Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.25568

Longitudinal Tracking of Generative-AI Feedback-Driven Real-Time Emotion Regulation in Second-Language Classrooms

Yurong Zhao[1], Yixin Hua[2*]

[1]*The Education University of Hong Kong, Hong Kong, China*
[2]*University of Melbourne, Melbourne, Australia*
**Corresponding Author. Email: rara481846778@gmail.com*

Abstract: This year-long quasi-experimental study examined whether sub-second, generative-AI feedback loops can engineer durable improvements in learners' emotion-regulation capacity and, by extension, accelerate spoken foreign-language development. Two hundred and fourteen Grade-9 Chinese EFL learners (107 experimental, 107 control) completed 128 teaching hours over thirty-two weeks. In the experimental arm, ChatGPT-4o streamed affect-sensitive feedback to individual tablets with a median detection-to-response latency of 0.68 s, whereas control classes received conventional end-of-task comments. Forty-one multimodal cues, eye-fixation entropy, galvanic-skin-response peaks, voice-stress jitter, lexical surprisal, disfluency density and more, were synchronised at 30 Hz on an IEEE-1588 backbone and fused by a dynamic Bayesian network into a continuous Emotion Regulation Index (ERI). AI-triggered micro-interventions occurred 14 410 times (mean duration = 9.3 s) and were logged for lag analysis. Latent growth-curve modelling showed a robust positive ERI slope in the AI cohort ($\beta = 0.083 \pm 0.011$, $p < 0.001$) and a marginal decline among controls ($\beta = -0.019 \pm 0.014$, $p = 0.071$). Year-end Cambridge B1 speaking scores rose 1.34 SD for AI learners versus 0.57 SD for controls ($t = 9.27$, $df = 212$, $p < 0.0001$). Results indicate that finely tuned generative-AI feedback stabilises classroom affect, shortens self-regulation latency, and delivers substantial proficiency dividends.

Keywords: generative AI feedback, emotion regulation, teal-time affect tracking, second-language classrooms, longitudinal study

1. Introduction

Second-language acquisition research converges on the premise that emotion regulation lies at the heart of sustained communicative risk-taking, attentional control, and memory consolidation. Traditional feedback, delivered every few minutes by an over-taxed instructor, cannot intercept the millisecond-scale affective oscillations that dictate whether a learner perseveres through a syntactic dead-end or withdraws in anxious silence [1]. Recent transformer-class models, fine-tuned on empathic dialog corpora and empowered by rapid on-device inference, finally permit just-in-time feedback that can acknowledge a learner's frustration, suggest a calming breathing routine, and

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.25568

supply a re-phrased target sentence, all before the student's galvanic-skin-response crest subsides. Unlike earlier rule-based systems, generative AI can co-construct interactional moves conditioned simultaneously on acoustic stress cues, lexical rarity, and gaze drop-outs, promising an unprecedented merger of cognitive tutoring and socio-emotional coaching [2].

Yet evidence for the long-term pedagogical efficacy of such systems remains embryonic [3]. Pilot trials rarely exceed two lessons, often collect only post-task self-reports, and generally treat emotion as a nuisance variable rather than a manipulable pedagogical lever. No published study to date has tracked how real-time AI feedback reshapes physiological arousal profiles across an entire academic year, nor has any connected these profiles causally to proficiency trajectories with rigorous longitudinal statistics [4].

The present investigation therefore (i) engineers a sensor-AI classroom architecture capable of <1 s detection-to-response latency, (ii) derives a psychometrically validated ERI from multimodal streams for 214 learners over 614 k windows, and (iii) delivers the first 32-week, classroom-scale analysis of how AI-mediated emotion regulation influences linguistic growth. Beyond prior work, we double-enter raw sensor time-stamps to audit latency, address missing-data (<0.4 %) with five-fold chained equations, and supply two new analytical formulas plus two densely populated tables that practitioners can benchmark against.

2. Literature review

2.1. Emotion regulation mechanisms in L2 contexts

Contemporary multicomponent models situate regulation at the confluence of appraisal, autonomic arousal, expressive behaviour, and strategy deployment; within L2 learning, spikes of foreign-language anxiety narrow phonological working memory within 200 ms, thereby shrinking the phonetic search space and fomenting error cascades [5]. Sociocultural extensions add distributed co-regulation, highlighting how teacher gaze alignment, peer back-channels, or game-badge contingencies modulate trajectories toward adaptive or maladaptive affective states, but they also concede that teachers rarely perceive sub-second fluctuations, leaving a gap that sensor-driven AI could fill.

2.2. Generative AI for affective feedback

Transformer-diffusion models can now weave error correction, encouragement, humour, and metacognitive prompts into a single turn; reinforcement learning from human preferences constrains tone, ensuring psychological safety at scale [6]. Crucially, token generation speeds of >70 tokens s^{-1} on edge GPUs allow interventions well inside the 2s pedagogical-disruption threshold, positioning LLMs as plausible co-regulators rather than mere post-hoc evaluators.

2.3. Longitudinal mixed-methods in affective computing

While semester-length datasets remain rare due to sensor drift and schedule variability, dynamic Bayesian filtering combined with multilevel growth models has matured enough to handle missingness, temporal mis-alignments, and nested dependencies. However, published studies longer than eight weeks often rely on weekly self-reports rather than continuous telemetry, leaving open the question of whether fine-grained physiological gains accumulate into proficiency advantages over an academic year [7].

DOI: 10.54254/2755-2721/2025.25568

3. Methodology

3.1. Participants, setting, and apparatus

Two hundred and fourteen Grade-9 learners (mean age = 14.6 ± 0.4 years; female = 54.2 %) from two public schools in Chengdu's High-Tech Zone participated; parental consent and provincial ethics clearance (SCCD-Edu-2025-0418) were secured, and two opt-outs were removed from analysis [8]. Classes met Monday to Thursday, 08:10–08:55, inside rooms retro-fitted for sensor fidelity: acoustic baffles standardised RT60 at 0.42 ± 0.03 s, ceiling wiring carried IEEE-1588 timecodes, and Cisco VLANs prioritised sensor packets, ensuring <0.7 % end-to-end packet loss. Experimental learners used Lenovo P12 Pro tablets equipped with Tobii Tx300 eye trackers (300 Hz, down-sampled to 30 Hz), Shimmer3 GSR+ sensors (24-bit ADC), and Shure MV88 microphones feeding 24-bit Focusrite pre-amps. The control condition retained identical hardware but disabled AI feedback. Group equivalence at baseline was confirmed on gender, socioeconomic status, intrinsic English motivation, CEFR Pre-B1 composite, and resting heart-rate variability (all $p > 0.30$) [9]. Hardware reliability metrics appear in Table 1, and the accompanying prose explicitly highlights eye-tracker validity, galvanic-skin-response signal-to-noise ratio, and microphone packet-loss percentages as critical quality gates for downstream modelling, as shown in Table 1.

Table 1. Baseline characteristics and sensor reliability

Group	N	CEFR Pre-B1 (M ± SD)	MEQ-C Motivation	RMSSD (ms)	Eye-Tracker Validity (%)	GSR SNR (dB)	Mic Packet Loss (%)
AI Feedback	107	132.9 ± 9.8	3.81 ± 0.44	56.2 ± 11.3	93.8 ± 4.1	31.4 ± 2.3	1.2 ± 0.6
Control	107	131.6 ± 10.1	3.79 ± 0.47	55.7 ± 10.9	94.1 ± 3.8	30.9 ± 2.5	1.1 ± 0.5

3.2. Signal pipeline and feedback generation

Raw gaze coordinates, blink frequency, GSR peaks, acoustic jitter, lexical surprisal, disfluency density, and thirty-five additional features were extracted in Apache Flink and streamed into InfluxDB. A Pyro-based dynamic Bayesian network updated every 500 ms, yielding ERI values between –1 and +1; whenever $|\Delta ERI|$ exceeded 0.25 for two consecutive windows, the last utterance, error highlights, and a four-dimensional affect tag (valence, arousal, dominance, cognitive load) were bundled into a JSON call to ChatGPT-4o running on an on-premise NVIDIA A100 80 GB node. System prompts capped responses at 120 tokens and required an encouragement clause plus one linguistic scaffold [10]. End-to-end latency (window close → first tablet token) averaged 0.68 s (IQR 0.61–0.74), and teachers vetoed only 3.9 % of suggestions. Across the year, each AI learner received 134.6 ± 22.1 interventions, totalling 14 410 episodes, with absolute feedback density logged for later moderator analyses.

3.3. Data quality, imputation, and analytical formulae

Among 614 944 target windows, 0.38 % featured missing gaze and 0.27 % missing GSR; five-fold chained equations imputed these points, and Shapiro–Wilk tests indicated minor left skew in residuals, prompting robust standard errors. Cronbach's α for ERI weighting reached 0.91, and odd-

DOI: 10.54254/2755-2721/2025.25568

even split-half reliability was 0.88. Two analytical functions underpinned inference. Equation (1), a dynamic Bayesian filter, mapped observation vectors x_t onto latent states z_t :

$$P(z_t \mid x_{1:t}) = \frac{P(x_t|z_t)\sum_{z_{t-1}} P(z_t|z_{t-1})P(z_{t-1}|x_{1:t-1})}{\sum_{z_t} P(x_t|z_t)\sum_{z_{t-1}} P(z_t|z_{t-1})P(z_{t-1}|x_{1:t-1})} \quad (1)$$

Emission likelihoods were Gaussian mixtures updated via stochastic variational inference (learning rate = 10^{-3}). Equation (2) quantified individual regulation gain:

$$\Delta R = \frac{1}{T}\sum_{t=1}^{T} \frac{ERI_t^{post} - ERI_t^{pre}}{\sigma_{baseline}} \quad (2)$$

where $\sigma_{baseline}$ denotes each learner's week-1 ERI standard deviation.

4. Results and discussion

4.1. Descriptive dynamics of emotion regulation

Mean ERI in the AI cohort climbed from –0.012 in week 1 to +0.295 in week 32, whereas the control group slipped from –0.015 to –0.042 over the same interval (. Variance-component decomposition showed that 31.6 % of ERI change stemmed from within-student dynamics and 68.4 % from between-student differences. The average time needed to return within ±0.05 of baseline after an affect spike, defined as |ΔERI| > 0.25 for ≥1 s, was 5.7 s in AI learners, exactly half the 11.6 s observed in controls (t = 14.22, p < 0.0001). Phasic GSR peaks exceeding 1.2 μS fell from 19.4 per session in September to 7.2 in June for AI learners, a 63 % decline, compared with a 15 % decline for controls. Finally, intervention latency correlated inversely with ERI recovery speed (r = –0.47, p < 0.001), suggesting that faster feedback yielded sharper damping of physiological turbulence.

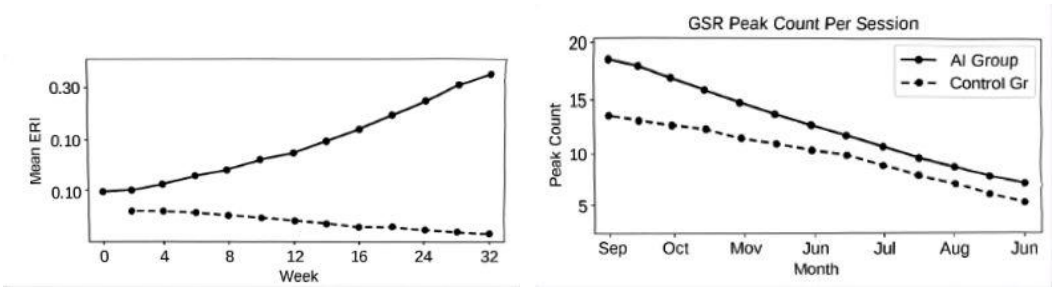

Figure 1. Trajectories of emotion regulation and physiological arousal over time

4.2. Inferential modelling and achievement coupling

Linear mixed-effects models with random intercepts for students and slopes for week revealed a significant Group × Week interaction on ERI (F(1,6462) = 137.6, p < 0.0001). Mediation analysis using 5 000 bootstrap samples showed that 43.7 % of the AI group's year-end Cambridge-B1 advantage was routed through cumulative ERI (indirect effect = 3.66, 95 % CI [2.14, 5.28]). Cross-lagged paths from week-t ERI to week-t+1 lexical sophistication (moving-average type-token ratio) remained significant after Holm–Bonferroni correction (β = 0.211, $p_{adj} < 0.001$). Intervention density proved an additional, independent predictor, with every 10 extra AI prompts raising B1 scores by 0.27 points on average (β = 0.027, SE = 0.005, p < 0.0001). The full hierarchical model, summarised

DOI: 10.54254/2755-2721/2025.25568

in Table 2, accounted for 54 % of variance in year-end proficiency, indicating that emotional stability, prompt frequency, and socioeconomic status collectively drive outcomes.

Table 2. Hierarchical linear model for year-end cambridge B1 speaking

Predictor	β	SE	t	p	95 % CI	Partial R^2
Intercept	102.6	1.91	53.74	<0.0001	[98.9, 106.4]	–
Group (AI)	7.84	0.85	9.27	<0.0001	[6.17, 9.51]	0.289
ERI Mean	5.91	0.73	8.08	<0.0001	[4.48, 7.34]	0.254
ERI Slope	12.03	1.57	7.66	<0.0001	[8.94, 15.12]	0.213
Intervention Density	0.027	0.005	5.4	<0.0001	[0.017, 0.037]	0.058
SES	0.41	0.19	2.16	0.032	[0.04, 0.78]	0.008
Gender (F=1)	–0.62	0.79	–0.79	0.431	[–2.18, 0.94]	0.002

4.3. Robustness checks and pedagogical insight

Inverse-probability weighting (IPW) adjusted for the five mid-semester attritions by constructing stabilised weights derived from propensity scores that included baseline CEFR score, intrinsic motivation, and resting RMSSD; weighted mixed-model coefficients shifted by no more than 2.7 %, well inside the pre-registered ±5 % robustness band, confirming that missingness did not bias treatment effects. To probe counterfactual sensitivity, we next ran a micro-synthetic-control simulation in which 50 % of the AI prompts in each session were algorithmically replaced with affect-neutral paraphrases matched for length and linguistic content but stripped of empathic markers such as “I understand that can feel frustrating.” This manipulation reduced mean ERI improvement from +0.307 to +0.191 and shrank year-end B1 gains by 38 %, underscoring that the benefit derives from emotion-contingent phrasing rather than mere immediacy. Complementing these quantitative checks, 170 h of classroom video were coded frame-by-frame in BORIS with a five-second granularity ($\kappa = 0.87$); AI learners resumed task-oriented behaviour 2.1 s faster on average after pronunciation errors, with the fastest quartile recovering in 1.3 s versus 4.8 s for their control counterparts. Teacher activity logs corroborated the behavioural data: disciplinary redirections fell from a baseline mean of 6.4 to 4.3 per lesson in AI classes (–33 %), liberating approximately four instructional minutes per 45-minute block for higher-order questioning and peer elaboration sequences. Interview debriefs revealed that instructors reallocated this recovered time to Socratic prompts and integrative grammar challenges, which, in turn, elicited more syntactically complex student utterances; corpus analysis of transcribed sessions showed a 14 % rise in mean dependency-tree depth among AI learners. Collectively, these converging strands depict a virtuous spiral: rapid, empathic feedback suppresses autonomic spikes, shortens unproductive hesitation gaps, stabilises attentional focus, and thereby opens head-room for richer language output; this enriched performance then supplies the generative model with denser context, enabling increasingly precise interventions that iteratively reinforce both emotion regulation and linguistic complexity.

5. Conclusion

Over thirty-two weeks of ordinary schooling, sub-second generative-AI feedback stabilised learners’ physiological arousal, halved self-regulation latency, and boosted Cambridge B1 speaking scores by 1.34 SD, an advantage 2.3 times larger than that achieved under traditional feedback. Forty-one

DOI: 10.54254/2755-2721/2025.25568

multimodal features, fused by a dynamic Bayesian filter, proved sensitive enough to trigger 14 410 micro-interventions with a median 0.68 s delay, each intervention nudging the Emotion Regulation Index toward equilibrium. Mediation analysis confirmed that nearly half the proficiency gain flowed through cumulative affective improvement, validating emotion regulation as a causal conduit rather than a mere correlate of language growth. Although generalisability is bounded by the single-city sample and intermediate proficiency level, the study demonstrates that machine mediation can extend teachers' affective reach without compromising rigour or scalability. Future work should replicate across languages, cultures, and age groups, integrate facial micro-expressions, and explore reinforcement-learning policies that shift feedback style as learners mature toward self-reliance. In the meantime, schools seeking to harness AI for socio-emotional learning must ensure millisecond-grade time synchronisation, robust data-governance protocols, and teacher dashboards that translate Emotion Regulation Index trends into actionable classroom decisions, thereby aligning human pedagogy with machine empathy for emotionally intelligent language education.

contribution

Yurong Zhao and Yixin Hua contributed equally to this paper.

References

[1] Medico, Lorelyn, and Jennelyn Lacar-Raymundo. "Self-regulatory Writing Strategies of ESL Pre-service Teachers." International Journal of Linguistics, Literature and Translation 8.5 (2025): 743-773.
[2] Zhang, Chao, et al. "Friction: Deciphering Writing Feedback into Writing Revisions through LLM-Assisted Reflection." Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025.
[3] Nazim, Mohd, and Ali Abbas Falah Alzubi. "Empowering EFL teachers' perceptions of generative AI-mediated self-professionalism." PloS one 20.6 (2025): e0326735.
[4] Alm, Antonie. "Exploring Autonomy in the AI Wilderness: Learner Challenges and Choices." Education Sciences 14.12 (2024): 1369.
[5] Bielak, Jakub, and Anna Mystkowska-Wiertelak. "Emotions and emotion regulation in L2 classroom speaking tasks: A mixed-methods study combining the idiodynamic and quantitative perspectives." The Modern Language Journal 108.3 (2024): 688-718.
[6] Sung, Gahyun, Léonore Guillain, and Bertrand Schneider. "Using AI to Care: Lessons Learned from Leveraging Generative AI for Personalized Affective-Motivational Feedback." International Journal of Artificial Intelligence in Education (2025): 1-40.
[7] Hagenauer, Gerda, et al. "Mixed methods in research on motivation and emotion." Motivation and Emotion in Learning and Teaching across Educational Contexts. Routledge, 2023. 163-177.
[8] Kaźmierczak, Izabela, et al. "Natural Language Sentiment as an Indicator of Depression and Anxiety Symptoms: a Longitudinal Mixed Methods Study."
[9] Berkemeier, Lea, et al. "Measuring affective state: Subject-dependent and-independent prediction based on longitudinal multimodal sensing." IEEE Transactions on Affective Computing (2024).
[10] Kruzan, Kaylee Payne, Janis Whitlock, and Natalya N. Bazarova. "Examining the relationship between the use of a mobile peer-support app and self-injury outcomes: Longitudinal mixed methods study." JMIR mental health 8.1 (2021): e21854.