TPM Vol. 32, No. 3, 2025
ISSN: 1972-6325
https://www.tpmap.org/

A HYBRID ANALYTICAL FRAMEWORK FOR EVALUATING AI-GENERATED FEEDBACK IN ENGLISH AS SECOND LANGUAGE WRITING

DR. AYESHA BIBI
ASSISTANT PROFESSOR, DEPARTMENT OF HIGHER EDUCATION COLLEGES, AJK, PAKISTAN
EMAIL: mphillinguistics4@gmail.com

DR HUMERA FARAZ
ASSISTANT PROFESSOR DEPARTMENT OF ENGLISH AIR UNIVERSITY ISLAMABAD
humera.faraz@au.edu.pk

DR. SAJJAD AHMAD
ASSISTANT PROFESSOR, DEPARTMENT OF ENGLISH, BACHA KHAN UNIVERSITY CHARSADDA,
EMAIL: sajjadahmad.eng@bkuc.edu.pk

ABSTRACT
The recent launch and escalation of generative Artificial Intelligence (AI) like the one called ChatGPT have revolutionized English as a Second Language (ESL) writing instruction to create automated feedback systems that offer linguistic and structural assistance at a moment's notice. Prior studies are still fragmented in that they tend to examine individual components like grammatical error, perceptions and/or usability aspects of feedback to writing with AI in isolation rather than providing an integrated evaluative approach. This conceptualization has shaped a gap in the methodology for comprehending the phenomenon of AI- feedback action in the several aspects of the quality of writing and pedagogy. To tackle this issue, the aim of this study is to design a multidimensional Hybrid AI Feedback Evaluation Model (HAFEM) for systematically assessing the AI feedback for ESL writing. Maintaining the linguistic accuracy, discourse quality, pedagogical value, AI reliability, and ethical integrity are five interconnected layers identified in the model reflecting Error Analysis Theory (EAT), Corpus Linguistics (Corkus), Discourse Theory, Sociocultural Learning Theory (SCT), natural language processing evaluation approaches (NLPEA), and ethical frameworks in AI (AI Ethics). The study states that considering it just a single dimension is not enough to understand the complexity of feedbacks that are produced by AI in education. According to implications, the proposed framework is of real benefit to ESL teachers, curriculum designers, practitioners in the field of artificial intelligence (AI) in language education, and policymakers by providing a structured tool to assess the efficacy, linguistic quality and ethical considerations of AI-enhanced writing systems. It also helps promote the proper and responsible use of AI in the language learning process to be more transparent. Generally, the model has implications for computational linguistics and applied linguistics, through its emphasis on interdisciplinary evaluation criteria in the context of AI-generated language feedback. The study further highlights the needs to evaluate AI-generated feedback for ESL writing from a more comprehensive perspective, beyond the fragmented one, and from a more theoretical one. By incorporating linguistic, discourse, pedagogical, technical, and ethical aspects in a single analytical framework, the HAFEM model overcomes the lack of a comprehensive modeling approach and paves the way for future interdisciplinary studies on the use of AI in L2 learning methodologies.
KEYWORDS: Artificial intelligence, ESL writing, ChatGPT, automated writing feedback, discourse analysis, error analysis, corpus linguistics, AI ethics, pedagogical feedback, Hybrid AI Feedback Evaluation Model (HAFEM)

INTRODUCTION

With the exceptional progress of AI, its impact on educational practices, especially on second language writing education, is getting more and more profound (Yang, Gao, & Shen 2024). The proliferation of generative AI systems like ChatGPT has spurred LAI to be more rapidly adopted in the English as a Second Language (ESL) and English as a Foreign Language (EFL) writing context (Yan & Zhang, 2024). The use of AI-based writing systems for automatic

written corrective feedback (AWCF), grammar correction, word or phrases improvement, coherence tips and support system revision is becoming more common for language learners (Escalante et al., 2023). As a result, the area of machine-generated feedback, which is part of artificial intelligence, has emerged as one of the fastest growing branches of applied linguistics, educational technology and computational linguistics (Mahapatra, 2024).

A recent study revealed the positive impact of comprehensively providing writing-focused feedback through AI on the grammatical accuracy of ESL learners, their fluency in writing, thesaurus use, and the level of learner engagement (Mahapatra, 2024). Another similarly potential application is revising the assessment process and practices of writing with the assistance of ChatGPT-based feedback systems in EFL classes (Li et al., 2024). On the other hand, AI tools for writing offer tailored, personalized, and rapid feedback, which might be difficult for teacher-led models to offer due to constraints on time and workload (Yang et al., 2024). The changes highlight a transformative approach in writing teaching from traditional corrective feedback to AI-based writing support systems. The changes signal a significant paradigm shift in pedagogical approach to writing from a conventional corrective feedback to AI-supported writing assistance systems.

Practical research in the context of learning English as a second language since the emergence of the AI feedback system have been widely underway, but the findings so far were theory and methodology, which were rough and irregular (Yan & Zhang, 2024). Recent studies tend to examine specific aspects of AI-driven feedback, including scores automated by the system or the ease of use for students, or student perceptions of quality but leave out more significant aspects including the pedagogical considerations in using the feedback, the discourse in the feedback, and ethical issues in feedback (Escalante et al., 2023). While there has been ample literature that focuses on the level of surface language correction, there has been less literature that focuses on the level of coherence in the discourse, learner scaffolding, the reliability of the feedback, and ethical issues related to the generative AI systems that provide feedback (Yoon et al., 2023). Therefore, a comprehensive analytical framework to systematically consider feedback from AI on multiple interconnected dimensions does not exist currently.

In addition, the sophistication of the features of generative AI has contributed to mounting risks for plagiarism, attribution issues, algorithmic bias, “hallucination,” overreliance on AI tools, and ethical misuse in academic settings (Sison et al., 2023). There has been increased focus on the merits of generative AI systems under a "whole human" (ethical and pedagogical) lens beyond just efficiency considerations (Khowaja et al., 2023). While AI tools are now coming into the mainstream of classroom use, essential characteristics like fairness, transparency, trustworthiness, and learner autonomy are features not always considered by evaluators (Sison et al., 2023). Therefore, a blended and interdisciplinary analysis of applied linguistics, corpus linguistics, discourse analysis, educational technology and AI ethics is required to assess AI-generated feedback. Furthermore, research literature studies conducted in recent years provide evidence of conflicts between the quality and usefulness of ChatGPT-written comments in the field of ESL writing. For instance, Yoon et al. (2023) have found the feedback on coherence and cohesion necessary when using ChatGPT for ELL writing was often too general and lacked enough specific revision criteria. Likewise, Yang et al. (2024) found that while the automated writing evaluation system developed by AI can offer prompt and accurate feedback, students often have difficulties in understanding and applying the automated suggestions when revising their writing. The findings establish the need to abandon the one-dimensional evaluation approach of the grammar favoring, to introduce the multidimensional approach that allows the evaluation of the linguistic, pedagogical, discourse, but also the ethical level of the language.

However, the lack of a common model has led to methodological differences in studying AI writing assistance in ESL contexts. This has resulted in methodological inconsistencies in the study of AI writing assistance for ESL contexts. The existing analytical approaches show significant differences regarding the criteria they use in assessing an AI system, which makes it difficult to compare them (Escalante et al., 2023). Certain research works focus on precision and understanding of the code while others increase retention and satisfaction from the users or technological effectiveness without any set analytical dimensions (Yan & Zhang, 2024). That methodological integration is lacking impedes theory building in relation to AI supported writing research and hampers future comparative research. In this sense, the growing demand from journals for innovations in methodology and adequate theoretically sound models for thoroughly assessing generative AI technology in educational settings is manifesting.

To address these theoretical and methodological deficiencies, the present paper suggests using the Hybrid AI-Fbk Eval (HAFEM): an interdisciplinary analytical model for evaluating AI-generated feedback to ESL writing systematically. In this article, I attempt to pragmaticize the vision of developing a conceptual and methodological construct for the use of future researchers in comparative studies of AI-assisted writing feedback. The framework proposed seems to be a combination of five major traditions which are Error Analysis, Corpus Linguistics, Discourse Analysis, Pedagogical Evaluation, and AI Ethics.

According to the HAFEM framework, the evaluation of generated textual feedback by AI can be conceptualized from 5 interconnected aspects where the linguistic accuracy of the content is measured, its discourse evaluated, its pedagogical value assessed, its reliability compared with the reliability of humans, and its ethical responsibility considered. The areas of linguistic accuracy consist of grammar and syntax corrections, vocabulary corrections, coherence, cohesion, organizational structure, scaffolding, clarity, learner support, AI hallucination and AI consistency, and ethical responsibility, namely: bias, dependency, transparency, authorship concerns. The model

proposed in this paper combines these dimensions into one comprehensive analysis model, striving to address the reductionism of previous studies on the evaluation of feedback in AI.

The main value of this article is that it brings together 'instabilities', 'social circuits' and 'family strategies' into a shared interdisciplinary discourse for future research. The conceptual guidance offered by the proposed model in this research can be used as a reference for providing feedback in German Level Writing by students on ELW and can guide future researchers systematically in analyzing German Level Writing feedback in ESL writing situations, where they need to analyze and evaluate the feedback by considering the linguistic aspects proposed by the model. With the rapid transformation in the global landscape of language teaching due to the rise of generative AI, it is even more important than ever to create evaluation models based on theory and principles to ensure proper use of technology in ESL instruction. Given the way language teaching is evolving globally with the advent of generative AI, there is a growing need to construct evaluation models that are grounded in theory and ethics to ensure responsible application of AI in ESL teaching.

Problem Statement

The advent of the generative AI (gAI) tools, including ChatGPT, into English as a Second Language (ESL) writing instruction has revolutionized the nature of automated written corrective feedback (AWCF) and digital writing support systems (Escalante et al., 2023; Yan & Zhang, 2024). It is noteworthy that the use of AI technologies for feedback has become commonplace in both ESL and EFL teaching and learning in education, as they can deliver personalized, immediate, and scalable support for grammar, vocabulary, coherence, and revision, thus contributing to students' writing development (Mahapatra, 2024). Recent studies also suggest that AI feedback can enhance writing fluency and revision ability, along with promoting the engagement of learners in foreign language writing contexts (Yang et al., 2024). Despite the increasing adoption of AI-based structures for receiving feedback in language education, the concept of language teaching and learning is still not very unified and consistent in the method in this field.

Generally, previous studies of AI-powered feedback generation for ESL writing have concentrated on specific characteristics like grammatical correctness, learner perception, technological user-friendliness, or automated scoring efficiency (Escalante et al., 2023; Banihashem et al., 2024). While many studies assess an AI system either at a narrow linguistic/technical level or independently from how broader discourse, pedagogical, and ethical aspects of feedback assessment pertain to students' teaching (Warschauer et al., 2023). As such, existing methods and tools are not necessarily integrated and do not have a structured interdisciplinary design that can systematically account for the linguistic accuracy, discourse quality, pedagogical value, AI reliability, and ethical responsibility of AI feedback in the context of ESL writing.

Furthermore, the advent of generative AI tools in language learning has raised various issues about illusion with feedback, algorithmic bias, overdependence on AI, ambiguous author claims, transparency, and academic integrity (Sison et al., 2023; Perkins et al., 2023). The study of AI for evaluation has also evolved over the last few years, with the various modern studies highlighting that current methods focus mainly on technical aspects and do not consider the ethical and human aspects of AI based learning in a comprehensive way (Khowaja et al., 2023). AI-driven assessment tools have also been shown to generate biased results for non-native writers of English, which also poses issues for fairness and inclusion in ESL education (Liang et al., 2023). However, even with these concerns, there is still limited theory that could be offered to future researchers to assess AI-generated feedback using the multidimensional and interdisciplinary analytical models.

Moreover, in recent literature, methodological criteria for comparing AI-generated feedback to other languages and educational contexts are still missing (Yan & Zhang, 2024). As different analytical approaches proposed to evaluate different analytical procedures in AI-assisted ESL writing research (Escalante et al., 2023), it is hard to systematically compare them from existing methods. This has been identified as a new imperative for methodological innovation and theorized interdisciplinary frameworks that can assess AI feedback, rather than grammar-based or perception-based approaches, and to this end, scholars have called for a shift toward more interdisciplinary methods that build on multiple elements of the learner's experience (Jeon, 2025).

Thus it is necessary to develop a comprehensive analytical framework that can take multiple evaluative aspects of the AI generated feedback into consideration for ESL writing. To fill this gap, the present study advocates for the proposed conceptual and methodological Hybrid AI Feedback Evaluation Model (HAFEM), which integrates Error Analysis, Corpus Linguistics, Discourse Analysis, Pedagogical Evaluation and AI Ethics to pave the way for ongoing research and comparative analysis of AI-generated feedback in future ESL writing contexts.

2. LITERATURE REVIEW

A. AI in ESL Writing

In recent years, the incorporation of AI in ESL writing teaching has grown considerably, especially after the invention of generative AI, including ChatGPT and automated writing evaluation systems (AWE). In ESL and EFL education, AI writing tools have become more prevalent recently for correcting grammar, enriching vocabulary, offering organizational insights, and offering real-time suggestions during writing (Escalante et al., 2023). The rise of AI

powered writing systems is an indicator of the wider changes in digital education and language instruction, as language learning becomes increasingly automated, individualized, and informable (Yang et al., 2024).

At the core of such advancements lies ChatGPT, OpenAI's groundbreaking conversational AI model, which has spurred a wave of innovation. Among the most revolutionary developments is the creation of ChatGPT by OpenAI, a conversational AI model that has driven innovation. ChatGPT is capable of generating context relevant responses, clarifying points, rewriting paragraphs, and emulating interactive teaching scenarios (Warschauer et al., 2023), which are different from previous grammar-checking software that had been targeted mostly for surface-level correction. According to recent studies, ChatGPT can assist ESL students in enhancing their grammatical appropriateness, words choice and fluency of writing (Mahapatra, 2024). Likewise, Yan and Zhang (2024), found that the ESL learners interacted with AWCF produced by ChatGPT while revising their sentence structures and their lexical selections. The results point to an emerging role of AI systems as language learning assistive systems that are more interactive with the learner and go beyond the role of editing.

Grammarly, another popular AI writing assistant in the ESL domain, is another crucial piece of software. Grammarly is another significant piece of software that uses AI to support writing in an ESL context. Grammarly offers instant feedback on grammatical, punctuation, clarity, tone, and sentence structure. Grammarly has been shown to have a positive effect on students' grammatical ability and revision strategies when they write, as it gives them instant corrective feedback during the process (O'Neill & Russell, 2019). Other researchers also contend that Grammarly is more concerned with the surface level of the accuracy of language and lacks the ability to focus on more important concerns such as discourse-level coherence, rhetorical organization, and development of critical thinking skills (Ranalli, 2021). That means Grammarly is a helpful tool for helping writers be accurate, but its role in the pedagogical process is still subject to debate in the field of applied linguistics.

There has also been growing interest in the general domain of automated writing evaluation (AWE) in the field of ESL writing. The tools available for prompt and scoring learner writing could be characterized as AWE systems (e.g., Criterion, Pigai, Grammarly, and tools based on ChatGPT) (Stevenson & Phakiti, 2019). These systems rely on natural language processing (NLP), machine learning, and corpus-based algorithms to identify mistakes and make recommendations for writing. Some researchers agree that AWE systems can lessen teaching workload in some instances and offer opportunities for students to receive instant feedback that is not possible in the conventional classroom setting (Link et al., 2022). Besides, AI-powered feedback systems promote student independence in writing practice and revision, allowing students to continuously refine their writing, and to correct it multiple times (Yang et al., 2024).

While AI feedback offers some benefits, there are issues with the pedagogical reliability and consistency of AI generated feedback. Research suggests that sometimes, an AI system might make the wrong correction, offer wrong explanations, or provide vague revision suggestions that cannot take into account students' writing context (Yoon et al., 2023). Moreover, with the advent of recent research, the quality of presentation exhibited by the feedback generated by AI systems has been an issue that has been given special attention as most of the discussion has focused on grammatical correctness but not discourse, cohesion, or cognition of learners (Warschauer et al., 2023). Thus, existing research on AI writing is scattered both across the linguistic and the technological and the pedagogical fields.

Another major concern is methodological inconsistencies among the AI-assisted ESL writing studies. Previous studies typically assess AI tools from a single lens, focusing on student satisfaction, grammatical correctness, or technology efficiency, but not considering other broader analytical aspects (Escalante et al., 2023). With the constant development of generative AI technologies, researchers have pointed out the need for overall frameworks and approaches that are comprehensive and interdisciplinary in scope to evaluate AI-generated feedback with a systematic approach, taking into account aspects such as linguistic, pedagogical, discourse and ethical considerations (Kohnke el al., 2023). As can be seen from the existing literature explored, the transformative potential of AI in the ESL writing education and the need for theoretically-oriented assessment approaches for further investigation are in urgent need.

B. Existing Feedback Models

As an important aspect of second language writing teaching, feedback has long taken an important place because it helps students discover their own mistakes in language, enhance the rhetorical structure of composition and build second language academic writing ability (Hyland & Hyland, 2006). In an ESL setting, feedback not only serves as a corrective tool, but also plays a role as a pedagogical instrument influencing the cognition, revision and language development of the ESL learner (Ferris, 2014). In the last few decades several theories and models regarding learners' processing and use of written corrective feedback (WCF) have been proposed. The principles of these models still shape ongoing debates related to the use of AI to provide feedback on writing and automated grading systems today.

Teacher feedback theory assumes a fundamental strategy for feedback in ESL writing, placing the teachers at the center of analysis and offering individualized and context-specific feedback. Hyland and Hyland (2006) maintain that teacher feedback is socially situated and dialogic, that is, feedback is a process involving a dialog between the teacher and the learner and not just a collection of corrections. Teacher feedback may make use of direct correction, metalinguistic explanation, revision suggestions and motivational support, which enables teachers to make adjustments based on the learner's level of proficiency and rhetorical setting (Ferris 2014). Teacher feedback has

consistently been demonstrated to be helpful to revision quality and long-term improvement of writing, as it represents a combination of pedagogical and linguistic elements (Bitchener & Ferris, 2012). Feedback from teachers is also tainted with some constraints: Quick return, inconsistency, teacher busywork and inability to support large populations of students (Lee, 2017).

The other key aspect of feedback is related to written corrective feedback (WCF), which involves answering small-level grammatical, lexical and syntactic mistakes in written tasks (Ferris, 2014). Corrective feedback has been subject to extensive discussion in L2 acquisition studies. However, Truscott (1996) has put forward the controversial argument that grammar correction can be potentially counterproductive, or even ineffective since learners are not always able to internalize corrections. However, in contrast to this, Ferris (1999) made forceful support of giving corrective feedback. He showed that adequate grammatical feedback could make learner error more accurate over time to some extent. Later research confirms this perspective, that if the feedback given is clear, focused and meaningful in the context of the task, it is powerful (Bitchener & Ferris, 2012). Researchers differentiate between direct corrective feedback, in which the errors are directly corrected, and indirect corrective feedback, in which the learner is asked to spot and correct his/her own errors (Ellis, 2009). These differences are of great significance to research on AI-generated feedback, as automated feedback may frequently be done by a direct approach that cannot fully support domain leaner reflection or metacognitive activity.

Sociocultural and process approaches are also key features highlighted in contemporary scholarship on feedback. A sociocultural perspective states that feedback is a sort of scaffolding that helps scaffold learners' learning in the zone of proximal development (Vygotsky, 1978). In this model, a successful feedback needs to cultivate autonomy, interaction and self-regulated learning in their learner rather than simply rectify grammatical mistakes (Hyland & Hyland 2006). Likewise, process-writing approaches regard feedback as a continuous part of an ongoing revision process in which students draft, revise, and perfect their writing (Lee, 2017). These viewpoints demonstrate the need not just to assess the accuracy of feedback, but also its pedagogical usefulness, engagement of learners, and developmental effects.

The advent of automated feedback systems has brought a significant change to conventional modalities in ESL writing. Automated writing evaluation systems (AWE) like Criterion, Grammarly, Pigai and ChatGPT-based feedback tools would give instant feedback on learner writing through the technologies such as AI, machine learning, and NLP (Link et al., 2022). Some proponents of automated feedback systems believe that these deliver several benefits such as quick turnaround time, scalability, consistency, and independent revision (Stevenson & Phakiti, 2019). Such systems are especially beneficial in a lot of mass education where teachers may not be able to give detailed personalized feedback. While benefits abound, previous studies have identified several problems with automated feedback systems. However, Ranalli (2021) suggests that it can be challenging for learners to accept or understand the suggestions that the AI system makes, because the suggestions can be very generic or not engage with the context of the writing. Likewise, Warschauer et al. (2023) state that automated systems tend to focus on the surface transfer and linguistic mistakes made at the grammatical level, which they believe lacks coherence, rhetorical appeal and growth in critical thinking. Research literature more recently also indicates that AI based feedback system might facilitate passive revision practices by motivating learners to apply the revision outputs passively (Yan & Zhang, 2024). As a result, automatic feedback models developed so far are mostly severely constrained in the emphasis they place on linguistic correctness and neglect pedagogical and discourse level objectives.

The literature shows that, as a whole, feedback models in ESL writing are shifting from teacher-centered approaches of correction to technologically mediated and AI-supported approaches. Current methods, however, are disjointed both linguistically and pedagogically and technologically. Teacher feedback theories are designed around dialogic interaction or scaffolding while automated feedback systems are heavily geared towards efficiency and correctness of the feedback. The fragmentation underscores the importance of inter-sciplinary approaches that can incorporate the linguistic, discourse, pedagogical, and ethical aspects into a holistic framework for analyzing AI-generated feedback in ESL writing.

C. Corpus and Error Analysis

Corpus linguistics and Error Analysis Theory have contributed significantly to the field of serious language writing in terms of knowledge of learner errors, language patterns and ESL writing development. The research and writing assistance tools that integrate AI systems often argue errors and provide the writers with suggestions with algorithms based on corpora, so these methods are all the more pertinent in the field of writing research in the modern era. Corpus and error analysis theories offer valuable theoretical frameworks for analyzing the linguistic quality and reliability of the AI-generated feedback tools like ChatGPT, Grammarly, and other automated writing evaluations (AWE).

Stephen Pit Corder's groundbreaking work has influenced thinking about second language errors and can be seen as forming the roots of Error Analysis Theory. Before Corder's work, the errors of the learner were considered negative, that is, as failures and as interference of the first language. In response to this, Corder (1967) maintains that errors made by learners are regular and are steps in the language learning process. In Error Analysis Theory errors investigated and analyzed can gain insights into the inter-language development of learners and their cognitive processing while acquiring SCL. Corder pointed out that error analysis helps the researcher to discover patterns in

language development, learning strategies adopted, and the needed areas of language facilitation for instruction. The theoretical terms of this shift were very influential in applied linguistics and led to the development of error analysis as a research strand in studies of ESL writing.

The prominent aim of Error Analysis Theory is exploring the L2 errors that learners make in written or spoken form of the language in terms of error identification, error classification, error description, and error explanation (Ellis & Barkhuizen, 2005). Errors are generally classified as either grammatical errors, lexical errors, syntactic errors, morphological errors, or discourse errors, (James, 2013). In ESL writing situations, error analysis has been extensively applied in analyzing students' challenges working with tense, subject-verb agreement, and use of articles, prepositions, sentence construction and choice of words (Ferris, 2011). It is suggested that a systematic error analysis will assist the teacher to plan teaching "corrective/study" strategies and/or corrective uses of interventions which are specifically designed to satisfy the linguistic needs of learners. Furthermore, error analysis is still an important language analysis tool in AI-assisted writing, as many automated tools focus on detecting and correcting deviations in learner language in their main function.

Another methodological approach that has gained considerable impact in recent years in second language writing studies is corpus linguistics. In Corpus Linguistics, patterns, collocations and grammatical structures of language are identified from the analysis of very large collections of authentic language data called corpora (McEnery & Hardie, 2012). One major reason that there were many contributions to the understanding of learner language through corpus-based approaches is because the researcher is able to study language patterns of use across a large corpus that are not found in other approaches, which might just be based on intuition or a single example (Biber et al., 2021). Recent research on ESL writing has been conducted using learner corpora like the International Corpus of Learner English (ICLE) and the British Academic Written English Corpus (BAWE).

It is developing lately into a fundamental asset of automated writing evaluation systems, which use a large number of linguistic corpora as basis for training statistic algorithms. Grammatical errors are corrected or word suggestions or stylistic changes are made using, for example, the large-scale language models learned by writing tools like Grammarly or ChatGPT and/or probabilities established from a text database (Warschauer et al., 2023). There has been debate about corpus-based AI systems' ability to detect language patterns and generate context-sensitive suggestions that are better than rule-based grammar checkers (Biber et al., 2021). As a result, there are technological and theoretical implications for the use of foundational information and concepts of corpus linguistics in many of the current Artificial Intelligence-based feedback systems.

Even with all these improvements, researchers have spotted some drawbacks to the corpus-based AI feedback systems. A significant difficulty is that it is often hard to ensure that the automated system focuses on a statistically likely chance of language over context and pedagogic relevance (Ranalli, 2021). AI-powered corrections might look correct at the level of grammar even though they neglect to take into account the communicative purpose, genre norms or discourse-level coherence of the learner's production (Yoon et al., 2023:168). Likewise, auto-generated by AI, chatbots can reinforce dominant linguistic patterns and may under-represent non-native varieties, leading to the creation of bias against ESL writers (Liang et al., 2023). These worries stress the need for quality assessment "in addition to a corpus-based statistical approach.

Moreover, the traditional error analysis theory has been questioned for its overemphasis on grammatical errors at the surface level and its lack of attention to the organization of the discourse level, rhetorical effect and the nature of the learner's cognition (James, 2013). Current researchers have argued that teaching of writing needs to shift away from focusing on simply checking grammar toward complex and analytical evaluation that takes into account the learner's language, text, subject matter, and text production strategies, as well as ethical issues (Ferris, 2011). This criticism is particularly pertinent in the context of AI-provided feedback research, which is dominant in most cases, and is still largely focused on sentence-level corrections.

The capacity to experience discourse and cohesion.

Since the grammaticality of writing in L2 is only a single aspect of writing quality, in addition to the grammatical problems, second language writing research has been given an important appreciation in formulating questions that are addressed in discourse analysis and L2 cohesion studies. Learners in writing contexts in English as a Second Language should be given not only grammatically sound sentences but also cohesive, coherent and rhetorically competent texts (Hyland, 2019). As such, discourse-level assessment in writing evaluation and feedback has grown in significance for both conventional writing assessment and present-day studies of feedback in AI-generated writing. Discourse quality is an integral aspect of writing competence and not only sentence-level errors can be analyzed in grammar in view of evaluating students' writing competence in L2 writing (Warschauer et al, 2023).

Cohesion in English (1976) by Michael Halliday and Ruqaiya Hasan is one of the pioneering works on discourse analysis and cohesion. Halliday proposed that cohesion is the linguistic devices used to link words, phrases, clauses and sentences to ensure (linguistic) unity and continuity in a text and that these devices are conceptualized as mechanisms. Conceptualized by Halliday and Hasan as the mechanisms of linguistic devices linking words, phrases, clauses and sentences to ensure unity and continuity in a text. The concept of cohesive ties, according to their framework, aids in readers' interpretation of relationships between elements of texts and grasping how the meaning

unfolds across discourse. They brought up a number of different categories of cohesion, and these were (Halliday and Hasan, 1976): reference, substitution, ellipsis, conjunction and lexical cohesion. The ideas get focal importance in the discourse analysis and also shape of the writing assessment in applied linguistics and language education.

The Halliday-Hasan theory of cohesion, one of the greatest theories in the field of writing, proved to be a radical change in the conception of the quality of writing, and it became evident that the process of good writing is related to the process of meaningful connections within the text and is not related to the process of grammatical correctness. Cohesion has to do with the flow which information demonstrates and its readability, and coherence has to do with the overall, logical, meaningful organization of the ideas in discourse (McNamara et al., 2010). In ESL writing, learners sometimes lack coherence and cohesion as their written sentences are grammatically correct, but fail to have logical development and organization between the sentences (Crossley & McNamara, 2011). As a result, the concepts of coherence and cohesion have been recognized as significant criteria to measure writing skills in the assessment of second language learners.

Coherence and cohesion studies during the past 20 years have grown significantly, especially with the use of corpora and computers. While they show that using effective and specific "cohesion" devices including transition markers, lexical repetition, referential awareness, and thematic structure makes written work of high quality and helps readers understand what is being read (Crossley et al., 2016), they are very useful to include if they are present in the written text. Also related to coherence is cognitive processing as reading and constructing meaning in discourse is supported by means of textual cues, McNamara et al's (2010) argue. In the field of ESL writing research, commonly employed techniques are coherence studies in which the organizational and argument structures, the thematic development and the use of cohesive devices in learners' texts are analyzed (Hyland, 2019).

The growing exploitation of AI in writing teaching has stimulated scholars' renewal interest in discourse-level feedback assessment. With the use of AI in writing teaching in recent years, scholars have once again been attracted by discourse-level evaluation of feedback. There is a growing number of systems nowadays, like Chat GPT and Grammarly that supposedly give your content guidance on topics like coherence, organization and clarity, as well as grammar. Generative AI systems are shown to be able to generate more reasonably coherent and fluent texts in recent studies because these systems are trained on a vast amount of linguistic corpus (Warschauer et al., 2023). Meanwhile, there also have been scholars who have pointed out that AI generated discourse feedback may fail to be specific, generic, or out of context in responding to higher order issues regarding writing (Yoon et al., 2023).

For instance, Yoon et al. (2023) found that results of coherence and cohesion feedback given in ChatGPT tended to be too general with little guidance to effective revision strategies for students. Likewise, Ranalli (2021) notes that automated feedback systems may be able to draw out discourse-level problems, but such suggestions are typically formulaic and usually fail to take into account rhetorical context or the intentions of communication. Such restrictions are significant in the context of academic writing in which argumentation, conventions of genres and the scope of expectations of a discipline is essential for coherence. Indeed, over the years it has become evident that multiple criteria are necessary to assess discourse quality and that criteria based on processing or grammar alone are not sufficient.

The other important one is the relationships between discourse analysis and automated writing evaluation (AWE). Most AI systems are still based on a statistic language model and only pick on very grammatically correct aspects of the text, and so do not fully grasp the text semantics (Crossley et al., 2016). This has the potential of causing AI feedback to misunderstand the coherence of a discourse and ascribe too high a grade to a text's quality, as lexical richness and sentence fluency do not necessarily correspond to quality. Consequently, AI feedback can misjudge the coherence of the discourse as well as rate the quality of the text based on its lexical richness and sentence fluency. Evaluation of a discourse, especially for its effectiveness, is a challenge for the automated system contextual meaning, rhetorical purpose, audience knowledge and logical organization are among the elements that need to be taken into account (Hyland, 2019).

In addition, the attention paid to discourse studies is growing steadily accentuates the role of writing in a social and communicative environment. Johns (2015) argues that, in terms of social and cultural domains, coherence might refer to more than just what can be found in a text; it is also (culture and) reader-dependent and context-sensitive, and it is a property of communicative interaction. From this standpoint, there may be problems with these systems of automatically generated feedback being able to account for the various rhetorical norms and linguistic backgrounds ESL learners have. Thus, the current discussion about discourse emphasizes the need to take linguistically, rhetorical, pedagogical, and contextual aspects of discourse into consideration in the framework used for evaluating writing.

E. AI Ethics

As AI technology proliferates in the educational field, there have been a number of questions that have surfaced about ethical practices. With the speed in which artificial intelligence (AI) technologies have been embraced in a learning environment, there are a number of concerns pertaining to these technologies and their ethical practices in the context of ESL writing instruction. It is increasingly common to use AI tools like ChatGPT, Grammarly and AI writing evaluation (AWE) in order to support language learning with grammar correction and writing suggestions as well as revision support. These technologies come with profound pedagogical benefits, however, and bring with them

significant concerns regarding bias, dependency, hallucinated feedback, transparency, fairness and academic integrity (Sison et al., 2023). For this reason, it becomes crucial to investigate the responsible and humanist use of the technologies of artificial intelligence in education and communication, while building a new interdisciplinary area of research and reflection called AI ethics. In this sense, AI ethics is becoming a field of studies that has gained relevance in the development of responsible and humanist use of AI technologies in education and communication and that aims at building a new interdisciplinary area of studies and reflection.

One of the most talked-about problems with using AI for writing is the possible risk of bias. Bias is when the outputs of AI systematically favor linguistic, cultural, and/or social groups at the expense of others (Bender et al., 2021). Generative AI systems are created from vast linguistic datasets (corpora) from online sources, academic databases and websites, digital communication platforms, etc. AI-generated outputs can inadvertently perpetuate existing biases in the training data, which is especially critical since such data can often be influenced by dominant language norms and social inequalities. These datasets tend to reflect dominant language norms and social inequalities, meaning that AI output will likely repeat biases found in the training data (Birhane et al., 2023). The use of such structures and rhetorical devices may also be a problem in the case of non-native English writers, who often rely on structures or tropes that are not the norm in standard written English in ESL writing contexts.

Recent research has revealed the potential for unintentional pedagogic bias against learners' SLLW (second language writer). Casual differences in writing style or lexical sophistication between texts from native and non-native speakers of English caused AI-induced text detectors to label many texts—even those from ESL learners in line with the previously discussed errors. Because of these variations, many ESL texts were mislabeled AI-generated, as found by Liang et al. (2023), who noted that AI-generated text detectors were significantly biased against non-native English writers. Likewise, the proponents of AI-assistance claim that the AI tools sometimes perpetuate dominant western academic discourses, and do not properly account for different linguistic and rhetorical traditions (Warschauer et al., 2023). They highlight potential issues of favoring localizing language patterns for the native speaker community and discarding linguistic diversity in educational contexts wherein ESL is taught.

One of the other ethics problem areas is over reliance on AI technologies by the learners. AI tools offer almost instantaneous feedback and editing, but there are concerns that learners might become overly reliant on the machine and lose their critical thinking and writing skills (Ranalli, 2021). Pupils might only be accepting self-marked in Feed the Text, not actively being involved in the reflective learning processes or in their own self-editing. It raises concerns from a pedagogical perspective about learner autonomy, their cognitive skills, and the sustainability and impact of AI-powered writing instruction.

A growing number of studies suggest that students often endorse corrections without independently assessing their reliability or applicability, as they can be easily fooled into mistakes or misunderstandings by having their hand the answers (Yan & Zhang, 2024). This phenomenon is of great concern as it is in ESL writing, as reasonable cognitive activity, metalinguistic awareness and revision practice is needed in language learning. The excessive reliance on Ai feedback might foster superficial revisioning activities and inhibit in-depth learning of the language and solving skills of the language in use (Kohnke et al., 2023). In this way, scholars point out the need for AI systems to be used as complementary learning tools and not in place of human teaching and students' thoughtful contemplation.

The ethical problem with generative AI is another one that is of critical importance. This is about hallucinations in generative AI systems. AI hallucination is a phenomenon that AI models confidently generate false, fabricated, distractive and irrelevant information (Ji et al., 2023).

The power of large language models, like ChatGPT, is that they respond probabilistically on patterns in the training data, which differ from traditional grammar-checking software, which may both verify and understand. So, AI feedback can sometimes offer incorrect grammar explanations, incorrect citations, fake citations or revisions suggestions.

In the classroom, misinformation can pose significant problems for language learners who might not be able to recognize false information in AI-generated answers. Yoon et al. (2023) noted that the feedback provided by ChatGPT in terms of coherence and cohesion sometimes consisted of vague and misleading suggestions which rarely had an effective impact on the quality of learners' writing. When it comes to that, Ji et al. (2023) say hallucinations are still one of the biggest drawbacks of generative AI systems, where their made-up information can sound sophisticated and very believable. This poses problems for the reliability and trust in the educational context, as well as problems on how to use AI responsibly in an educational environment.

Issues related to authorship and academic honesty are becoming more prominent too, given the increasing adoption of generative AI in writing instruction. The role of AI-generated writing has blurred the lines between human writing and AI-generated content (Perkins et al., 2023). When using AI in writing for ESL students, there can be significant issues related to authenticity, originality, and plagiarism as students use generators to turn out whole texts and/or revisions. The current landscape of educational institutions and their expectations of acceptable use of AI in research and writing continue to be a challenge.

Additionally, the researchers stress transparency and explainability of the AI-driven feedback systems. There are numerous AI technologies that can be characterized as "black box" technologies, where the user cannot fully understand how outputs are produced. (Khowaja et al., 2023) The absence here makes it harder to assess the reliability

and fairness of feedback and the accountability of feedback givers. Ethically, there is a need to gain a deeper understanding about the restrictions, decision making processes, and perspectives entrenched in the AI system applications in learning environments.

Research Gap

Although the field of artificial intelligence (AI)-assisted writing tools in English as a Second Language (ESL) learning has seen a significant surge in research regarding writing's nature, a variety of theories, methods, and studies have led to a disparate literature in terms of theoretical orientation, methodological design, and evaluative scope (Yan & Zhang, 2024; Warschauer et al., 2023). However, there has quickly been a growing number of research efforts on generative AI tools like ChatGPT and automated writing evaluation (AWE) tools like Grammarly, but these studies typically focus on examining specific aspects of feedback analyses, such as grammatical correctness, learner satisfaction, or usability (Escalante et al., 2023; Mahapatra, 2024). Therefore, there remains no coherent or systematic framework to consider AI-produced feedback in terms of language and pedagogical levels, discourse and ethical aspects.

Many existing studies rely on narrow, sheltered definitions of what 'AI feedback' means, and are focused solely on one aspect of AI feedback. In addition, numerous empirical research investigations only measure AI tools at the level of language corrections, particularly grammatical and vocabulary correctness, and fail to measure higher-order features of writing like coherency, arguments, or rhetoric (Yoon et al., 2023). Other research studies are likewise limited to learner perceptions of AI tools without considering the pedagogical quality or discourse-level effectiveness of feedback that is provided (Ranalli, 2021). In fact, systematic reviews of automated writing evaluations corroborate that the current research literature focuses primarily on technical rather than pedagogical and/or interpretive aspects of feedback quality (Link et al., 2022; Stevenson & Phakiti, 2019).

Moreover, while emerging studies continue to point towards the possibilities generative AI can bring to ESL framework development, inconsistencies in evaluation practices emerge at the same time. We want to highlight just one example: ChatGPT feedback has been found to enhance revision process and writing fluency, although its impact on the coherence, rhetorical organization and academic argumentation has not been consistent (Yan & Zhang, 2024; Warschauer et al., 2023). Likewise, some systematic review studies were conducted recently, which reported that AI tools are being adopted globally for formative assessment or writing assistance and the evaluation criteria applied across diverse systematic review studies are not uniform, therefore making comparison and generalization difficult (Settiawan, 2025; Crosthwaite & Sun, 2025). This non-standardization confirms the large methodological gap in this field.

A second key research missing is the lack of more multi-layered assessment procedures combining accuracy of language, quality of discourse, pedagogical relevance, reliability of the AI, and ethical dimension into a unified model of assessment. In existing research, these these dimensions tend to be analyzed separately from each other and not as integrative and interrelated parts of one system. In particular, the study of linguistic errors does not necessarily take overall discourse analysis into account (Kouchi et al., 2023; Kuo et al., 2023), and ethical issues like bias, hallucination, and overdependence are treated independently of overall discourse analysis in AI ethics literature sources (Lima et al., 2022; Gianorgas et al., 2023). This disintegration makes it difficult to cultivate even general observations on whole analyses of the AI-generated feedback in the ESL writing domain. This fragmentation hinders the researchers in building overall observations regarding the AI-generated feedback in the ESL writing domain.

Furthermore, the concepts of AI ethics when discussed has become relevant in the domain of writing but has not been implemented as a part of a model in empirical or analytical writing evaluation. Some of the major ethical concerns identified in the study of generative AI include algorithmic bias, hallucinated feedback, dependency by learners on the AI tools, and lack of transparency (Bender et al., 2021; Ji et al., 2023). However, the above said concerns are usually treated as individual theoretical frameworks of discussion, rather than being introduced as a part of evaluation frameworks for ESL writing feedback in practice. The separation between the theory of ethics and the applied research in education is highlighted.

Moreover, the evaluation of a discourse level is still undeveloped in the field of AI feedback research. Although Textual coherence and cohesion are considered key determinants of the quality of writing by Halliday and Hasan (2016); discourse aspects were not considered to be the primary focus of most of the AI systems (Crossley et al., 2016; Hyland, 2019). Recent studies have revealed that AI-generated feedback sometimes lacked comprehensive feedback on the coherence, cohesion, and rhetorical development, and rather gave general or superficial feedback (Yoon et al., 2023). This is one point that reveals a vast disconnect between the theory of language and the implementation of AI in the assessment of writing.

Proposed Model

Naming the Model

The current study aims to respond to the theoretical dilemma of the fragmentation in the field of AI-assisted ESL writing studies by proposing a new integrative model called Hybrid AI Feedback Evaluation Model (HAFEM). The model can be considered a multidimensional analytical framework for systematically analyzing the feedback that is created by AI in English as a Second Language (ESL) writing contexts. The designation “Hybrid” stems from its

interdisciplinary basis that incorporates ideas from applied linguistics, corpus linguistics, discourse analysis, pedagogical theory, and the ethics of artificial intelligence. In recent years, the increasing use of generative AI, like ChatGPT, in the education sector has created a dire need for models that integrate the evaluation of the capabilities and knowledge gained from such technologies into broader educational institutions (Warschauer et al., 2023; Yan & Zhang, 2024).

1. Conceptual Foundation of HAFEM

The HAFEM model does not rely on merely one-dimensional measurement, indicating the recognition of the fact that feedback about writing for the L2 community needs to be expressed by humans, particularly in an ESL environment. The HAFEM model is based on the knowledge that artificial intelligence (AI) produced feedbacks in ESL writing cannot be effectively measured in one dimension. Current research is mainly centered on independently studying aspects of a system's output, whether they refer to grammatical accuracy and/or learner-perception, or the usability system, not connecting discourse, pedagogy, and ethics in a single analytical framework (Escalante et al., 2023; Mahapatra, 2024). These disconnections hinder the creation of stronger theoretical models that integrate the way that AI feedback works with varying levels of proficiency in writing.

The recent literature points to a tendency in generative AI models like ChatGPT for outputs that are coherent, naturalistic in their language but at times may fall short in terms of both accuracy and context (Yoon et al., 2023; Ped agonists and Gooff, 2023). Likewise, the work conducted on automated writing evaluation (AWE) systems indicates that automated systems can boost the accuracy of the surface-level performances but cannot assist the higher-order ones like building coherence and developing arguments (Link et al., 2022; Stevenson & Phakiti, 2019). The results justify establishment of an integrated evaluation model on theoretical level.

The concept of HAFEM Model is illustrated as below

The Hybrid AI Feedback Evaluation Model (HAFEM) consists of 5 analytical dimensions, each connected to the rest of the model. Each dimension is a key lens used to critique the writing received from AI in ESL.

The degree to which the language you use is accurate.

The accuracy of the language used (i). This dimension assesses correct grammar, word choice, sentence structure and precision of terms. It is rooted in, and extends, traditional error analysis models (Corder, 1967; Ellis, 2009) and draws on corpus-based linguistic model (Biber et al., 2021). The accuracy of AI feedback in identifying and editing language errors, while maintaining clarity and avoiding any distortion or overcorrection, is evaluated.

The selected items are from the Discourse Quality Dimension.

This dimension looks at coherence, overall paragraph organization, and rhetorical development. It is based on the modern theories in discourse studies, which focus on the unity and argumentation of discourse (Hyland, 2019), and Halliday and Hasan's cohesion theory (1976). The AI-generated feedback is assessed based on its capacity to assist with coherent organization of the text (with the exception of single sentence correctness).

This dimension evaluates how effective the given results of AI regarding instruction, such as scaffolding, explanations, guiding learners and support in revision, are on the instructional level.This dimension examines the instructional value of the results given by AI: scaffolding, explanation, guiding learners, revision support. It is based on socio-constructivist learning theory (1978 Vygotsky) and research on feedback that focussed on dialogic and developmental learning processes (2006 Hyland & Hyland). Research revealed that proper feedback could have an impact to learners' autonomy rather than passive acceptance of the errors made by the learners (Ranalli, 2021).

(iv) AI Reliability Dimension

In this dimension, AI-provided feedback is looked at for consistency, accuracy and trustworthiness. It includes the ability to recognize hallucinated outputs, contradictory suggestions, and/or contextually inappropriate corrections. AI tools powered by generative AI also have been shown to be capable of generating convincing but wrong answers (Ji et al., 2023; Sison et al., 2023), which can create complications for educational reliability.

(v) 3-Ringed Ethical Integrity Dimension.

This dimension relates to fairness, bias, transparency, writer concerns and risk of dependency from the learners. Buccino et al. (2023) and Bender et al. (2021) found in their studies on artificial intelligence ethics that language models can be influenced by social and language biases in their access to training data, and that this can lead to disproportionate impacts on behalf of non-native English writers. This dimension aims to assess logistical systems based on AI for their ethical responsibility in an educational setting alongside its technical performance.

Model integrated theoretically

The use of different theoretical traditions is a strength of HAFEM that is integrated into it. These elements are Error Analysis Theory, corpus linguistics, discourse analysis, pedagogical theory, and AI ethics properly equip the learner to fully embrace the enigma of errors and their underlying causes. Altogether these elements set the groundwork for the learner to fully confront the enigmatic nature of errors and what may be motivating them.

HAFEM approaches these domains as parts of a larger whole, as opposed to treating them individually like is done by current models. For instance, a grammatical correct suggestion is linguistically correct, but it might not be pedagogically effective because it makes the message incoherent (discourse dimension), and/or it might be ethically disturbing because it supports problematic language norm (ethical dimension). The integration is contrasted by recent proposals of multidimensional AI assessment proposals in educational research literature that emphasize the

evaluation of AI implementation and its impact from multiple dimensions (Warschauer et al., 2023; Yan & Zhang, 2024).

7. Test Procedures for HAFEM

The HAFEM model is not only intended for future empirical applications in both qualitative and quantitative studies, it is also intended for evaluating the model to enable the application to other work units. The model can be applied to the analysis of texts via rubric-based evaluation, corpus comparison methods, discourse mapping, or mixed methods analysis. In each dimension, there can be different metrics which can be measured separately and combined into a total "AI feedback".

For example:

Linguistic errors could be studied by analyzing the frequency of the errors made in the corpus.

The cohesion indices can be used to measure the quality of discourse.

Learner uptake studies are possible means for measuring pedagogical value

HAFEM has three key roles to fulfil in the field. First, it offers a multi-structured model to rate feedback that goes beyond grammar, giving more than a one-dimensional perspective on the value of the AI-generated feedback. Second, it includes more than pedagogical, discourse, linguistic and ethical aspects under one theoretical umbrella. Finally, it fills a gap in methodology concerning the research on the use of AI in ESL writing instruction, by providing a standard measurement of the tasks for future empirical investigation.

Much like reports' capacity to connect with homework, generative AI's integration is also transforming the language education landscape, making it vital to adopt theoretically sound models to ensure that Language AI feedback systems are used responsibly, transparently, and pedagogically effectively. The current calls for interdisciplinary evaluation of AI in the learning field urge researchers to develop frameworks that reflect new learning opportunities for people to learn, engage, and act with AI tools while measuring continuous impact. HAFEM addresses this need by providing an extensive blueprint that aligns with the present academic literature on using interdisciplinary approaches for assessing AI tools in education (Escalante et al., 2023; Sison et al., 2023; Warschauer et al., 2023).

HAFEM Components

The Hybrid AI Feedback Evaluation Model (HAFEM) would be a five-layer analytical model to systematically analyze the AI-generated feedback in the context of ESL writing. They combine the in-depth insights of applied linguistics, corpus research, pedagogy, computational linguistics and AI ethics to reflect three interrelated layers of the quality of writing and feedback (Warschauer et al., 2023; Yan & Zhang, 2024).

Component 1 - Linguistic Layer

At the micro level, The Linguistic Layer assesses the accuracy of students' Language (Grammar, Syntactic and Word choice) in the AI-generated feedback given to students. This component evaluates the ability of AI systems to accurately recognize and correct errors in language (e.g. agree with the noun/pronoun, the tense of verbs, sentence structure, and word selection). The theory about this layer is derived from Ellis' (2009) and Corder's (1967) conception of Error Analysis Theory, which views L2 learner errors as systematic developmental stages in L2 learning instead of random errors. Furthermore, Corpus Linguistics can reveal authentic frequencies and patterns of errors in learner corpora on a large scale, which has also served as empirical evidence of Corpus Linguistics (Biber et al., 2021). According to Mahapatra (2024), academic authorship is another important dimension to assess the accuracy of the feedback given by AI systems like ChatGPT and Grammarly as they use probabilistic models of language for their text production.

Component 2 - Discourse Layer

The Discourse Layer is concerned with the way a text is organised at the macro level that is cohesion, coherence, paragraph types and order of ideas. This component assesses if AI-generated feedback advances learner's ability to create coherent and organized texts, or just fixes grammatical and sentence-level errors. Halliday and Hasan's (1976) cohesion theory has been used as the theoretical framework for this layer, which explanations how the unity of text is related to linguistic devices which are concerned with reference, conjunction, substitution and lexical cohesion. Modern theories of discourse also highlight the idea that coherence does not solely lie in the structure of a discourse, but also comes from the interpretation and construction of meaning by the reader as well as the way interpretation and construction are put into practice within the discourse. (Hyland, 2019). In recent research, AI-generated feedback has been found to often lack rhetorical support and frequently only gives general organization advice at the discourse level (Yoon et al., 2023; Warschauer et al., 2023).

Component 3 - Pedagogical Layer

The Pedagogical Layer takes into account the possible impact of AI feedback on the effectiveness of instruction, as well as scaffolding, learner autonomy, and whether it is helpful for writing growth. This will evaluating the use of AI-generated feedback to help with learning decisions instead of just to fix grammatically. The theoretical basis is D'Entremont's (1985) sociocultural theory that focuses on the idea that learning takes place with the help of other people and supports within the learner's zone of proximal development as defined by Vygotský (1978). Feedback to students in ESL writing contexts should support students' self-correction, reflection and independence (Hyland et al., 2006). AI-generated corrections, however, were found to be superficial as opposed to significant and often offered in

a passive manner without encouraging deeper thinking and reflection, which could make students less independent in their learning process and become more reliant on AI systems (Ranalli, 2021; Yan & Zhang, 2024).

Component 4 (AI Reliability Layer)

The AI Reliability Layer evaluates the factual correctness of the AI-generated feedback, risk of hallucination, and stability of the AI's answer. This layer assesses whether the AI system consistently gives sensible, contextually appropriate and non-contradictory feedback for writing tasks. The theoretical ground is employed measures of NLP evaluation (such as error rate, consistency, and accuracy for outputs produced by a model) (Ji et al., 2023). One important challenge that lies in this layer is that of "AI hallucination" in which generative models generate plausible, but inaccurate or fabricated information (Ji et al., 2023). These inconsistencies can drastically reduce the trustworthiness of the feedback provided by AI in educational environments, especially when students are not aware of errors (Bosma et al., 2023).

Component 5 - Ethical Layer

The Ethical Layer caters to issues of the AI generated feedback systems that impact on fairness, bias, dependency, transparency and academic integrity. This part assesses the ability of AI tools to ensure that all learners are treated fairly, do not commit a language-biased or culturally-biased approach, and contribute to judicious use in education. The theoretical aspect is rooted in AI ethics principles, such as fairness and accountability, transparency and human-centered design, which highlight the importance of Steering Artificial Intelligence systems with these four principles (Bender et al., 2021; Khowaja et al., 2023). AI systems have been found to have biases toward writers of native English, and can perpetuate dominant linguistic norms, which can harm ESL learners (Liang et al., 2023). We also wish to share ethical concerns regarding learner overdependence and lowering the level of academic honesty, adding to the significance of incorporating ethical evaluation in AI feedback studies (Perkins et al., 2023).

The framework is aimed at future empirical research with feedback in ESL writing tasks that are generated by AI. This model can enable the researchers to do systematic and multidimensional evaluation, combining the method of qualitative and quantitative.

6. Implications of the Study

The proposed model called Hybrid AI Feedback Evaluation Model (HAFEM) has far-reaching impacts on ESL teaching, the use of AI in writing tasks, computational linguistics research, and educational policymaking. With the introduction of more and more generative AI models like ChatGPT in language education, it is essential to have a systematic and multi-dimensional approach to evaluating their pedagogical impact (Warschauer et al., 2023; Yan & Zhang, 2024).

ESL Relevance

Within the ESL classroom, the HAFEM framework offers a lenses-based approach that helps the teacher analyze AI-generated feedback that goes beyond mere grammatical correctness. In recent research, writing pedagogy and discourse-level development, learner autonomy, and pedagogic scaffolding are explicitly linked to the importance of traditional writing feedback, emphasizing linguistic accuracy (Hyland, 2019; Hyland & Hyland, 2006). Through the use of linguistic, discourse, and pedagogical elements, the model invites instructionally savvy ESL professionals to apply more strategically with AI tools for writing instruction instead of using them as generic correction devices. Moreover, it helps educators to evaluate if AI-based feedback helps or hinders students or if it replaces the teacher's assessment for the students with AI-recommended suggestions (Ranalli, 2021). That has direct consequences for classroom work, as teachers have to master technological processes and find the appropriate ways to juxtapose it with pedagogical ones.

Training applicable to AI Assistant writing

The HAFEM model will serve as a holistic lens for grasping the advantages and drawbacks of automated feedback systems in AI-assisted writing environments. AI tools like Grammarly and ChatGPT have been proven to improve writing fluency and grammatical correctness, but not consistently in terms of offering feedback on the discourse level or being context aware (Escalante et al., 2023; Yoon et al., 2023). Multidimensional evaluation can help uncover strengths and weaknesses in AI systems, facilitating improvement and development. This involves more coherent feedback, pedagogically better explanations and more generic unrelated responses. In addition, the model focuses on the importance of designing writing tools ethically in promoting learner autonomy and not dependence on a tool for automatic grammar correction.

It outlines several possible implications for Computational Linguistics.

The HAFEM framework plays a key role in the evolution of more sophisticated evaluation techniques for natural language processing (NLP) systems in the field of education. Existing metrics to assess the quality of an artificial intelligence system's performance have tended to emphasize accuracy, fluency, or statistical performance, while overlooking pedagogical and discourse-level aspects of AI quality (Ji et al., 2023). The proposed model adds to computational evaluation through combining the linguistic error analysis, discourse structure assessment and ethical considerations with AI feedback evaluation. This interdisciplinary integration is consistent with recent calls for more human-centered assessments for AI with reference to its applicability in real-world educational contexts, beyond

technical measure of performance (Bender et al., 2021). This is why HAFEM can help direct future computational linguistics work towards meaningful educationally focused design and testing of AI systems.

Consequences for Educational Policy.

The growing significance of AI in the teaching of writing, from a policy standpoint, poses pertinent regulatory, ethical, and institutional considerations. At present, there is a lack of consistent policies and practices to assess and incorporate AI feedback into educational curriculum in many schools. The HAFEM framework can offer a structured, evidence-informed approach to creating guidelines for the wise use of AI in language learning. Specifically, the ethical aspects of the model reflect the existence of bias, transparency, dependency, and academic integrity, which are paramount concerns in the context of AI in education (Liang et al., 2023; Perkins et al., 2023). The use of multidimensional evaluation criteria can guarantee that AI technologies are utilized in a way that improves learning outcomes, fairly, and ethically, preventing potential plagiarism and AI-driven cheating.

7. CONCLUSION

The study has introduced the Hybrid AI Feedback Evaluation Model (HAFEM) as an all-encompassing, interdisciplinary framework to evaluate AI-produced ESL writing feedback. Generative AI technologies in education are under rapid development, which gives rise to both opportunities and challenges for language learning, and especially automated feedback systems for writing. Many existing studies have shown that AI tools can improve the accuracy and speed of writing, but the studies have also identified many limitations regarding discourse-level support, pedagogical effectiveness, and ethical reliability (Warschauer et al., 2023; Yan & Zhang, 2024).

The fact that this study has reached is multidimensional analytical frameworks are needed in the future of ESL writing evaluation using AI instead of single-dimensional, one-facet evaluation. The extant research is limited in scope, focusing on the linguistic level, the perception of the learner, or the performance of the system rather than on a set of integrated pedagogical, discourse and moral issues (As for Hyland, 2019; Escalante et al., 2023). This division makes it difficult to cultivate a comprehensive grasp of the value of AI generated feedback for real-world education scenarios. The suggested HAFEM framework aims to overcome this methodological gap by combining five related aspects: Linguistic accuracy, discourse quality, pedagogical relevance, AI reliability, and ethical integrity. The model integrates error analysis, corpus linguistics, discourse theory, sociocultural learning theory and AI ethics to offer a cohesive framework for assessing AI-generated feedback in ESL writing. This integration facilitates more systematic, transparent, and thorough analysis of AI feedback systems.

Moreover, the study underscores the role of the HAFEM model in fostering interdisciplinary research development, as it connects various domains within applied linguistics, computational linguistics, education, and AI ethics. With the continuous development of AI, there is a growing demand for collaborative mechanisms that can deal with technical performance and human-centered concerns in schooling. Considering the ongoing evolution of AI, it is crucial to foster collaborative structures that can support both technical performance and human-centred concerns in education. Thus the model brings a methodological novelty, but also to the academic field of responsible use of AI in language teaching, in general.

To sum up, this study recommends moving away from the traditional single-factor and highly-granular approach and introducing a multidimensional, theoretically informed and ethically concerning model when assessing the effectiveness of feedback produced by AI. The HAFEM model offers a starting point for such progress by providing a framework for future empirical studies, education, and development of policy on AI use in ESL writing.

REFERENCES

1. Banihashem, S. K., Taghizadeh Kerman, N., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? International Journal of Educational Technology in Higher Education, 21(23).
2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. Proceedings of FAccT, 610–623.
3. Biber, D., Conrad, S., & Reppen, R. (2021). Corpus linguistics: Investigating language structure and use. Cambridge University Press.
4. Birhane, A., Prabhu, V. U., Kahembwe, E., & Njoroge, S. (2023). The values encoded in machine learning research. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 173–184. Bitchener, J., & Ferris, D. (2012). Written corrective feedback in second language acquisition and writing. Routledge.
5. Corder, S. P. (1967). The significance of learners' errors. International Review of Applied Linguistics in Language Teaching, 5(4), 161–170.
6. Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. Proceedings of the Annual Meeting of the Cognitive Science Society, 33(33), 1236–1241.

7. Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. Behavior Research Methods, 48(4), 1227–1237. Crosthwaite, P., & Sun, S. (2025). Generative AI and L2 written feedback studies: A scoping review. RELC Journal.
8. Ellis, R. (2009). A typology of written corrective feedback types. ELT Journal, 63(2), 97–107.
9. Ellis, R., & Barkhuizen, G. (2005). Analysing learner language. Oxford University Press.
10. Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing. International Journal of Educational Technology in Higher Education, 20(57).
11. Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). Journal of Second Language Writing, 8(1), 1–11.
12. Ferris, D. R. (2011). Treatment of error in second language student writing (2nd ed.). University of Michigan Press.
13. Ferris, D. R. (2014). Responding to student writing: Teachers' philosophies and practices. Assessing Writing, 19, 6–23.
14. Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. Longman.
15. Hyland, F., & Hyland, K. (2006). Feedback on second language students' writing. Language Teaching, 39(2), 83–101.
16. Hyland, K. (2019). Second language writing (2nd ed.). Cambridge University Press.
17. Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. Language Teaching, 39(2), 83–101.
18. James, C. (2013). Errors in language learning and use: Exploring error analysis. Routledge.
19. Jeon, E.-Y. (2025). Artificial intelligence in ESL/EFL education: Evidence from recent reviews (2024–2025). International Journal of Learning, Teaching and Educational Research, 24(1).
20. Ji, Z., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.
21. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.
22. Johns, A. M. (2015). Genre awareness for the novice academic student: An ongoing quest. Language Teaching, 48(4), 499–511.
23. Khowaja, S. A., Khuwaja, P., Dev, K., Wang, W., & Nkenyereye, L. (2023). ChatGPT needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) evaluation: A review. arXiv.
24. Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. RELC Journal, 54(2), 537–550.
25. Lee, I. (2017). Classroom writing assessment and feedback in L2 school contexts. Springer.
26. Li, J., Huang, J., Wu, W., et al. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. Humanities and Social Sciences Communications, 11, 1268.
27. Liang, W., et al. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), 100779.
28. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), 100779.
29. Link, S., Dursun, A., Karakaya, K., & Hegelheimer, V. (2022). Automated writing evaluation and second language writing. Language Teaching Research, 26(3), 389–413.
30. Link, S., et al. (2022). Automated writing evaluation and second language writing. Language Teaching Research, 26(3), 389–413.
31. Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. Smart Learning Environments, 11(9).
32. McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.
33. McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. Written Communication, 27(1), 57–86.
34. O'Neill, R., & Russell, A. (2019). Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. Australasian Journal of Educational Technology, 35(1), 42–56.
35. Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2023). The AI assessment scale (AIAS): A framework for ethical integration of generative AI in educational assessment. arXiv.
36. Ranalli, J. (2021). L2 engagement with automated feedback. Journal of Second Language Writing, 52, 100816.
37. Settiawan, D. (2025). Leveraging ChatGPT in EFL writing assessment. Journal of English and Arabic Language Teaching.
38. Sison, A. J. G., Daza, M. T., Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). ChatGPT: More than a weapon of mass deception, ethical challenges and responses from the human-centered artificial intelligence perspective. arXiv.
39. Stevenson, M., & Phakiti, A. (2019). Computer-generated feedback effects. Assessing Writing, 19, 51–65.
40. Truscott, J. (1996). The case against grammar correction in L2 writing classes. Language Learning, 46(2), 327–369.

41. Vygotsky, L. S. (1978). Mind in society. Harvard University Press.
42. Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.
43. Warschauer, M., et al. (2023). Affordances and contradictions of AI-generated text. Journal of Second Language Writing, 62, 101071.
44. Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. Journal of Second Language Writing, 62, 101071.
45. Yan, D., & Zhang, S. (2024). ChatGPT feedback in ESL writing. Humanities and Social Sciences Communications, 11, 1086.
46. Yan, D., & Zhang, S. (2024). L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. Humanities and Social Sciences Communications, 11, 1086.
47. Yang, H., Gao, C., & Shen, H.-Z. (2024). Learner interaction with, and response to, AI-programmed automated writing evaluation feedback in EFL writing: An exploratory study. Education and Information Technologies, 29, 3837–3858.
48. Yoon, S.-Y., et al. (2023). Evaluation of ChatGPT feedback on coherence and cohesion. arXiv.
49. Yoon, S.-Y., Miszoglad, E., & Pierce, L. R. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. arXiv.