International Journal of English Language Studies
ISSN: 2707-7578
DOI: 10.32996/ijels
Journal Homepage: al-kindipublisher.com/index.php/ijels/index

Stance and Feedback Focus in English Writing: A Corpus-Assisted Analysis of Human and GPT-4-Augmented Comments

Asjad Ahmed Saeed Balla[1], Nehad Salah Mohammed Ramadan[2], and Prof. Omer Elsheikh Hago Elmahdi ✉[3]

[1] Department of English Language, Faculty of Education, Gedarif University, Sudan Department of English Language and Literature, College of Languages and Humanities, Qassim University, Saudi Arabia

[2] Department of Foreign Languages, Faculty of Education, Mansoura University, Egypt Department of English Language and Literature, College of Languages and Humanities, Qassim University, Saudi Arabia

[3] The Open University of Sudan, Affiliated to Department of Languages & Translation, College of Arts & Human Science, Taibah University, Saudi Arabia

Corresponding Author: Prof. Omer Elsheikh Hago Elmahdi**, E-mail:** Ohago65@gmail.com

ABSTRACT

Generative AI can reshape English writing feedback, yet how human-informed augmentation changes its wording remains insufficiently understood, and textual change should not be mistaken for pedagogical improvement. This study examines how GPT-4 augmentation recontextualizes stance and metalinguistic focus, providing evidence for critically evaluating AI-supported feedback. A retrospective corpus-assisted analysis compared 4,917 complete human–augmented feedback pairs from the second LEAF release; the human comments came from a single educational consultant. Methods combined paired lexical profiles, markup, quotation and deduplication sensitivity checks, length-aware lexical diversity, an exploratory two-model audit of 64 pairs, and four source-linked close readings. Under the original tokenization, augmented comments averaged 165.2 tokens longer; hedging-associated and positive-evaluation densities increased in 82.0% and 79.1% of pairs, respectively. Task-reference density fell in 67.5% of pairs despite greater task-lexicon presence, while grammar/form and argument/meaning densities increased. Conventional type–token ratios fell, but MATTR-50 increased. Contextual analyses indicated that criticism and task-directed recommendations persisted. The findings describe expansion and reformulation, not demonstrated feedback quality, learner agency or learning gains; their generalizability is limited by the single-contributor, human-conditioned pipeline.

KEYWORDS

Writing feedback; LEAF; stance; metalinguistic discourse; human-informed GPT-4 augmentation; corpus-assisted analysis.

ARTICLE INFORMATION

ACCEPTED: 07 October 2026 **PUBLISHED:** 07 November 2026 **DOI:** 10.32996/ijels.2026.8.5.4

1. Introduction

Generative language models can produce written feedback at scale, renewing questions about its educational value and risks (Elmahdi et al., 2024). A feedback comment does more than identify a writing problem: it can evaluate a textual choice, suggest an action, and address the writer. Such wording may frame a concern as an obligation, suggestion, option or observation, and may foreground form, argument or task requirements. These textual choices are resources for interpretation; they do not by themselves show what a learner understood, revised or learned.

This study examines the second release of LEAF (Language Learners' English Essays and Feedback Corpus), which contains Essay Forum essays and feedback (Behzad et al., 2024b). Its human field retains comments from one educational consultant, Mary Rose. GPT-4 was prompted to revise previously generated feedback by incorporating elements of the human comment, producing the released augmented field. The paired texts therefore represent a particular human-conditioned pipeline, not independent human and AI responses or a representative sample of teachers. Prompt and rubric design, prior feedback, inherited diagnoses and model-mediated reformulation may all contribute to the observed differences.

An earlier study compared online human feedback with GPT-3.5 rubric-based feedback (Behzad et al., 2024a). The present comparison instead describes how the LEAF augmentation pipeline recontextualizes existing comments through stance and

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

metalinguistic focus. Its contribution is an account of textual change, not a claim that augmentation improves feedback quality or learning.

2. Theoretical framework

2.1 Textual affordances, stance, and engagement

Hyland and Hyland (2001) show that praise and criticism must be interpreted in context: praise can mitigate criticism, while negative evaluation can identify a substantive issue. Hyland's (2005) model of stance and engagement provides a lens for examining how a text presents positions and addresses readers. Although developed for academic discourse rather than online writing feedback, it helps interpret modality, evaluation, reader address and invitations to consider alternatives. Isolated markers do not establish authority, politeness or learner agency.

Textual affordances must be distinguished from learners' actions. Feedback literacy concerns the understandings, capabilities and dispositions through which students interpret and use feedback (Carless & Boud, 2018; Boggs & Manchón, 2023). Length, positivity and softened recommendations cannot serve as proxies for uptake. Reasons and options may help a writer evaluate advice, but do not demonstrate that the writer made a choice or revised effectively.

Finite dictionaries provide a transparent surface-level operationalization of stance/address and writing-target language. The stance/address lexicons screen for forms associated with directives, hedging, certainty, positive or negative evaluation, and writer address or choice. Separate lexicons screen for grammar/form, lexis/style, cohesion/organization, argument/meaning and task-reference language. These labels describe literal matches, not verified speech acts, intentions, diagnostic accuracy, coverage or priority. Positive evaluation is an interpersonal category, not a writing-target domain; the same expression can serve different functions in context.

Source-linked close readings of selected matched comments use pseudonymous record locators and relevant essay context to examine how targets, recommendations and reasons are expressed. The selected cases illustrate and complicate lexical contrasts rather than estimate their prevalence. They are interpretive readings of corpus examples, not independent human validation. Together, the lexical profiles and contextual readings describe textual recontextualization without equating discourse markers with outcomes.

Determining whether feedback improves learning, agency or co-construction requires evidence from writers, revisions, dialogue and independently validated judgments of feedback.

2.2 What existing evidence does—and does not—establish

Work on ChatGPT in language education identifies potential uses alongside risks such as inaccurate responses and the need for informed use (Barrot, 2023; Kohnke et al., 2023). Li et al. (2024) review the first year of research on ChatGPT in language education. Such studies provide context for evaluating feedback, but do not establish that a lexical profile measures feedback quality or learning.

Conventional automated writing evaluation (AWE) provides a related but distinct evidence base. Ngo et al. (2024) synthesize AWE research, while Sari and Han (2024) examine AWE in relation to EFL learners' writing-related outcomes. Zhai (2026) studies automated written corrective feedback. These studies cannot by themselves establish how generative AI reformulates the LEAF feedback or how writers respond to that reformulation.

Studies of ChatGPT-assisted writing have examined writing and motivation (Mahapatra, 2024; Song & Song, 2023), while Zhang et al. (2025) compare generative-AI and hybrid feedback arrangements. Other work considers conversational AI for particular learner populations (Elkot et al., 2025) and ethical and cultural issues in EFL education (Elmahdi et al., 2025). These contexts differ from the LEAF augmentation procedure and do not establish that differences in feedback wording cause learning gains.

Uptake must be observed in writers' responses. Yan and Zhang (2024) examine L2 writers' engagement with ChatGPT corrective feedback, and Zou et al. (2026) compare uptake of teacher- and ChatGPT-generated feedback. Research on dialogic peer feedback likewise examines interaction between participants (Zhao et al., 2024). Such evidence is distinct from counts of pronouns or modal verbs. Feedback may also have adverse effects (Yu et al., 2021); neither positivity nor criticism is inherently beneficial or harmful in isolation.

2.3 A corpus-assisted account of textual change

Written feedback is shaped by the relationship between writer and recipient and by its social and institutional context (Hyland & Hyland, 2006a, 2006b). Stance in learners' essays varies across writers and contexts (Afzaal et al., 2025; Black, 2022), while supervisory feedback occurs within a distinctive institutional relationship and revision history (Zhang & Hyland, 2022). Findings from these settings should not be transferred uncritically to online feedback.

Corpus-assisted analysis can use frequency patterns to guide contextual reading rather than treating counts as substitutes for interpretation (Baker et al., 2008). That precedent does not validate the LEAF categories. The present study combines a common-cohort paired profile with sensitivity checks and source-linked readings to distinguish visible recommendations and evaluations from unobserved pedagogical effects.

2.4 Research questions

1. How do displayed-text length and the densities of literal stance/address and writing-target lexemes differ between human and GPT-4-augmented comments in the complete-pair LEAF release, using pooled rates and within-record comparisons?

2. How robust are these differences to markup and quotation handling, exact-source and exact-essay deduplication, and length-aware lexical-diversity measurement?

3. What convergent or divergent contextual readings of evaluation, recommendations, options and writing targets emerge from a uniformly selected 64-pair exploratory two-model audit and source-linked examples?

The first two questions describe the pinned release; the third relies on model coding and interpretation, not independent human validation or observed learning effects.

3. Methods

3.1 Design, source and comparison

This retrospective corpus-assisted study compares original human feedback with human-informed GPT-4-augmented feedback on the same essay records in the second public LEAF release. It is not a randomized intervention or a comparison of independently generated human and AI comments. According to Behzad et al. (2024b), GPT-4 was asked to revise prior generated feedback using elements of the human comment. The dated GPT-4 snapshot and complete per-record generation logs are not recoverable from the source paper and are not reconstructed here.

The source is the Educational Testing Service repository at commit 685ad279d0b42d6c1e2b1b70ac8aae5a6d8fb60f, dated 18 September 2024 (Educational Testing Service, 2024). Repository documentation identifies Mary Rose as the consultant represented in the human-feedback field. These comments are not a representative sample of teachers. The pinned release, rather than the approximate size of an earlier LEAF release, defines the present corpus.

The Essay Forum material is relevant to English writing feedback, but the writers' first languages, educational backgrounds, EFL status and proficiency were not independently verified for this analysis.

3.2 Preprocessing and unit of analysis

The unit of analysis is an essay linked to a human comment and an augmented comment. Comparative analyses retain records with a nonempty essay and both feedback fields; feedback fields are checked for alphabetic tokens, and zero-token cases are tracked separately. All comparative indicators use the same complete pairs. HTML entities and markup are processed, and Unicode and whitespace are normalized. Word tokens are alphabetic sequences with optional internal apostrophes. The original tokenizer also counted literal backslash-n paragraph markers in the human field as 'n' tokens, so the reported word counts and density denominators include this formatting artifact. Quoted essay material and deletion-marked text remain in the displayed-text measure; this is not a validated provider-only commentary layer. Processing does not alter the original source records.

Sensitivity analyses examine raw strings retaining markup, displayed text with explicitly tagged quotations removed, and a conservative layer that additionally masks balanced double-quoted spans. The last layer cannot reliably distinguish learner quotations from example rewrites or other quoted material and is a detector sensitivity check, not a validated provider-only corpus. The lexical analysis does not measure sentence- or clause-level feedback moves.

3.3 Literal lexicons and paired estimands

Six finite stance/address lexicons screen for directive-, hedging- and certainty-associated forms, positive- and negative-evaluation words, and writer-address/choice forms. Five writing-target lexicons screen for grammar/form, lexis/style, cohesion/organization, argument/meaning and task-reference language. Unlisted synonyms and inferred inflections are not counted.

Within each category, normalized lowercase matching proceeds longest-first and left-to-right, with no overlapping matches inside that category. A multiword expression counts once without an additional nested hit for a component in the same category. Cross-category overlap is retained because a form can occur in more than one lexicon. These literal counts are not validated measures of speech acts, agency, politeness, authority or diagnostic accuracy; contextual evidence helps interpret ambiguity, including negation and quotation.

For each category, the pooled rate is 1,000 × total hits divided by total words in the retained field. A document-level rate is 1,000 × hits divided by words in that document; the paired difference is the augmented rate minus the human rate for the same record. Means, medians, quartiles, presence proportions, ties and the proportions of positive and negative differences describe the finite corpus. Pooled rates weight tokens, whereas summaries of paired differences weight records. These comparisons do not estimate a generic human-versus-AI provider effect.

The conventional type–token ratio (TTR) is reported alongside the moving-average type–token ratio with a 50-token window (MATTR-50) on the common eligible cohort (Covington & McFall, 2010). These measures describe lexical diversity, not feedback quality or template use.

3.4 Corpus dependence and robustness

Normalized source URLs and exact essay texts are checked for duplication. The main analysis retains the release as a census; sensitivity profiles retain one record per source or per exact essay using a reproducible source-order rule. Near-duplicate similarity flags are audited rather than used as an unvalidated deletion rule. Corpus split labels are inherited metadata, not separate learner populations. These descriptive and robustness analyses cannot overcome the limits of one contributor, a human-conditioned generation pipeline or restricted external validity.

3.5 Contextual machine-coding audit and close reading

A uniformly random sample of 64 rows was selected without replacement from the full 4,918-row handoff using seed 20261017. All sampled rows contained complete fields, and no outcome-based replacements were made. Seed 20261018 randomized assignment of the two feedback fields to Text A and Text B. These are sampling and assignment seeds, not documented model-generation seeds. Inputs omitted provider labels, source URLs, essay titles and writer identities; style and quotation metadata could nevertheless reveal the source, so blinding was procedural rather than guaranteed.

The audit used the requested model aliases gpt-5 and gemini-3.1-pro-preview as separate conditions, each receiving paired texts and the same codebook without access to the other's output. The run took place on 17 October 2026, from 17:15:11 to 17:22:13 UTC. Reasoning effort was set to low; the GPT-5 request used max_completion_tokens = 6,000 and Gemini used max_tokens = 9,000. The request code does not set temperature. These model aliases are not immutable snapshots and are distinct from the historical GPT-4 generator. The accompanying analysis archive contains the exact system prompt, JSON schema, sample design, request code and per-call requested and returned identifiers.

The schema codes each text for obligations or change commands, suggestions, meaningful options, positive and negative evaluation, reasons for recommendations, and explicit priority. For each of five writing-target domains, labels are absent, mention-only, diagnosis, and diagnosis with action. A diagnosis is a feedback judgment, not proof of correctness. Source-tagged quotations are identified to reduce attribution of learner wording to the provider. Non-absent decisions require an exact substring of at most 35 words; absence requires an empty evidence string. The evidence gate rejects violations, which remain missing rather than being imputed. The accompanying archive provides the verbatim prompt (sources/context_system_prompt.txt), schema (sources/context_schema.json), request construction (code/context_audit.py), sample mapping and saved outputs.

The audit describes model decisions and intermodel consistency; it is not a gold-standard accuracy assessment, human intercoder reliability study or validation of the constructs. Performance in other annotation tasks does not establish validity for LEAF (Gilardi et al., 2023). Independent expert validation remains necessary.

Four deterministically selected paired concordances supplement the census: typical expansion, a strong evaluative contrast, a reversal of the task-reference tendency, and a quotation/lexicon ambiguity. Their pseudonymous locators and short excerpts are used to inspect local context and counterexamples, not to estimate prevalence. These AI-assisted readings do not establish learner intentions, responses, diagnostic truth or co-construction.

LEAF is released under CC BY-NC 4.0 with attribution to its creators and Essay Forum. The accompanying source-data archive contains the original release and license; the analysis archive contains source metadata, lexicons, code, numeric data, audit evidence and frozen results. The corpus includes public source URLs and should not be represented as anonymous.

4. Results

4.1 Corpus and length-aware profile

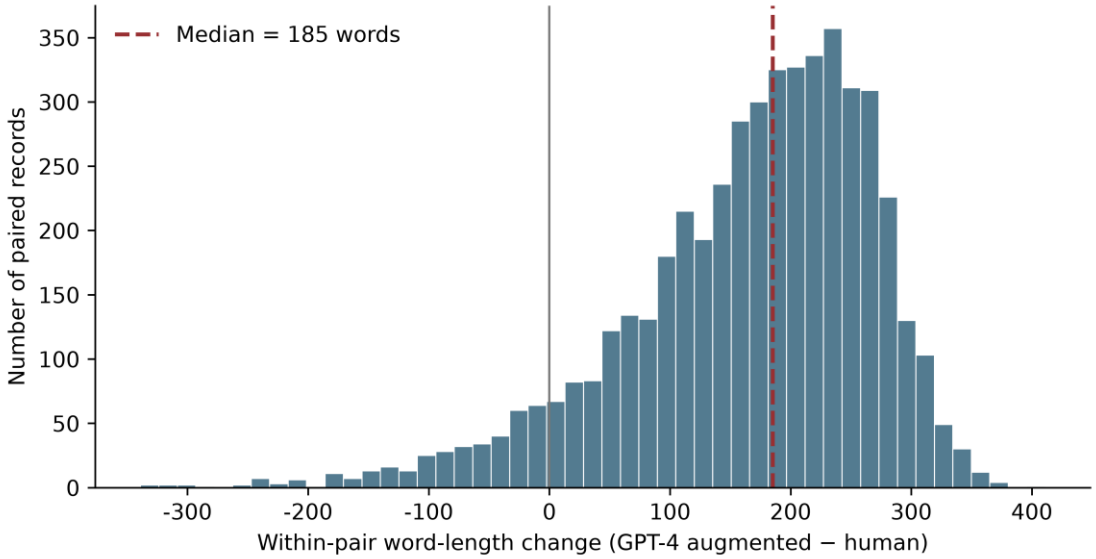
The pinned release contains 4,918 records (4,018 training, 400 development and 500 test) and 4,917 complete pairs; one row has null essay and augmented-feedback fields. No complete pair has a zero-token feedback field. No exact duplicates of normalized source URLs or essay texts were detected, and all main comparative indicators use the same pairs. This does not establish that all related submissions are distinct: a bounded character-5-gram screen produced 578 candidate pairs and identified no non-exact pairs with cosine similarity ≥ 0.95.

Table 1. Common-cohort displayed-text profile

Measure	Human	GPT-4 augmented
Complete feedback pairs	4,917	4,917
Total word tokens	1,154,057	1,966,507
Mean words	234.71	399.94
Median words [Q1, Q3]	212 [159, 287]	400 [373, 427]
Mean TTR	0.550	0.530
Mean MATTR-50	0.776	0.824

Augmented feedback is longer in 92.3% of pairs. The mean within-pair difference is 165.23 words and the median is 185 (Q1 = 110, Q3 = 241); the ratio of mean lengths is 70.4% higher for augmented feedback. Figure 1 shows the paired distribution.

Figure 1. Paired difference in feedback length



Note. The dashed line marks the empirical median; positive values indicate a longer augmented comment for the same record.

Mean TTR is lower for augmented feedback, whereas mean MATTR-50 is higher, with a paired mean increase of 0.047 and a positive change in 91.7% of pairs. All 4,917 pairs meet the two-sided 50-token condition. The opposing measures do not support an inference of greater formulaicity from unadjusted TTR alone; MATTR measures local lexical diversity, not feedback quality or the absence of templates.

4.2 Stance/address lexemes and within-pair variation

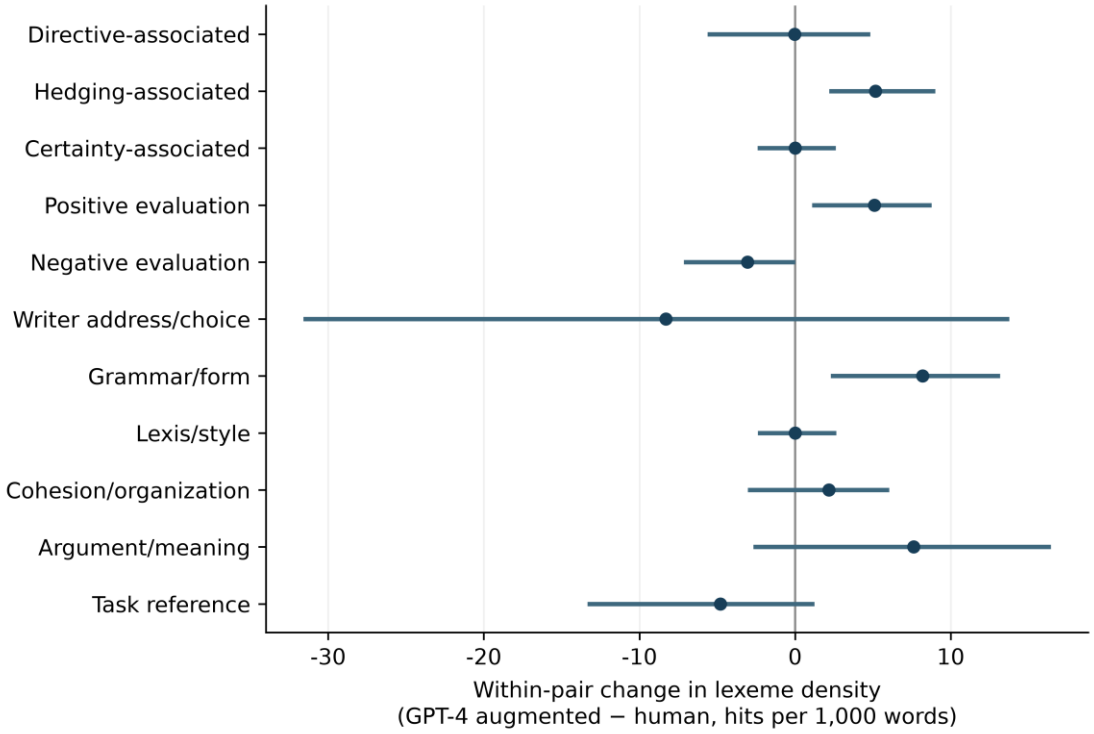
Table 2. Stance/address literal rates and paired differences

Lexicon	Human pooled	Aug. pooled	Mean Δ	Median Δ	Pairs Δ>0 (%)
Directive-associated lexemes	8.85	8.02	-0.76	-0.03	48.5
Hedging-associated lexemes	2.30	7.68	5.42	5.16	82.0
Certainty-associated lexemes	2.86	2.77	-0.03	0.00	44.4
Positive-evaluation lexemes	3.72	8.58	4.78	5.10	79.1
Negative-evaluation lexemes	5.13	1.20	-4.28	-3.05	12.2
Writer-address/choice lexemes	52.72	42.98	-8.15	-8.31	40.1

Note. Rates and Δ are hits per 1,000 words; Δ = augmented minus human. Pooled rates weight tokens, whereas mean and median Δ weight paired records. These are literal lexeme screens, not validated speech acts.

Hedging-associated density is higher in 82.0% of pairs and positive-evaluation density in 79.1%. Negative-evaluation density is lower in 60.1% and tied in 27.7%. Directive-associated differences are mixed (48.5% higher and 50.1% lower). Figure 2 shows the paired lexeme-density differences. These literal rates do not measure actual authority or learner agency.

Figure 2. Paired lexeme-density differences



Note. Points show empirical medians and lines show interquartile ranges across 4,917 matched records, not confidence intervals. The accompanying analysis archive provides the underlying summaries, including minima, maxima, means, quartiles and zero/tie counts.

4.3 Writing-target lexemes: density is not coverage

Table 3. Writing-target literal rates and paired differences

Lexicon	Human pooled	Aug. pooled	Mean Δ	Median Δ	Pairs Δ>0 (%)
Grammar/form lexemes	6.83	13.80	7.10	8.20	80.8
Lexis/style lexemes	4.28	3.90	-0.40	0.00	46.6
Cohesion/organization lexemes	8.38	9.33	1.00	2.17	58.5
Argument/meaning lexemes	21.06	27.91	6.05	7.62	69.7
Task-reference lexemes	15.56	8.93	-7.34	-4.79	29.0

Grammar/form and argument/meaning densities are higher in 80.8% and 69.7% of pairs, respectively. Task-reference density is lower in 67.5% of pairs, although a larger proportion of augmented comments contains at least one task lexeme (91.9% versus 83.5%). Increased literal presence does not establish functional coverage, and lower density does not mean that task advice is absent. Task-based instruction provides a broader pedagogical context (Elmahdi, 2016), but these data do not establish its causal influence on the observed distribution. Case C3 illustrates task-oriented human advice without an exact task-dictionary match.

4.3 Preprocessing and duplication sensitivities

Table 4. Selected pooled rates across text layers (human/augmented)

Lexicon	Displayed	Tagged quotations removed	Conservative quote mask
Hedging-associated lexemes	2.30/7.68	2.29/7.68	2.29/7.82
Positive-evaluation lexemes	3.72/8.58	3.74/8.58	3.75/8.79
Negative-evaluation lexemes	5.13/1.20	5.14/1.20	5.19/1.16
Grammar/form lexemes	6.83/13.80	6.88/13.80	6.95/14.27
Task-reference lexemes	15.56/8.93	15.69/8.93	15.89/9.23

The human field contains 1,444 opening quotation tags, compared with one in the augmented field. Removing explicitly tagged quotations removes 10,007 human-field and 15 augmented-field tokens. The additional conservative double-quotation mask may

remove example rewrites as well and does not create a validated provider-only layer. The directions of the selected pooled contrasts in Table 4 persist across the displayed and quotation-handling layers. Exact-source and exact-essay deduplication did not change the 4,917-pair cohort. Each sensitivity profile contains 4,917 pairs; the conservative MATTR-50 comparison has 4,916 jointly eligible pairs. These checks do not establish generalizability or diagnostic truth.

4.4 Exploratory contextual audit

The 64 sampled rows yielded complete pairs. Of 128 planned model invocations, 127 returned parseable structured decisions; one GPT-5 response was unusable and was not replaced or imputed. The evidence gate excluded 92 of 3,048 returned decisions (3.0%). These are evidence-format failures, not independently identified semantic errors, so valid denominators vary by feature, model and field.

Table 5. Visible feedback moves in the model-coded audit

Feature	GPT-5 H/A	Gemini H/A	Agreement (%)	κ
Obligation/change command	46/63; 49/62	53/64; 63/64	82.4 (125)	0.393
Suggestion	30/63; 59/60	28/64; 64/64	86.2 (123)	0.658
Meaningful option	3/62; 20/63	6/64; 35/64	82.4 (125)	0.538
Positive evaluation	13/63; 61/63	16/64; 63/64	96.8 (126)	0.934
Negative evaluation	62/63; 63/63	63/63; 64/64	99.2 (125)	0.000
Reason for recommendation	49/60; 62/62	43/63; 64/64	93.4 (122)	0.699
Explicit priority	14/63; 7/63	12/64; 11/64	85.7 (126)	0.505

Note. H/A entries give human and augmented presence counts over valid labels, separated by a semicolon. Agreement pools texts with valid labels from both models; its denominator is in parentheses. Cohen's κ is chance-corrected agreement between the two models, not human intercoder reliability or validity. Highly imbalanced category prevalence and marginal distributions can make κ low even when raw agreement is high; interpret κ alongside prevalence and raw agreement.

Both models coded positive evaluation more often in augmented feedback and identified negative evaluation in nearly all valid texts on both sides. Thus, lower negative-word density does not imply that criticism disappeared. For negative evaluation, raw agreement is 99.2% but κ = 0: one model labeled every jointly valid text as present, making κ uninformative in this prevalence-dominated case. The models both identified more options in augmented feedback, but their counts differed substantially. Options and address do not demonstrate exercised agency; obligations and suggestions can coexist. Neither model found explicit priority more often in augmented feedback, and numbering or domain density does not establish a priority hierarchy.

Table 6. Writing-target diagnoses in the model-coded audit

Feature	GPT-5 H/A	Gemini H/A	Agreement (%)	κ
Grammar/form	16/63; 56/60	14/63; 52/55	90.4 (114)	0.849
Lexis/style	22/63; 49/62	19/63; 54/58	83.1 (118)	0.721
Cohesion/organization	49/61; 52/59	43/58; 53/59	72.7 (110)	0.491
Argument/meaning	47/62; 58/58	47/64; 55/55	74.3 (113)	0.552
Task reference	53/58; 50/60	52/56; 47/52	75.8 (99)	0.590

Note. H/A entries are human and augmented counts over valid labels, separated by a semicolon; numerators combine diagnosis and diagnosis with action. Agreement and Cohen's κ use the exact four-level labels, including mention-only, rather than the collapsed counts; the number of jointly valid labels is in parentheses. κ is chance-corrected intermodel agreement, not human reliability or construct validity. Uneven label prevalence and marginal distributions can depress κ despite high raw agreement, so both agreement and the distribution of labels should be considered.

The models frequently coded diagnoses or diagnoses with action for grammar, lexis and argument in augmented comments. Task-directed moves remained common in both fields despite lower task-lexeme density. These are model assignments, not verified diagnoses or validated measures of functional coverage. Joint presentation of paired comments, stylistic cues, quotation metadata and shared model biases limit the audit. The readings below make particular interpretations inspectable but do not convert model agreement into independent human validation. Ethical and cultural considerations also matter when implementing AI-supported feedback (Elmahdi et al., 2025).

4.5 Paired readings

4.6 C1 — Typical word expansion (P-Row-A)

Human feedback: "Don't lose focus of the actual topic for discussion." Augmented feedback: "Each paragraph should start with a clear topic sentence". The human comment identifies form problems and directs the writer to keep the discussion focused on open-plan designs. The augmented feedback adds organized discussion of topic sentences, supporting details and examples.

These excerpts illustrate expansion, but do not establish improved feedback quality or writing. Positive/negative/task-reference rates are 0.00/4.48/4.48 for the human field and 12.25/0.00/4.90 for the augmented field. No human-tagged quotations were reported for this record; the absence of a tag does not prove the provenance of every span.

4.7 C2 — Strong evaluative contrast (P-Row-B)

Human feedback: "The reader will be confused about the presented information". Augmented feedback: "The essay provides a thoughtful examination". The human comment links underdeveloped reasoning to possible reader confusion; the augmented comment opens more positively while elaborating recommendations about the causal argument and its consequences. Positive/negative/task-reference rates are 0.00/47.62/0.00 for the human field and 6.98/0.00/4.65 for the augmented field. A separate quotation is tagged in the human field, so source provenance must be considered when interpreting its lexical counts. These differences describe the record, not learner outcomes.

4.8 C3 — Reversal of the task-rate tendency (P-Row-C)

Human feedback: "You did not give any instructions about how you use your apartment." Augmented feedback: "there's a missed opportunity to give detailed instructions about using the apartment." Both comments address missing practical instructions, although the human field has no exact task-dictionary hit. Task-reference rates are 0.00 for the human field and 25.07 for the augmented field. Positive/negative rates are 3.31/6.62 for the human field and 8.36/0.00 for the augmented field. This case illustrates why literal task-word density cannot be equated with task-directed advice.

4.9 C4 — High known-quotation fraction (P-Row-D)

Human-field tagged learner quotation, not a provider statement: "There are numerous problems behind this phenomenon, and several solutions should be adopted to address those problems." Augmented feedback: "The essay addresses the topic very well". The human feedback discusses generic wording and recommends clearer direct answers and topic sentences; the augmented text also addresses claims and organization. Positive/negative/task-reference rates are 6.13/6.13/18.40 for the human field and 8.95/0.00/6.71 for the augmented field. Forty-eight of the 163 original-tokenized human-field words are explicitly tagged quotations (0.2945). Because this field mixes learner quotation with feedback, its lexical profile cannot be treated as a pure provider statement.

5. Discussion

5.1 Recontextualizing criticism rather than replacing correction

The clearest change is in the presentation of criticism, not a demonstrated move from correction to co-construction. The models coded criticism in almost every valid sampled text, while the census showed more hedging-associated and positive-evaluation language in augmented comments. The selected readings show added explanations and recommendations alongside retained task-related concerns. This pattern is compatible with the interpersonal mitigation of criticism discussed by Hyland and Hyland (2001), but a positive opening or longer explanation does not guarantee effective feedback.

A lower density of negative lexemes does not establish less corrective work or a different learner experience. Because augmentation was conditioned on human feedback and earlier generated material, the contrasts cannot be attributed to GPT-4 as an independent provider; elaboration and rubric-related framing may also contribute.

5.2 Density, lexical presence and metalinguistic work

Task-reference density and presence measure different properties: density falls in most pairs, while a greater proportion of augmented comments contains a task lexeme. Longer texts can dilute a category's rate even when a term is present. In C3, human advice about apartment use is task-oriented despite having no exact task-dictionary hit. The audit and close readings qualify the literal profile without providing a validated coverage measure.

Naming a domain, diagnosing a problem and proposing an action are distinct moves. The audit and selected readings indicate diagnosis-related actions in augmented texts, but whether those diagnoses fit the essays remains untested. Corpus-assisted interpretation can reveal textual patterns without treating keyword counts as semantic attention or instructional priority (Baker et al., 2008).

5.3 Affordances for appraisal, not evidence of agency

An expressed option is a textual affordance; exercised choice requires evidence of a writer's subsequent judgment. The two models identified different numbers of options despite using the same codebook, so their shared direction is not a stable measure of agency. Feedback literacy likewise concerns how learners assess and use information (Carless & Boud, 2018; Boggs & Manchón, 2023). Additional words may offer reasons or alternatives but may also increase processing demands; cognitive load, comprehension and uptake were not assessed.

For practice, the findings motivate questions rather than establish an intervention effect: Does a recommendation address the identified problem? Are its reasons and alternatives clear? What remains at the writer's discretion? Such questions focus attention on the essay and feedback rather than equating fluency, positivity or numbered lists with usefulness. The corpus alone cannot show improved student revision.

5.4 Methodological and external-validity limits

Lower TTR and higher MATTR-50 show why an unadjusted, length-dependent ratio is insufficient evidence of greater formulaicity (Covington & McFall, 2010). Fixed-window diversity does not measure originality, quality or template use. Quotation and suggested rewrites also complicate authorship; conservative masking does not create a validated provider-only layer. The single human contributor, human-conditioned pipeline, model-coded audit and absence of learner-response data limit external and interpretive validity.

6. Conclusion

In this LEAF release, human-informed GPT-4 augmentation changes the length and distribution of stance/address and writing-target lexemes. Lower negative-word density does not mean criticism disappeared, lower task-word density does not establish missing task advice, and reader address does not demonstrate agency. Length-aware measures, sensitivity checks, exploratory model coding and paired readings qualify these textual contrasts. Because the human field comes from one educational consultant and the augmented field draws on human feedback, the findings cannot be generalized to teachers or independent human–AI comparisons. They do not demonstrate improved feedback or learner outcomes. Future work should validate coding with independent human experts and examine writers' responses and revisions.

6.1 Declarations

Data and code availability: The source is the second LEAF release at the pinned repository commit cited above. The accompanying LEAF_Verified_Analysis_Sources_and_Appendices.zip contains the lexicons, analysis code, prompt and schema, audit outputs, model settings, case locators and results. LEAF_Original_Licensed_Source_Data.zip contains the original source file, documentation, and applicable license. The historical audit summaries can be reproduced from saved outputs without new model calls; the archive's README explains the procedure. The dataset is not anonymous, and redistribution must comply with its license.

Ethics and attribution: This is a secondary analysis of publicly released material. Attribution to the corpus creators and Essay Forum and the applicable CC BY-NC 4.0 source terms should be retained.

AI assistance: The submitted manuscript reports AI assistance with source checking, coding, drafting, and source-linked interpretation. The exploratory contextual audit used GPT-5 and Gemini 3.1 Pro Preview.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Afzaal, M., El-Dakhs, D. A. S., Mardini, N., Ambreen, F., & Ismail, H. (2025). A corpus-based analysis of stance markers in upper and lower proficiency level argumentative essays by Saudi learners. *Frontiers in Education*, *10*, 1589004. https://doi.org/10.3389/feduc.2025.1589004

[2] Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, *19*(3), 273–306. https://doi.org/10.1177/0957926508088962

[3] Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

[4] Behzad, S., Kashefi, O., & Somasundaran, S. (2024a). Assessing online writing feedback resources: Generative AI vs. Good Samaritans. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 1638–1644). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.144/

[5] Behzad, S., Kashefi, O., & Somasundaran, S. (2024b). LEAF: Language learners' English essays and feedback corpus. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)* (pp. 433–442). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-short.36

[6] Black, K. E. (2022). Variation in linguistic stance: A person-centered analysis of student writing. *Written Communication*, *39*(4), 531–563. https://doi.org/10.1177/07410883221107884

[7] Boggs, J. A., & Manchón, R. M. (2023). Feedback literacy in writing research and teaching: Advancing L2 WCF research agendas. *Assessing Writing*, *58*, 100786. https://doi.org/10.1016/j.asw.2023.100786

[8] Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

[9] Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

[10] Educational Testing Service. (2024). *LEAF: Language learners' English essays and feedback corpus (Version 2; commit 685ad279d0b42d6c1e2b1b70ac8aae5a6d8fb60f) [Data set]*. GitHub. https://github.com/EducationalTestingService/LEAF/tree/685ad279d0b42d6c1e2b1b70ac8aae5a6d8fb60f

[11] Elkot, M. A., Youssif, E., Hago Elmahdi, O. E., AbdAlgane, M., & Ali, R. (2025). Generative conversational AI: Active practices for fostering students with mild intellectual disabilities to improve English communication skills. *Contemporary Educational Technology*, *17*(1), ep549. https://doi.org/10.30935/cedtech/15688

[12] Elmahdi, O. E. H. (2016). The impact of task-based approach on EFL learner's performance. *World Journal of Educational Research*, *3*(2), 300–320. https://doi.org/10.22158/wjer.v3n2p301

[13] Elmahdi, O. E. H., AbdAlgane, M., & Balla, A. A. S. (2024). AI's role in EFL: Optimizing opportunities while mitigating risks. *International Journal of Linguistics, Literature and Translation*, *7*(10), 21–37. https://doi.org/10.32996/ijllt.2024.7.10.4

[14] Elmahdi, O. E. H., Allehyani, F. S., & AbdAlgane, M. (2025). Exploring the ethical implications and cultural considerations in using AI technologies for EFL education in the Saudi context. *Nsukka Journal of the Humanities*, *33*(2), 1–30. https://www.njh.com.ng/admin/img/paper/NJH-33-2-2025-1-30.pdf

[15] Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

[16] Hyland, F., & Hyland, K. (2001). Sugaring the pill: Praise and criticism in written feedback. *Journal of Second Language Writing*, *10*(3), 185–212. https://doi.org/10.1016/S1060-3743(01)00038-8

[17] Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, *7*(2), 173–192. https://doi.org/10.1177/1461445605050365

[18] Hyland, K., & Hyland, F. (Eds.). (2006a). *Feedback in second language writing: Contexts and issues*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524742

[19] Hyland, K., & Hyland, F. (2006b). Feedback on second language students' writing. *Language Teaching*, *39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

[20] Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, *54*(2), 537–550. https://doi.org/10.1177/00336882231162868

[21] Li, B., Lowell, V. L., Wang, C., & Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education: Examining research on ChatGPT's use in language learning and teaching. *Computers and Education: Artificial Intelligence*, *7*, 100266. https://doi.org/10.1016/j.caeai.2024.100266

[22] Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, *11*, 9. https://doi.org/10.1186/s40561-024-00295-9

[23] Ngo, T. T.-N., Chen, H. H.-J., & Lai, K. K.-W. (2024). The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. *Interactive Learning Environments*, *32*(2), 727–744. https://doi.org/10.1080/10494820.2022.2096642

[24] Sari, E., & Han, T. (2024). The impact of automated writing evaluation on English as a foreign language learners' writing self-efficacy, self-regulation, anxiety, and performance. *Journal of Computer Assisted Learning*, *40*(5), 2065–2080. https://doi.org/10.1111/jcal.13004

[25] Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

[26] 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. https://doi.org/10.1057/s41599-024-03543-y

[27] Yu, S., Geng, F., Liu, C., & Zheng, Y. (2021). What works may hurt: The negative side of feedback in second language writing. *Journal of Second Language Writing*, *54*, 100850. https://doi.org/10.1016/j.jslw.2021.100850

[28] Zhai, Y. (2026). Automated written corrective feedback: Fostering writing accuracy, self-regulation, and motivational dynamics in EFL learners. *BMC Psychology*, *14*, 46. https://doi.org/10.1186/s40359-025-03765-y

[29] Zhang, Y., & Hyland, K. (2022). Responding to supervisory feedback: Mediated positioning in thesis writing. *Written Communication*, *39*(2), 171–199. https://doi.org/10.1177/07410883211069901

[30] Zhang, Z., Aubrey, S., Huang, X., & Chiu, T. K. F. (2025). *The role of generative AI and hybrid feedback in improving L2 writing skills: A comparative study. Innovation in Language Learning and Teaching*. Advance online publication. https://doi.org/10.1080/17501229.2025.2503890

[31] Zhao, Y., Zhang, F., Schunn, C. D., He, P., Li, D., & Zhao, Y. (2024). Feedback, feedback-on-feedback and re-feedback: Effects of written dialogic peer feedback on English as a foreign language writing. *Assessment & Evaluation in Higher Education*, *49*(4), 523–538. https://doi.org/10.1080/02602938.2023.2278017

[32] Zou, S., Guo, K., Wang, J., & Liu, Y. (2026). Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: A comparison study. *Computer Assisted Language Learning*, *39*(4), 1062–1091. https://doi.org/10.1080/09588221.2024.2447279