.* Guilford Press.

Zhai, N., Ma, Y., & Zhang, J. (2023). The effectiveness of automated writing evaluation on EFL/ESL writing: A meta-analysis. *Journal of Educational Computing Research, 61*(8), 1903–1933. https://doi.org/10.1177/07356331221127300

Wei, P., Zhang, L., & Chen, W. (2023). The impact of automated writing evaluation on second language writing performance: A randomized controlled trial. *Frontiers in Psychology, 14,* 1249991. https://doi.org/10.3389/fpsyg.2023.1249991

Lee, S. (2025). Generative AI in the language classroom: A systematic review (2023–2024). *Interactive Learning Environments.* Advance online publication. https://doi.org/10.1080/10494820.2025.2498537

Wu, X. Y., Zhou, Y., & Chen, H. (2024). Artificial intelligence in L2 learning: A meta-analysis of effectiveness. *System, 125,* 102132. https://doi.org/10.1016/j.system.2024.102132

Lyu, B., & Xiao, J. (2025). Effectiveness of chatbots in improving language learning: A meta-analysis. *International Journal of Applied Linguistics.* Advance online publication. https://doi.org/10.1111/ijal.12668

Li, B., & Wang, Y. (2024). A systematic review of the first year of publications on ChatGPT in language learning. *Journal of Computer-Assisted Learning, 40*(6), 1503–1526. https://doi.org/10.1016/j.caeai.2024.100199