

A Corpus-Centric Deep Learning Framework for AI-Assisted English Language Learning: NLP-Based Analysis of Learner Errors and Personalized Feedback

Shabnam Thakur¹, Shalini Sharma², Sukhmani Kaur³

¹ Research Scholar, Rayat & Bahra University, Mohali, Punjab, India, thakur.shabnam1@gmail.com

² Department of English, UILAH, Chandigarh University, Mohali, Punjab, India, shalinisharma.cgc@gmail.com; shalini@cumail.in

³ Department of Commerce, USB, Chandigarh University, Mohali, Punjab, India, sukhahal313@gmail.com

Abstract

The integration of artificial intelligence (AI), natural language processing (NLP), and large language models (LLMs) in English language learning is an emerging trend that is reshaping how the language is taught and learned. Artificial intelligence (AI), natural language processing (NLP), and large language models (LLMs) are revolutionizing the field of English language learning by facilitating automated assessment, conversational interactions, personalized feedback, and adaptive learning paths. Most of the current studies, however, study individual AI tools or pedagogical perceptions without providing a corpus-centric computational framework that links learner language data, linguistic error analysis, deep learning, and personalized feedback. English as an additional language learners make systematic errors with a first language structure and most AI writing tools give general corrections that do not identify errors in the learner's L1 structure and cannot evolve based on it. It is proposed in this paper that (a) the error corpus can be annotated with L1 and learners' proficiency levels and used to train a multi-task transformer that (b) identifies and classifies errors, (c) represents the L2 error trajectory for each learner and (d) provides a teacher-supervised system that generates personalised explanatory explanatory, bilingual-ready feedback. A Phase 1 survey of 287 learners across five districts was analysed as a problem baseline. Overall, perceptions of AI support were low (63.3% of all ratings 1–2 on a 5-point scale; 12.9% were 4–5), were not correlated with self-reported proficiency, were lower among AI users than among non-users (rank-biserial $r = .19$, $p = .007$), and showed one dominant dimension (42.4% of variance; $\alpha = .95$). The best predictor of perceived skill effectiveness was perceived personalised feedback ($\beta = .44$, $R^2 = .52$). Perception data alone was not sufficient to predict proficiency—neural classifiers failed to achieve a AUC score $> .55$ using the 26 survey items. Phase 2 (corpus and model) is completely planned and Phase 3 (quasi-experiment) has pre-planned outcomes. No Phase 2-3 result are reported.

Keywords: *Learner Corpus, Grammar Error Correction, Deep Learning, Natural Language Processing, Personalized Feedback, L1 transfer, EAL, India.*

1. Introduction

Errors made by L2 speakers are not accidental noises; they are indications of a developing interlanguage (L1 + the L2 learner's partially acquired knowledge of L2 rules) (Corder, 1967; Selinker, 1972). Articles in English do not correspond with any word in Hindi or Pahari; prepositions are used as postpositions and tense–aspect is not the same. So, if it is an omission of an article or a substitution of a preposition or a translation word-for-word, these can be expected and are not accidental. Learner corpora enable such patterns to become measurable and can be used to serve individual learners' needs (Granger, 2015; Gilquin, 2024), with proficiency-rated learner corpora enabling a connection between error patterns and developmental stage (Forti, 2024).

Meanwhile, the power of machines to do with such data has been revolutionized by a new technology called natural language processing (NLP). For the last few years, grammatical error correction (GEC) shifted from rules and statistical classifiers to neural machine translation and sequence-tagging systems (Bryant et al., 2023), with shared tasks and datasets like BEA-2019 (Bryant et al., 2019). Large language models (LLMs) can now produce text-based corrective feedback that matches trained human raters on a number of criteria in controlled comparisons (Steiss et al., 2024), and can be used in essay feedback without reducing learning outcomes (Escalante et al., 2023), though there are concerns about accuracy, over-reliance, and equity (Barrot, 2023; Kasneci et al., 2023; Pack et al., 2025). There is evidence that AI-assisted teaching improves achievement, motivation and self-regulation for EFL learners (Qiao & Zhao, 2023; Wei, 2023), and the research agenda in EFL is now increasingly focused on the concept of personalisation (Huang et al., 2023).

Though this is a step forward, three issues are severe in places where English is a second language, and digital resources are not available in all areas, including Himachal Pradesh. However, in the context of English being an additional language and digital resources unevenly distributed, three issues are severe in places such as Himachal Pradesh. First, the "general purpose tools" correct the forms of the surface, but rarely explain why a given error was made by a Hindi/Pahari speaker. Secondly, they do not capture if and when the same error surface again, thus feedback is not cumulative. Thirdly, models are trained and tested on learner populations that generally do not contain these learners, and their accuracy and fairness for

these learners is not known. Multilingualism is emphasized in the National Education Policy 2020 (Ministry of Education, 2020), as is the importance of technology-supported learning, raising the bar for what is at stake in getting this right.

Therefore, the current study proposes and tests a Corpus-Centric Deep Learning Framework (CC-DLF) which ties together four layers: learner-error corpus, deep-learning error detector/classifier, longitudinal learner model and feedback generator. A first-phase survey (n = 287; Section 4) quantifies the baseline problem: that is, that most of the sample rate learners in AI support are perceived as largely ineffective. The study asks:

- RQ1. What is the perception of non-native learners of AI support for language skills and feedback and how is this different across proficiency, AI use, L1, district, field and age? (Phase 1 – answered here.)
- RQ2. What types of errors and L1-transfer patterns are found in their writing, and can a multi-task transformer that is trained with L1 (along with proficiency) perform better at detecting and categorising these errors than the L1-agnostic baselines? (Phase 2.)
- RQ3. Does a learner-model-driven, explanatory feedback system, that is bilingual ready, positively affect grammatical accuracy, error recurrence and feedback quality as perceived by the learner compared to generic AI feedback and business-as-usual teaching? (Phase 3.)
- RQ4. Are performance levels fair by L1, district and proficiency subgroups? (Phases 2–3.)

It has three contributions: (1) an annotated L1-aware learner corpus for an under-represented Indian context; (2) a reproducible architecture that connects error detection with learner modelling and pedagogically structured feedback; and (3) an evaluation design that provides accuracy, learning gain and subgroup equity.

2. Literature review and research gap

This study is a convergence of three artistic endeavors. Error analysis and corpus-based analysis of learner errors reveal that errors are measurable and systematic (Corder, 1967; Selinker, 1972; Granger, 2015). Computational GEC has yielded good detectors and correctors (Omelianchuk et al., 2020; Bryant et al., 2023). Motivation and achievements in AI-assisted language learning (Qiao & Zhao, 2023; Wei, 2023) and LLM feedback as good but not perfect (Escalante et al., 2023; Steiss et al., 2024) are reported. Table 1 summarises 15 sources and the resulting gaps are stated in Table 2.

Table 1. Literature matrix

Authors (year)	Venue	Approach	Main finding	Limitation relative to this study
Bryant et al. (2023)	<i>Computational Linguistics</i> (Scopus)	Narrative survey: rules → SMT → neural MT / tagging	Neural approaches dominate; shared tasks and artificial error generation drove progress	Optimised for correction accuracy; little on explanation, learner modelling or L1- specific populations
Rozovskaya et al. (2017)	<i>Computational Linguistics</i> (Scopus)	Pre-neural classifiers adapted to L1-specific error patterns with minimal supervision	Adapting to the writer's L1 error distribution improves correction	Pre-deep- learning; correction only, no feedback generation or longitudinal modelling; no Hindi/Pahari data

Authors (year)	Venue	Approach	Main finding	Limitation relative to this study
Escalante et al. (2023)	<i>Int. J. Educ. Technol. in Higher Educ.</i> (Scopus)	Quasi-experiment: GPT-4 vs human-tutor feedback	AI-generated feedback can be incorporated into ENL essay feedback without harming outcomes; preferences varied	Short horizon; generic prompts; no error taxonomy, L1 conditioning or recurrence tracking
Steiss et al. (2024)	<i>Learning and Instruction</i> (Scopus)	Rubric-based comparison of feedback quality	Trained humans better on 4 of 5 criteria; ChatGPT close; quality did not vary by language status	Feedback quality only; no learning outcomes; not an L1-specific or low-resource context
Pack et al. (2025)	<i>Int. J. English for Academic Purposes</i> (supplied reference; verify Scopus status)	Evaluation of GPT-4 written corrective feedback	Assesses how well GenAI can provide WCF (see paper for accuracy figures)	Single-source feedback evaluation; no learner model or Indian L1 context (inferred from title/abstract)

Authors (year)	Venue	Approach	Main finding	Limitation relative to this study
Wei (2023)	<i>Frontiers in Psychology</i> (Scopus)	Mixed-methods quasi-experiment, AI-mediated vs traditional teaching	Higher achievement, L2 motivation and self-regulated learning in AI group	Small, intact classes; platform not error-driven; no feedback-level mechanism analysis
Qiao & Zhao (2023)	<i>Frontiers in Psychology</i> (Scopus)	RCT using Duolingo (NLP, speech recognition, personalised feedback)	Improved speaking skills and speaking self-regulation	Speaking focus; commercial app is a black box; no written-error diagnosis; long-term effects untested
Huang et al. (2023)	<i>Educational Technology & Society</i> (Scopus)	Systematic review of trends, issues, applications	Maps AI applications and research issues; indicates need for more learner-centred evidence	Secondary synthesis; no empirical error modelling
Barrot (2023)	<i>Assessing Writing</i> (Scopus)	Critical/conceptual analysis of ChatGPT	Identifies potentials (feedback, ideation) and pitfalls (accuracy, integrity)	Conceptual; no data from non-native learners

Authors (year)	Venue	Approach	Main finding	Limitation relative to this study
Kasneji et al. (2023)	<i>Learning and Individual Differences</i> (Scopus)	Interdisciplinary position paper	Opportunities for personalisation alongside risks of bias, over-reliance and reduced critical thinking	Conceptual; no language-learning corpus or intervention
Ranalli (2018)	<i>Computer Assisted Language Learning</i> (Scopus)	Study of learners' uptake of automated written corrective feedback	Usefulness depends on diagnostic accuracy and how well learners can act on feedback	Rule-based AWE era; no deep-learning diagnosis or personalisation
Gilquin (2024); Forti (2024)	<i>Babylonia; Int. J. Learner Corpus Research</i> (supplied references; verify Scopus status)	Conceptual / corpus-methodological papers	Learner corpora can serve individual needs; proficiency ratings enrich corpora	Corpus resources are not coupled to neural feedback generation
Behzad et al. (2024)	NAACL 2024 short papers (conference; supplied reference)	Corpus resource (LEAF)	Provides learner essays paired with feedback for NLP use	Not India-specific; no L1-transfer annotation; resource, not a framework

Authors (year)	Venue	Approach	Main finding	Limitation relative to this study
Yang et al. (2025)	<i>Int. J. Web- Based Learning and Teaching Technologies</i> (supplied reference; verify Scopus status)	Review/study of AI for writing enhancement	AI tools support EFL writing enhancement	Focus on tool use rather than error-level modelling (inferred from title)

Note. “Limitation” entries are the author’s synthesis relative to this study’s aims.

Research gap

To the best of the author's knowledge, none of the studies reviewed provide a learner-error corpus annotated with L1 and proficiency level, a deep-learning detector and classifier based on L1 and proficiency, a long-term model of error recurrence in the learner and a teacher-supervised, explanatory and bilingual-ready feedback, and assess its impact on learning gain and subgroup equity. Research on LLM-feedback is designed to evaluate generic feedback in other populations (Escalante et al., 2023; Pack et al., 2025; Steiss et al., 2024), while research on AI-learning reports results without error-level mechanisms (Qiao & Zhao, 2023; Wei, 2023). Studies about LLM-feedback do not involve any error-level mechanisms and studies about AI-learning report results without error-level mechanisms (Bryant et al., 2023; Rozovskaya et al., 2017; Qiao & Zhao, 2023; Wei, 2023). The Phase 1 data provide first-hand insights into the issue: a learner's current AI support is perceived as ineffective, and perceptions do not correlate with proficiency.

The claim of absence is restricted to the sources reviewed and has been substantiated with the search string in Appendix B.

Table 2. The gaps, evidence and response to the study

#	Gap	Evidence from the literature and Phase 1	Response in this study
G1	No L1-aware learner-error corpus for Hindi/Pahari-speaking learners in Himachal Pradesh	Reviewed studies use corpora from other populations (Behzad et al., 2024; Bryant et al., 2019); L1-adaptation exists (Rozovskaya et al., 2017) but not for this population	Annotated HP learner-error corpus with L1-transfer tags (Phase 2)
G2	Detection and correction are decoupled from explanation: tools return <i>what</i> to change, not <i>why</i> for this learner	Surveys centre on correction accuracy (Bryant et al., 2023); LLM-feedback studies assess quality of generic feedback (Steiss et al., 2024; Pack et al., 2025)	Type + transfer-aware, rule-grounded explanations in Hindi–English (Layers 2 and 4)
G3	No learner model: error recurrence and mastery are not tracked, so feedback is not cumulative	Intervention studies measure outcomes at pre/post only (Escalante et al., 2023; Wei, 2023)	Longitudinal sequence model of error events (Layer 3)
G4	Evidence is perception- or tool-level; it does not link perceived usefulness to error-level data	Our Phase 1 data: perceptions are unidimensional and cannot predict proficiency (Section 4)	Joint analysis of survey, corpus and learning gain
G5	Equity and fairness for non-native subgroups are rarely reported	Few reviewed studies report subgroup performance; Kasneci et al. (2023) flag bias risks	Pre-declared subgroup gap criterion (Section 5)
G6	Over-reliance and trust are under-examined where teacher capacity is limited	Barrot (2023) and Kasneci et al. (2023) warn of pitfalls; few designs include teacher-in-the-loop review	Human-in-the-loop dashboard; FB7/FB11 items

2.3 Conceptual framework and hypotheses:

The framework is based on the structure of feedback as a three question process: Where are they going, How are they going and Where next? (Hattie & Timperley, 2007) and applies it computationally. Layer 1 (corpus) provides evidence; Layer 2 (detector) provides answers to the question 'how am I going?'; Layer 3 (learner model) provides answers to the question 'where next?'; Layer 4 (generator) provides answers to the question 'how does this explain it?'. The hypothesis is written for each phase, H1 is tested using the data from Phase 1 and the others are pre-declared.

H1. The relationship between perceived language-skill effectiveness and perceived personalised feedback and error diagnosis remains after controlling for other perceptions. (Tested – supported, Section 4.4.)

H2. The conditioning of L1 and proficiency increases the span level F0.5 compared to an L1-agnostic baseline, particularly for transfer error types.

H3. The adjusted accuracy of learners who receive CC-DLF feedback is higher than those who receive generic AI feedback which is higher than the business-as-usual group.

H4. CC-DLF is used to reduce the rate of reoccurring error types.

H5. The mean score of CC-DLF feedback is $\geq 4/5$ and hallucinated-correction rate is $< 5\%$.

H6. There are no more than 5 F0.5 points between the subgroup performance gap (L1, district, proficiency).

3. Methodology

3.1 Design

This study is an explanatory sequential mixed methods design consisting of three phases. Phase 1 (Completed) is a cross sectional survey which sets a problem baseline. Phase 2 generates a learner-error corpus and trains/evaluates the deep-learning components. Phase 3 is an eight week, three arm quasi-experiment that includes a delayed post test and follow-up interviews. Phase 1 data are secondary analyses of the uploaded survey export; Phases 2–3 need to be new data collected using the instruments in Appendix A.

3.2 Context and participants

Phase 1 participants were 287 learners from Hamirpur (66.6%), Mandi (15.3%), Kangra (12.9%), Shimla (4.9%) and Una (0.3%), aged 18–26 (73.5% aged 18–20). Self-rated proficiency was Beginner (48.1%), Intermediate (46.0%) or Advanced (5.9%); reported L1 was Hindi (55.4%), English (39.0%) or a local dialect (5.6%); fields were STEM (43.9%),

professional roles (23.3%), competitive-exam preparation (19.5%) and engineering (13.2%); 64.8% reported using AI tools for English. Convenience sampling was used to obtain the sample. Only 39% of respondents reported English as their first (L1) language, which is not plausible in this region and is likely to be the “medium of instruction” and therefore treated with caution as L1. If randomisation is by class, we plan 210 (70 per arm) which equates to an expected attrition rate of around 20% for Phase 3, a priori power analysis on a one-covariate ANCOVA with 3 groups (Cohen's $f = .25$, $\alpha = .05$, power = .80).

3.3 Instruments

The original 26 question survey (five blocks: AI/global employability, core skills, global-cultural competence, experiential limits, professional/entrepreneurial competence) was re-mapped into six study constructs (Table 3) and supplemented by five new modules (E–J; 40 fields; Appendix A and the accompanying workbook). New modules include self-reported error profile (EP1- EP9), feedback experience (FB1- FB11), access and context (AC1-Ac8), writing and diagnostic tasks that produce the corpus (W1- W2, GT), and self-efficacy (SE1- SE3). The original items focus on employability and negotiation, and not so much on error feedback, so they are used here as a general baseline of AI perception, rather than a measure of error feedback quality.

Table 3. Re-mapping of the 26 existing items to study constructs

Code	Construct	Original items	Role
LS	Language-skill effectiveness	CS1, CS2, CS3	Perceived grammar/vocabulary, writing/listening, pronunciation gains
PF	Personalised feedback & error diagnosis	CS4, CS5, AI6	Core to the framework: personalisation and diagnosis of L2 input
LC	Linguistic-cultural sensitivity	GC1, GC4, GC5	Idioms, figurative language, learner linguistic preferences
CE	Communicative & employability transfer	AI1–AI5, GC2, GC3	Downstream transfer to interviews, negotiation, employability

Code	Construct	Original items	Role
EX	Experiential & creative limits	EL1–EL4	Perceived limits of AI in judgement, creativity
PE	Professional / entrepreneurial competence	PEC1–PEC6	Decision-making, leadership, simulations

Note. EL1 and PEC4 are worded as limitations but correlate positively with all other items ($r > .35$) and were not reverse-scored. Full mapping with question wording is in the workbook.

3.4 Corpus construction and annotation (Phase 2)

About 300 consenting learners on 4 prompts (2 pre test, post test and delayed post test) will give 20 minute timed texts (150-200 words) which will be added to the Himachal Pradesh Learner-Error Corpus (HP-LEC) yielding 1200 texts. Texts are anonymised, tokenised and double annotated at the span level in accordance with an ERRANT-compatible scheme (Bryant et al., 2017) and categorised at the study level, which includes 8 study categories – article, preposition, tense/aspect, subject–verb agreement, word order, spelling/punctuation, lexical/collocation, other – and a transfer flag indicating errors resulting from the Hindi/Pahari pattern (e.g., omission of the article, choice of preposition based on its postposition, SOV word order, etc.). Inter-annotator agreement is presented as Cohen's κ (Cohen, 1960) of .75 or greater, and resolution of disagreements. External pre-fine-tuned data is from BEA-2019 collection (Bryant et al., 2019) and, if licensed, LEAF (Behzad et al., 2024).

3.5 The CC-DLF architecture

Table 4. Layers of the Corpus-Centric Deep Learning Framework

Layer	Component	Specification	Output
1. Corpus	HP-LEC + external data	Annotated spans, error type, transfer flag, learner metadata (L1, proficiency, district)	Training/evaluation data

Layer	Component	Specification	Output
2. Detector	Multi-task transformer encoder	BERT-style encoder (Devlin et al., 2019) with GECToR-style edit tagging (Omelianchuk et al., 2020); heads for 8-way type and transfer flag; learner-profile embedding (L1, proficiency) concatenated to token states; loss $L = L_{tag} + \lambda_1 L_{type} + \lambda_2 L_{transfer}$	Edit tags, type, transfer probability per token
3. Learner model	Sequence model over error events	GRU/Transformer over each learner's chronological error-event vectors, analogous to deep knowledge tracing (Piech et al., 2015); predicts probability of recurrence per category	Mastery vector; top-2 focus areas
4. Feedback generator	Instruction-tuned LLM with retrieval	Inputs: sentence, detected edit, type, transfer flag, mastery vector, preferred explanation language, retrieved teacher-authored rule card. Output: what / why (with L1 contrast) / try-it micro-task. A verifier rejects any suggested edit that conflicts with the detector. Teacher dashboard allows approval	Structured, bilingual-ready feedback

The design solutions address the gaps directly. L1 operationalizes the L1-adaptation insight of Rozovskaya et al. (2017) in a neural model; feedback is cumulative in the learner model to prevent information loss; retrieval from teacher-authored rule cards and a verifier mitigates fabricated corrections which could arise with LLM feedback (Barrot, 2023; Pack et al., 2025); and restricting feedback to two focus areas per text prevents information overload. All the raw learner text is processed on institution-controlled infrastructure with open-weight models and no text is sent to third-party APIs without explicit consent.

3.6 Training and evaluation protocol

The models are pre-fined on external data and then adapted to HP-LEC through five-fold cross validation twice (split by learner to avoid leakage). Baselines are (B1) an L1-agnostic detector, (B2) a rule-based checker, and (B3) a prompted general LLM. Ablations eliminate L1 conditioning, the learner model and retrieval of the rules. Detection scores are based on span-level precision, recall and F0.5 (GEC convention; Bryant et al., 2023); classification scores are macro-F1; confidence intervals are 1,000 bootstrap resample; model comparisons are performed using paired permutation tests with Holm correction (Holm, 1979). Feedback quality is rated blind by two trained raters on the five criteria used by Steiss et al. (2024) — accuracy, clarity of directions, prioritisation, supportive tone and criteria alignment — and on a binary hallucination check. Subgroup metrics are provided for L1, district and proficiency.

3.7 Intervention (Phase 3)

Learners are randomly assigned to three arms: (1) CC-DLF feedback; (2) generic AI feedback through a generic, default prompt used by a general purpose chatbot; (3) business-as-usual teacher feedback. The 8-week programme consists of one writing task each week, which receives feedback and is revised in the week after, and a delayed post-test in week 12. Primary outcome: adjusted post-test grammar accuracy (GT_post, covariate GT_pre) and error rate per 100 words in W1_post. Secondary outcomes: error-recurrence rate, perceived personalised feedback (PF), self-efficacy (SE1, SE2, SE3), and affect (FB9). Reflexive thematic analysis (Braun & Clarke, 2006) is used for the analysis of semi-structured interviews with approximately 20–24 learners and 4–6 teachers.

3.8 Analysis plan

Table 5. Research questions, data and analytic methods

RQ	Data	Method	Decision criterion
RQ1	Survey (26 items + demographics)	Reliability (α , ω), parallel analysis/EFA, Mann–Whitney, Kruskal–Wallis, OLS with robust SEs, k-means, MLP vs logistic regression	$\alpha \geq .70$; α -level .05; effect sizes reported; Holm for post-hoc
RQ2	HP-LEC annotations	Multi-task transformer; ablations; bootstrap CIs; permutation tests	F0.5 and macro-F1 vs baselines; $\kappa \geq .75$
RQ3	Pre/post tests, error logs, PF, interviews	ANCOVA / linear mixed model (class random intercept); Poisson mixed model for recurrence; paired Wilcoxon; thematic analysis	Adjusted group difference, $f \geq .25$ targeted; 95% CI excluding 0
RQ4	Detector and learning outcomes by subgroup	Subgroup metrics, interaction terms	Gap ≤ 5 F0.5 points; no significant arm $\times$ subgroup interaction

3.9 Ethics and data protection

For data collection, written informed consent, participation in the study being voluntary with the option to withdraw at any time, and the removal of e-mail addresses and names (usernames in the Phase 1 export were deleted before analysis) is required for Phases 2–3. The storage and processing is compliant with Digital Personal Data Protection Act, 2023 (Government of India, 2023). Teachers have final control over feedback provided to students and a bias audit by subgroup is reported.

4. Analysis of Phase 1 data (executed)

4.1 Data screening and procedures

There were 287 full responses in the export, with no missing values. Five usernames were repeated and 16 users provided all ratings of the 26 items the same; these were retained and a sensitivity analysis eliminated straight-liners ($n = 271$). Likert data were treated as ordinal for group tests and as approximately interval for composite scores, which is acceptable since composite scores are summed scales (Norman, 2010). Cronbach's α (Cronbach, 1951) was used to determine the reliability, while McDonald's ω (McDonald, 1999) was used as well. The Kaiser–Meyer–Olkin index (Kaiser, 1974), Bartlett's test and Horn's parallel analysis (Horn, 1965) was used to examine the dimensionality followed by minimum-residual factor analysis. Group effects: Mann–Whitney U (rank-biserial r ; Kerby, 2014); Kruskal–Wallis (ϵ^2); profiles: k-means with silhouette width (Rousseeuw, 1987) Predictive modelling involved comparing the performance of an L2-regularized logistic regression with a two-hidden-layer MLP (32–16 units, early stopping) via a 5×5 repeated stratified cross-validation (AUC).

4.2 Descriptive findings

Perceptions were uniformly low: item means ranged 2.13–2.37 (SD 0.99–1.17); 63.3% of the 7,462 ratings were 1 or 2, 23.8% were 3, and 12.9% were 4 or 5 (Figure 1). The worst-rated areas were related to the effectiveness of grammar use (CS1, $M = 2.13$), the ability to provide professional skill insights (PEC1, 2.14), and the ability to identify the major issues in second-language input (CS5, 2.15): all of these are the diagnostic function targeted by the framework. The most highly rated (but still below the midpoint on the scale) included simulated negotiation practice (PEC6, 2.37) and idioms/figurative language (GC5, 2.34).

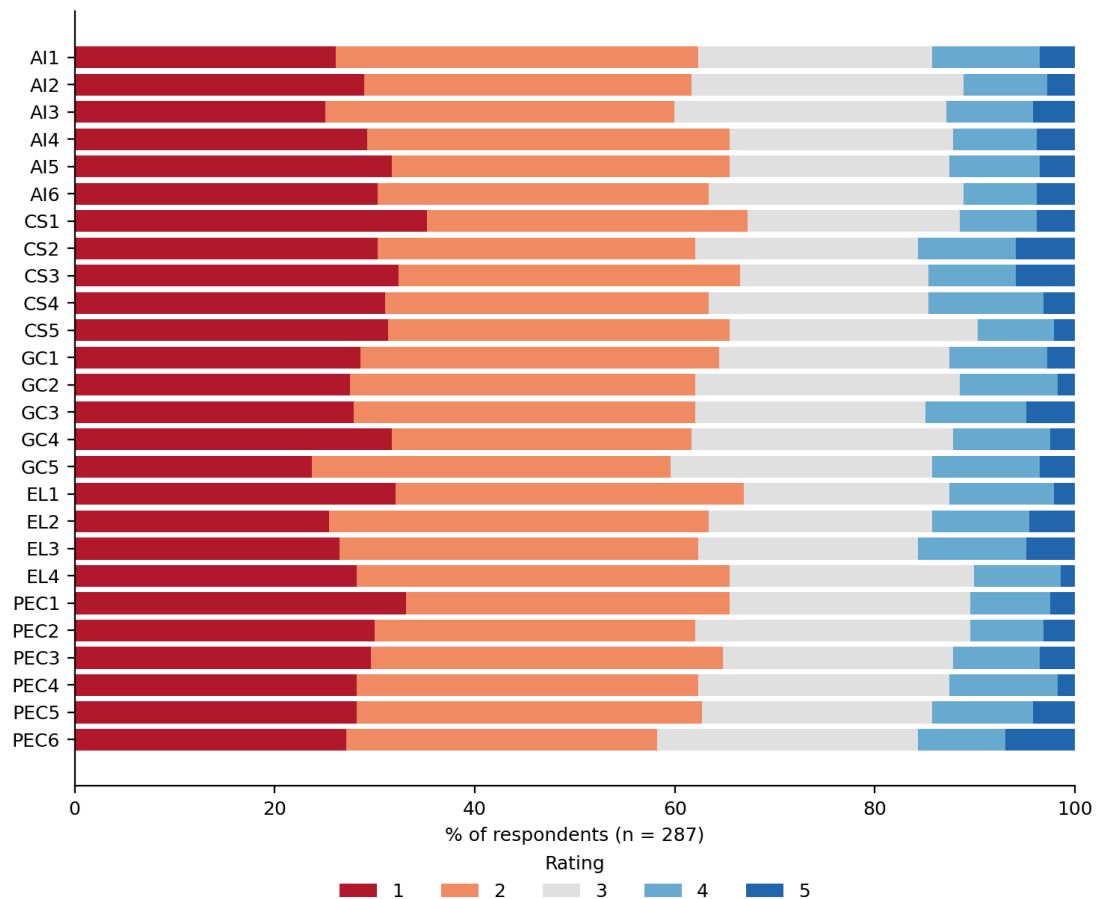


Figure 1. Distribution of responses to the 26 items ($n = 287$). Anchors were not stated in the source form; 1 = lowest, 5 = highest is assumed.

Table 6. Construct descriptives and reliability ($n = 287$)

Construct	Items	<i>M</i>	<i>SD</i>	α	ω
LS Language-skill effectiveness	3	2.21	0.92	.73	.74
PF Personalised feedback & diagnosis	3	2.20	0.84	.70	.70
LC Linguistic-cultural sensitivity	3	2.26	0.85	.71	.72
CE Communicative & employability transfer	7	2.25	0.81	.88	.88
EX Experiential & creative limits	4	2.24	0.82	.77	.77
PE Professional / entrepreneurial	6	2.24	0.80	.83	.84
All 26 items	26	2.24	0.72	.95	.95

Note. Original five-block alphas: AI = .86, CS = .82, GC = .83, EL = .77, PEC = .83. Construct alphas of .70–.77 for three-item scales are acceptable but modest.

4.3 Dimensionality

Data were suited to factor analysis ($KMO = .945$; Bartlett $\chi^2(325) = 4,019.6$, $p < .001$). The first eigenvalue (11.59) was greater than the parallel-analysis threshold (1.69), while the second (1.48) was not greater than the parallel-analysis threshold (1.56), which led to the acceptance of one dominant dimension that accounted for 42.4% of variance, with loadings ranging from .55–.74 (Figure 3a). The six a-priori constructs have correlations between themselves ranging from $\rho = .59$ to .81 and thus cannot be considered as empirically distinct constructs. Indeed, the respondents have a more global perception of “usefulness of AI” than individual perceptions regarding skills, feedback and employability (a specific measurement limitation that is addressed by the new, specific modules (E–F)).

4.4 Who perceives what? Group differences and predictors

Table 7. Group differences in the overall perception score (1–5)

Grouping	Test	Statistic	<i>P</i>	Effect size / direction
AI-tool use (yes vs no)	Mann–Whitney	$U = 11,212.5$	.007	$r_{rb} = .19$; users 2.16 < non-users 2.39
Field of study	Kruskal–Wallis	$H(3) = 20.20$	< .001	$\epsilon^2 = .06$; Engineering 1.93 < STEM 2.15 < Competitive 2.39 < Professional 2.45
Proficiency	Kruskal–Wallis	$H(2) = 3.03$	.220	$\epsilon^2 = .004$; Beginner 2.16, Intermediate 2.32, Advanced 2.28
L1	Kruskal–Wallis	$H(2) = 4.37$	.113	$\epsilon^2 = .008$; English 2.17, Hindi 2.27, local dialect 2.44

Grouping	Test	Statistic	<i>P</i>	Effect size / direction
District (Una excluded)	Kruskal–Wallis	$H(3) = 3.72$	.293	$\epsilon^2 = .003$
Age band	Kruskal–Wallis	$H(3) = 2.39$	.496	$\epsilon^2 \approx 0$

Note. After Holm correction, Engineering differed from Professional Roles ($p = .0006$) and Competitive-Exam learners ($p = .005$); STEM differed from Professional Roles ($p = .011$). Excluding straight-liners left conclusions unchanged (AI-use $p = .009$; proficiency $p = .255$).

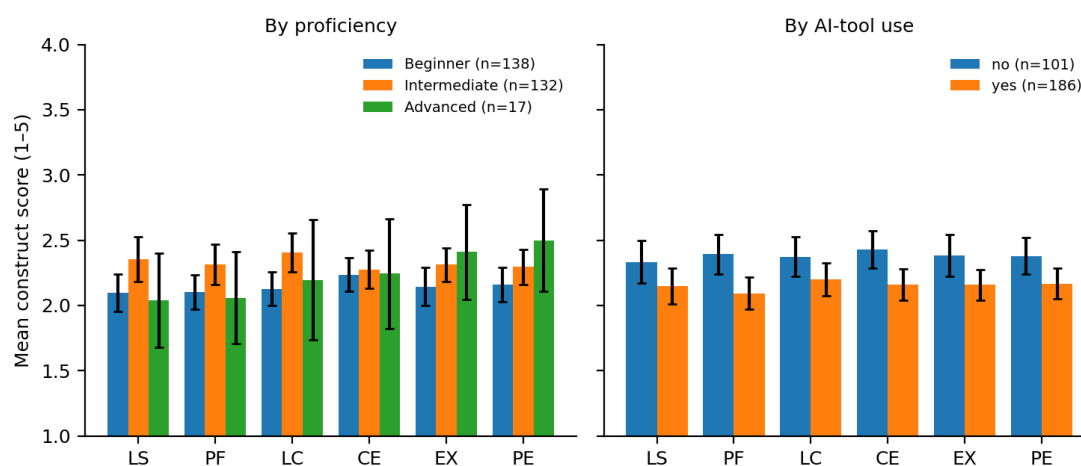


Figure 2. Mean construct scores by proficiency and AI-tool use (95% CI).

For the problem statement there are 2 results that are relevant. First, users of AI were less positive than non-users on all constructs (with the greatest difference on PF). A plausible explanation is expectation–experience gap: users of the tools have reached their limits while non-users give a hypothetical answer to the question; it will not be tested here. Second, this does not get better with practice: weaker and stronger learners are both equally unpersuaded and tools do not discriminate by needs. The use of AI was correlated with proficiency ($\chi^2(2) = 7.91$, $p = .019$, Cramér's $V = .17$): 58.0% of Beginners, 68.9% of Intermediate and 88.2% of Advanced learners reported using AI — the most proficient learners reported the least use. When using the OLS model (adj. $R^2 = .025$, F-test $p = .032$), AI use remained negative ($B = -0.25$, $SE = 0.09$, $p = .006$), and the 22–24 age band scored lower ($B = -0.40$, $p = .038$); proficiency, L1 and district were not significant.

Table 8. Regression of perceived language-skill effectiveness (LS) on other perceptions (H1)

Predictor	<i>B</i>	β	<i>p</i>
PF Personalised feedback & diagnosis	0.478	.436	< .001
LC Linguistic-cultural sensitivity	0.222	.203	.004
CE Communicative & employability transfer	0.006	.006	.942
EX Experiential & creative limits	0.088	.078	.293
PE Professional / entrepreneurial	0.098	.085	.296

Note. $R^2 = .52$. Predictors are intercorrelated ($p = .64-.81$), so coefficients should be read as indicative.

The most strong association of believing that AI can improve grammar, vocabulary, writing and pronunciation ($\beta = .44$) is with the perception of AI personalising feedback and diagnosing second-language input, followed by linguistic-cultural sensitivity ($\beta = .20$). This is correlational evidence, but it's the pattern the framework is meant to exploit.

4.5 Learner perception profiles and prediction

Two clusters (.43) were preferred for the silhouette width and a single cluster (.31) was kept for interpretability. The three profiles differ in level rather than shape (Figure 3b): low ($n = 123$; overall $M = 1.56$), moderate ($n = 108$; 2.49) and favourable ($n = 56$; 3.26). Profile membership was related to AI use ($\chi^2(2) = 7.19$, $p = .028$; users were 73.2% of the low profile versus 56.5% and 62.5% in the others) but not to proficiency ($\chi^2(4) = 7.57$, $p = .109$). Four in ten (42.9%) learners have uniformly low perceptions and only 19.5% are in the favourable profile.

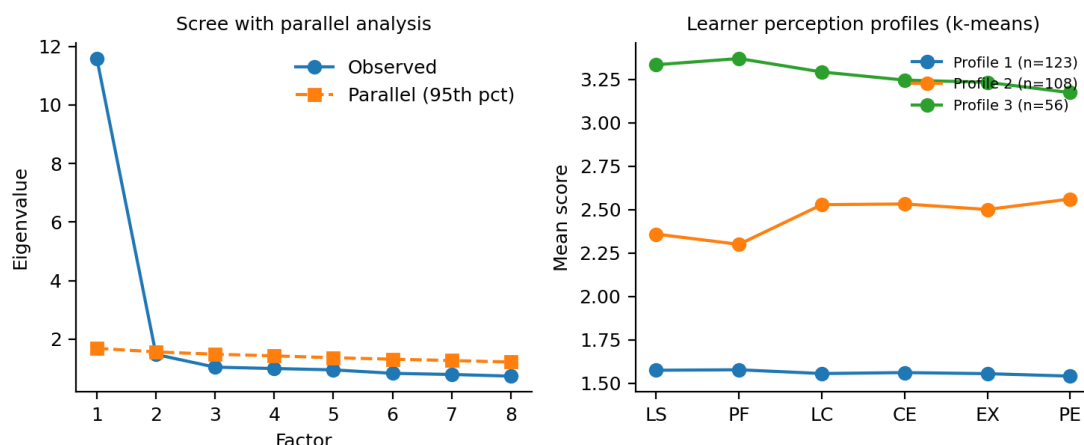


Figure 3. (a) Scree plot with parallel-analysis threshold; (b) profile means across the six constructs.

Table 9. Cross-validated prediction from the 26 survey items (AUC, mean of 5×5 folds)

Target	Logistic regression	MLP (32–16)	Majority class
Proficiency (Intermediate/Advanced vs Beginner)	.57 (SD .02)	.52 (SD < .01)	51.9% prevalence
Uses AI tools for English (yes vs no)	.56 (SD .01)	.51 (SD .01)	64.8% prevalence

Both models weren't meaningfully discriminative and the neural network did not perform better than the linear model. However, perception ratings provide very little information as to what a learner can or cannot do. This is a non-modelling finding, pointing to the need of training deep learning algorithms for personalised feedback on learner language (errors in context), not on attitude surveys.

4.6 Synthesis: the problem in non-native settings

- Learners rate the current AI support as weak (63.3% of all rating are ≤ 2), especially in diagnostics, which is required for personalised feedback.
- Dissatisfaction is global and is not tied to proficiency so current tools do not adapt to need, users are less satisfied than non-users.

- The lever most closely correlated with perceived effectiveness is perceived personalisation ($\beta = .44$).
- The ability to perceive does not correlate with proficiency (AUC $\sim .5$) and error level data are essential.

5. Desired outcomes and success criteria for Phases 2–3

The following is a list of pre-declared targets, not results. They translate the problem in Section 4 into measurable criteria which are not met and reported.

Table 10. Desired outcomes, metrics and targets

Outcome	Metric	Target	Link to Phase 1 problem
O1 Detection	Span-level F0.5 on held-out HP-LEC (learner-split CV)	$\geq .55$ and $\geq +3$ points over L1-agnostic baseline on transfer categories	No existing HP-specific model (G1)
O2 Classification	Macro-F1, 8 categories	$\geq .70$	Diagnosing “core issues” rated lowest (CS5)
O3 Annotation quality	Cohen’s κ	$\geq .75$	Corpus validity
O4 Feedback quality	Mean of 5 rubric criteria; hallucination rate	$\geq 4.0/5$, non-inferior to generic LLM; $\leq 5\%$ fabricated corrections	Trust and accuracy concerns (G2, G6)
O5 Learning	Adjusted post-test accuracy; error-recurrence rate	CC-DLF > generic AI > control; $f \geq .25$; $\geq 20\%$ lower recurrence than control	Feedback is not cumulative (G3)

Outcome	Metric	Target	Link to Phase 1 problem
O6 Perception	PF construct (baseline $M = 2.20$; $\approx 12\%$ of PF-item ratings are 4–5)	Post-intervention $M \geq 3.5$; $\geq 50\%$ rating 4–5	Low baseline perceptions (Section 4.2)
O7 Equity	Subgroup F0.5 gap (L1, district, proficiency)	≤ 5 points	Equity gap (G5)
O8 Practicality	Teacher review time; latency	≤ 2 min per text; < 5 s per sentence	Limited teacher capacity

6. Limitations

The instrument measured perceptions of employability and negotiation, rather than error feedback; 35.2% of respondents who don't use AI tools still rated AI, meaning some answers are hypothetical; the single dominant factor implies a halo effect and acquiescence effect; and the design is cross sectional, which means that the associations in Tables 7–8 are not causal. The AI-user result interpretation as expectation gap is a hypothesis. The deep-learning framework, corpus and intervention have not yet been implemented and the literature table was filled with verified bibliographic records and supplied references, venue indexing should be checked on Scopus prior to submission. An overview of a systematic search string is included in Appendix B.

Conclusion

The Phase 1 evidence indicates a definite need. LHS students rate the AI support they receive as weak, not differentiated by need, and least effective for diagnosis, and consider personalisation to be essential to be useful. In all, there were 63.3% rating 1 or 2 in 287 respondents. There was no difference in dissatisfaction as ability progresses, so the current tools do not seem to adjust to who is in need most. Users of AI are not as optimistic as those who do not use it, indicating that exposure to AI tools reveals its shortcomings. The highest correlation with perceived effectiveness was perceived personalisation ($\beta = .44$), signifying the places where improvement will have the greatest impact. Since the survey data does not address

what misconceptions learners may have (cross-validated AUC from .52 to .57), the next steps involve the construction of the L1-aware corpus, training the detector–learner-model–generator pipeline and testing it on the basis of pre-determined accuracy, learning, perception and equity criteria against generic AI feedback. The framework proposed is thus a testable solution to a documented problem, rather than a demonstrated solution.

Future implications

A framework might enable teachers who have classes that are too large to build error profiles themselves to remain in control by reviewing the system's suggestions, and up to the targets they pre-declare themselves. The explanations in the bilingual version might be extended to other Hindi and Pahari speaking parts and the annotated corpus might be made available as an open set of tools to audit the fairness of the English-learning tools for under-represented learners. The subgroup results would help policy-makers who are putting the National Education Policy 2020 into practice to determine whether instruction with AI is helping learners who are most in need or exacerbating disparities. Future research should be extended to the other areas learners found the least competent in (speech and pronunciation), follow learners over longer periods to test for retention, and explore whether accuracy improvements transfer onto the employability outcomes that motivated many of the participants in this study. Results obtained through different language backgrounds and state replicates with probability samples would indicate the extent to which results can be generalised.

References

- Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57, Article 100745. <https://doi.org/10.1016/j.asw.2023.100745>
- Behzad, S., Kashefi, O., & Somasundaran, S. (2024). LEAF: Language learners' English essays and feedback corpus. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-short.36>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>

- Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 793–805). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P17-1074>
- Bryant, C., Felice, M., Andersen, Ø. E., & Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 52–75). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-4406>
- Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., & Briscoe, T. (2023). Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3), 643–701. https://doi.org/10.1162/coli_a_00478
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching*, 5(4), 161–170. <https://doi.org/10.1515/iral.1967.5.1-4.161>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- 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* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20, Article 57. <https://doi.org/10.1186/s41239-023-00425-2>
- Forti, L. (2024). Proficiency-rated learner corpora. *International Journal of Learner Corpus Research*. <https://doi.org/10.1075/ijlcr.00045.for>

- Gilquin, G. (2024). Learner corpora to cater for learners' individual needs. *Babylonia Journal of Language Education*. <https://doi.org/10.55393/babylonia.v2i.386>
- Government of India. (2023). *The Digital Personal Data Protection Act, 2023* (Act No. 22 of 2023). Ministry of Law and Justice.
- Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7–24. <https://doi.org/10.1075/ijlcr.1.1.01gra>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185. <https://doi.org/10.1007/BF02289447>
- Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. *Educational Technology & Society*, 26(1), 112–131. [https://doi.org/10.30191/ETS.202301_26\(1\).0009](https://doi.org/10.30191/ETS.202301_26(1).0009)
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31–36. <https://doi.org/10.1007/BF02291575>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, 3, Article 11.IT.3.1. <https://doi.org/10.2466/11.IT.3.1>
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.
- Ministry of Education, Government of India. (2020). *National Education Policy 2020*.
- Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education*, 15(5), 625–632. <https://doi.org/10.1007/s10459-010-9222-y>

- Omelianchuk, K., Atrasevych, V., Chernodub, A., & Skurzhashnyi, O. (2020). GECToR – Grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 163–170). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.bea-1.16>
- Pack, A., Hartshorn, K. J., Escalante, J., & [additional author(s) — complete from publisher record]. (2025). How well can GenAI (GPT-4) provide written corrective feedback on English-language learners' writing? *International Journal of English for Academic Purposes*. <https://doi.org/10.3828/ijeap.2025.2>
- Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in Neural Information Processing Systems*, 28, 505–513.
- Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: Illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Frontiers in Psychology*, 14, Article 1255594. <https://doi.org/10.3389/fpsyg.2023.1255594>
- Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, 31(7), 653–674. <https://doi.org/10.1080/09588221.2018.1428994>
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Rozovskaya, A., Roth, D., & Sammons, M. (2017). Adapting to learner errors with minimal supervision. *Computational Linguistics*, 43(4), 723–760. https://doi.org/10.1162/COLI_a_00299
- Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1–4), 209–231. <https://doi.org/10.1515/iral.1972.10.1-4.209>
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, Article 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>

Wei, L. (2023). Artificial intelligence in language instruction: Impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology, 14*, Article 1261955. <https://doi.org/10.3389/fpsyg.2023.1261955>

Yang, N., Gao, N., & Zhang, T. (2025). Artificial intelligence in EFL writing enhancement. *International Journal of Web-Based Learning and Teaching Technologies*. <https://doi.org/10.4018/ijwltd.390442>

Appendix A. New questionnaire and task items added for the study

All items are rated 1–5 unless stated (frequency items: 1 = Never, 5 = Always; agreement items: 1 = Strongly disagree, 5 = Strongly agree). These fields are present but empty in the accompanying workbook; no responses have been simulated.

Code	Item wording	Module
EP1	How often do you omit or misuse articles (a / an / the) when writing English?	E
EP2	How often do you choose the wrong preposition (e.g., "discuss about", "depend from")?	E
EP3	How often do you make mistakes in verb tense or aspect?	E
EP4	How often do you make subject–verb agreement mistakes?	E
EP5	How often do you translate word-for-word from Hindi/Pahari, producing unnatural English or word order?	E
EP6	How often do spelling or punctuation mistakes appear in your writing?	E
EP7	How often do you use a wrong word or unnatural word combination (collocation)?	E
EP8	I can notice my own mistakes before anyone points them out.	E
EP9	The same mistakes keep returning even after they have been corrected.	E
FB1	AI feedback explains WHY something is wrong, not only what the correction is.	F

Code	Item wording	Module
FB2	I understand the language in which AI explains my mistakes.	F
FB3	AI feedback is matched to my level of English.	F
FB4	AI feedback addresses mistakes that are typical for Hindi/Pahari speakers.	F
FB5	I would learn more if explanations were given in Hindi or in both Hindi and English.	F
FB6	After receiving AI feedback I revise my text.	F
FB7	I accept AI corrections without checking them.	F
FB8	I prefer teacher feedback to AI feedback.	F
FB9	AI corrections make me anxious or discouraged.	F
FB10	I would use a tool that tracks my repeated mistakes and sets practice tasks for them.	F
FB11	I would trust such a tool if my teacher could review its suggestions.	F
AC1	Reliable internet access at home/college (1 Never – 5 Always).	G
AC2	Own smartphone/laptop available for study (0 = shared, 1 = own).	G
AC3	Hours per week spent using AI tools for English.	G
AC4	Tools used (tick all): ChatGPT / Gemini; Grammarly; Duolingo; Google Translate; Other; None.	G
AC5	Medium of school instruction (Hindi / English / Mixed).	G
AC6	Type of institution (Government / Private).	G
AC7	Years of formal English instruction.	G
AC8	Does a teacher guide your use of AI tools? (Never – Always)	G
W1_pre	Timed writing task, Prompt A, 20 min, 150–200 words (pre-test). Paste text.	H
W2_pre	Timed writing task, Prompt B, 20 min, 150–200 words (pre-test).	H

Code	Item wording	Module
W1_post	Post-test Prompt C (after 8-week intervention).	H
W2_post	Post-test Prompt D (delayed, week 12).	H
GT_pre	40-item grammar/usage diagnostic covering the 8 error types (score 0–40), pre-test.	H
GT_post	Same diagnostic, parallel form, post-test.	H
SE1	I am confident I can write a correct English paragraph without help.	I
SE2	I can fix my own grammar mistakes when I know where they are.	I
SE3	I feel anxious when I must write in English.	I
ARM	Assigned arm: 1 = Framework feedback; 2 = Generic AI feedback; 3 = Teacher/business-as-usual (set by researcher).	J
RATER_A	Error-annotation by rater A (file ID).	J
RATER_B	Error-annotation by rater B (file ID).	J

Appendix B. Systematic Scopus search string (to confirm the gap)

TITLE-ABS-KEY (("learner error" OR "grammatical error correction" OR "error analysis") AND ("deep learning" OR transformer OR "large language model" OR "neural") AND (personali* W/3 feedback) AND (English OR EFL OR ESL OR "second language")) AND PUBYEAR > 2019 AND (LIMIT-TO (DOCTYPE , "ar"))

Add a second line restricted with AND (India OR Hindi OR "Himachal") to test G1. If this returns no study that combines an L1-annotated learner corpus, deep-learning error classification, learner modelling and personalised feedback in an Indian setting, the gap claims in Table 2 are confirmed; if it returns such studies, they must be added to Table 1 and