



Castledown

 OPEN ACCESS

Language Education & Assessment

ISSN 2209-3591

<https://www.castledown.com/journals/lea/>

Language Education & Assessment, 9, 103750 (2026)

<https://doi.org/10.29140/lea.2026.103750>

Evaluating Prompting and Fine-Tuning Strategies for GPT-Based Grammatical Error Correction



JONATHAN PAUL SUMNER^a

THOMAS DILLON^b

^a*Daegu Catholic University, Republic of Korea*
sumnerjonathan@icloud.com

^b*Daegu Catholic University, Republic of Korea*
dillon@cu.ac.kr

Abstract

This study investigates aligning Large Language Models (LLMs) to “minimal-edit” grammatical error correction (GEC) for L2 learners, contrasting with their default “fluency-first” behavior. We used a $2 \times 2 \times 2$ Linear Mixed Models analysis to compare prompting strategies against a task-specific fine-tuned GPT-4o model on 1,000 Korean EFL writing diaries. We operationalized the “minimal-edit” goal using a profile of metrics: $F_{0.5}$ (precision), GLEU (similarity to the original text), and Perplexity (fluency). The LMM revealed fine-tuning was decisively superior. The fine-tuned model (FT-Complex) achieved desirable Perplexity and GLEU scores and the lowest total edit count. In contrast, the base GPT-4o, even with a complex pedagogical prompt, produced disfluent, over-corrected text, driven by a significant Model $\times$ Prompt interaction (high Perplexity). A second one-way LMM on fine-tuning sample sizes ($N = 0$ to 2,617) revealed a trade-off: a small dataset ($N = 500$) achieved the highest minimal-edit adherence (GLEU), while larger datasets ($N = 2,617$) maximized precision ($F_{0.5}$) and fluency. This study demonstrates fine-tuning is an accessible methodology for educators to create AI tools aligned with specific pedagogical goals.

Keywords: Grammatical error correction (GEC), large language models (LLMs), fine-tuning, pedagogical alignment, minimal editing

Copyright: © 2026 Jonathan Paul Sumner & Thomas Dillon. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within this paper.

Introduction

The integration of LLMs into second language (L2) writing support offers immediate feedback. However, the feedback these models provide is a critical unexamined, pedagogical variable. This creates tension within L2 writing pedagogy: should feedback consist of comprehensive “fluency-oriented” rewrites modelling native-like production, or are more targeted and interpretable “minimal edits” preferable? Östling et al. (2023), frame “minimal edits” as isolating specific problems, while “fluency edits” expose learners to more complex, idiomatic structures.

Drawing from Second Language Acquisition (SLA) principles such as “noticing” and “focus on form” (Ellis, 2010; Schmidt, 2012), this study focuses on the pedagogical benefit of minimal edits. Targeted corrections preserving the learner’s original intent are more likely to be “noticed”, thereby facilitating long-term acquisition, whereas wholesale rewrites can obscure the specific error and overwhelm the learner (Ekanayaka & Ellis, 2024).

While generative models excel at “fluency-oriented” rewrites, their alignment with the specific pedagogical requirements of autonomous learners remains unclear. This study addresses this gap by investigating LLM configurations within a Self-Directed Learning (SDL) context. The data consist of writing diaries produced through an online writing workshop. Learners received “minimal-edit” feedback—surgical corrections paired with brief explanations—designed to promote Individual Processing (Yan, 2024). This approach is pedagogically prioritized over holistic rewrites because, in the absence of real-time teacher scaffolding, over-correction can exceed the learner’s Zone of Proximal Development and lead to cognitive overload. Detailed descriptions of the dataset and the specific learning context are provided in Section 3.1.

We operationalize the ‘minimal-edit’ objective through a multi-dimensional metric profile—using the Generalized Language Evaluation Understanding (GLEU) described by Napoles, Sakaguchi, Post, and Tetreault (2016) where higher scores demonstrate more minimal, targeted editing, Perplexity (PPL) to assess linguistic fluency, and $F_{0.5}$ to quantify convergence with an optimized reference correction. Through a $2 \times 2 \times 2$ Linear Mixed Model (LMM) analysis, we provide a comparative framework identifying optimal prompting strategies and data-scaling thresholds required to achieve stable pedagogical alignment.

Literature Review

The Evolution of Grammatical Error Correction

GEC is a subfield of natural language processing concerned with automatically detecting and correcting linguistic errors in learner writing. Historically, GEC research has been dominated by system-oriented paradigms prioritizing accuracy and benchmark performance, often conceptualizing the task as a transformation from non-native to native-like text. Within this framing, progress has been driven largely by advances in modeling architectures and evaluation on shared datasets, rather than by pedagogical considerations.

The field of GEC has evolved significantly from early rule-based systems. A dominant early paradigm was Statistical Machine Translation (SMT), which framed GEC as a “translation” task from an “incorrect” source language (L2 English) to a “correct” target language (native English) (Brockett et al., 2006). This approach was benchmarked in influential shared tasks like the Help Our English (HOO) series (Dale & Kilgariff, 2011), which stimulated broader interest in statistical methods.

This SMT paradigm was later supplanted by Neural Machine Translation (NMT) models (Yuan & Briscoe, 2016) and, more recently, by pre-trained LLMs (Ouyang et al., 2022). These models, trained on vast web-scale data, have demonstrated powerful zero-shot and few-shot GEC capabilities, often achieving state-of-the-art results on standard benchmarks (Bryant, 2023).

Pedagogical GEC: Minimal Edits vs. Fluency Rewrites

A central challenge in GEC is aligning feedback with distinct instructional goals. While learners in product-oriented contexts may benefit from “fluency-oriented” rewrites as target-like models, those in learning-oriented contexts—focusing on iterative practice and reflection—require different constraints. LLMs and tools like Grammarly tend toward comprehensive rewrites (Katinskaia & Yangarber, 2024; Davis et al., 2024; Wu et al., 2023). While this can boost learner confidence (Tungka, 2024), it is often suboptimal. Steiss et al. (2024) found ChatGPT performs significantly worse than human evaluators in prioritization and tone, frequently failing to distinguish high-impact errors from minor stylistic issues.

This lack of prioritization is problematic because feedback must align with individual learner needs (Gozali et al., 2024). Lower-proficiency learners are often confused by vague or overly technical feedback (Shadiev & Feng, 2024). Consequently, LLMs’ default “overcorrection” tendency can erase the learner’s voice, obscure original errors, and introduce linguistic forms beyond the learner’s current proficiency, leading to disengagement (Myers, 2003; Leng, 2014).

Conversely, SLA research provides a theoretical foundation for a “minimal edit” approach. Schmidt’s (2012) Noticing Hypothesis argues conscious attention to linguistic forms is essential for acquisition, and Leki (1991) found students expect errors to be marked and value transparent feedback. Minimal edits facilitate this “noticing” by preserving original sentence structure and highlighting specific errors. LLM generated fluency rewrites, by contrast, function as “implicit recasts” often overlooked by learners (Lyster & Ranta, 1997) whereas explicit feedback leads to self-repair opportunities and therefore uptake. This consensus is supported by Sociocultural Theory, which views feedback as scaffolded mediation within the learner’s Zone of Proximal Development (ZPD) (Vygotsky, 1978; Lantolf & Thorne, 2007). A full rewrite (b) that is too complex sits outside this zone and cannot be internalized, undermining “learner fit” and meaning focus (Chapelle, 2001).

This instructional logic matches research on Dynamic Written Corrective Feedback (DWCF), studying feedback frequency in iterative writing contexts. Hartshorn et al. (2023) demonstrate frequent, low-volume feedback on short writing tasks leads to significant gains in accuracy and fluency. Crucially, these gains emerge not from comprehensive single-draft correction, but from the accumulation of small, interpretable corrections across repeated drafting cycles. DWCF thus provides a strong precedent for minimal editing, especially in autonomous writing contexts where feedback must remain cognitively manageable to support noticing and uptake.

The effectiveness of AI-generated feedback is also dependent on the social context of the learning environment. Yan (2024) showed individual learners (Individual Processing; IP) achieving immediate task improvements through “surgical” interaction with AI, they struggle to internalize long-term rules without external scaffolding. Conversely, the greatest long-term learning gains are achieved when feedback is scaffolded by a teacher (Teacher-Scaffolded; TS) who validates and interprets complex suggestions. Therefore, for the autonomous learner, minimal edits are a structural necessity to ensure feedback remains manageable.

Furthermore, overly critical or opaque feedback can heighten learner anxiety, raising the “affective filter” (Krashen, 1982; Horwitz, 2001; Lin & Lin, 2020). This can undermine the principles of autonomy

and competence central to Self-Determination Theory (Deci & Ryan, 2012; Reinders & White, 2011), whereas minimal edits lower the affective filter counteracting the ‘stylistic over-correction’ of AI that learners find discouraging (Hartshorn, 2008).

Recent learner perception studies confirm this affective component. While learners see ChatGPT as a “companion” (Teng, 2024) and find AI feedback motivating, concerns about overreliance, accuracy, and stylistic over-correction are expressed (Coyne et al., 2023; Slamet, 2024; Poláková & Ivenz, 2024). Ekanayaka and Ellis’s (2024) Engagement with Feedback Model synthesizes this, highlighting cognitive, affective, and behavioral engagement; minimal, interpretable edits are more likely to sustain all three.

However, even when prompted explicitly to minimize edits, LLMs frequently revert to overcorrection (Lin, 2024; Nguyen et al., 2025). This study, therefore, opposes the high-fluency bias of current LLMs and focuses on “minimal edits.” As Staruch et al. (2025) note, the adaptation of LLMs for minimal-edit GEC remains “still underexplored.” This is a critical gap: while many studies evaluate LLMs for fluency-based GEC, few have empirically tested configurations to achieve a pedagogically motivated, minimal style. Our study addresses this gap.

LLM Configuration Strategies

LLM behavior is dynamic; it can be guided through prompting or fundamentally altered through fine-tuning, a central question in recent AI research. Peng (2024) found that prompting is effective, but model tuning (fine-tuning) is superior for complex tasks requiring specialized knowledge, mirroring the hypothesis of this study. To evaluate the extent to which LLM behavior can be steered toward pedagogical goals, this study employs a factorial design testing three distinct prompting configurations and assesses the volume of training data required to fine-tune a model to our minimal edit standard.

Prompt Design and Calibration Techniques

Prompting, which steers a model’s output at inference time, ranges from simple directives to complex, example-based instructions (Li et al., 2023). The “simple instruction” method, uses concise directives without scaffolding. For instance, Katinskaia and Yangarber (2024) found short commands like “Improve grammar in the following sentence” led to better GEC than longer prompts when using a fine-tuned model. Likewise, Loem et al. (2023) demonstrated brief, task-aligned instructions outperforming vague ones, and Davis et al. (2024) validated this, finding the best zero-shot prompt was a clear request: “Reply with a corrected version of the input sentence with all grammatical and spelling errors fixed.”

More “complex instruction” prompts attempt to scaffold the model’s reasoning without examples. This can involve providing explicit rubric rules, as seen in frameworks like AERA, which uses iterative prompting to generate rationales for its decisions (Li et al., 2023). Other methods use structural cues, such as <input> and <output> tags to prevent the model from completing unfinished text (Fang et al., 2023). Scaffolded chain-of-thought (CoT) prompts can instruct the model to analyze an essay in sequential steps (Xia, Mao, & Zheng, 2024) or issue separate, specialized prompts for individual writing traits before aggregating the results (Lee et al., 2024).

The third method, “few-shot” prompting, provides one or more in-context exemplars of the task (Brown et al., 2020). This approach has been shown to yield more accurate and stable evaluations, such as by using sample essays with scored annotations (Li et al., 2023) or examples tagged with Common European Framework of Reference for Languages levels to anchor the model’s scoring

(Yancey et al., 2023). Katinskaia and Yangarber (2024) also found that few-shot CoT prompts consistently outperforming minimal-edit prompts with an “out-of-the-box” model, supporting the general efficacy of example-based prompts (Wei et al., 2022).

For this method, the quality of the examples is critical. Song et al. (2023), for instance, explored few-shot prompting for Grammar Error Explanation (GEE), finding that providing examples of human-curated “atomic” edits—single, isolated corrections—boosted F1 scores dramatically from 0.636 to 0.843. This concept of “atomic” edits directly parallels the output of the ERRANT tool used in our study, suggesting models are highly sensitive to the granularity of the examples they are shown.

However, prompting alone is insufficient for complex tasks. Recent empirical studies on learner corpora showed, even when given an explicit “minimal edit” prompt, LLMs tend to overcorrect and default to fluency rewrites (Lin, 2024; Nguyen et al., 2025). This strongly suggests a model’s inherent training, not its prompt, may be the dominant factor—precisely what our study investigates.

Fine-Tuning and Model Adaptation

An alternative to prompting is fine-tuning, which updates a model’s parameters using a domain-specific, task-specific dataset (Li et al., 2023). The necessity and efficiency of fine-tuning are debated. Mayfield and Black (2020) questioned its cost-effectiveness, while Eguchi and Kyle (2024) advise researchers to test prompting first.

However, the *amount* of data required is a key variable. Eguchi and Kyle (2024) achieved human-comparable accuracy in tagging grammatical features using just 11,856 labeled spans, and Kim and Lu (2024) also achieved high performance after fine-tuning on only 80 documents. This suggests data *quality* and *specificity* may be more important than raw quantity. This unresolved question of precisely *how much* high-quality data is “enough” to achieve pedagogical goals is a primary motivation for the second research question of this study.

This quality-over-quantity argument is supported by Katinskaia and Yangarber (2024), who provided a crucial comparator, finding fine-tuning on a large but *general* GEC corpus (68,000 sentences from FCE/W&I) yielded *lower* performance than zero-shot prompting. This strongly implies fine-tuning on a large, broad-domain corpus is not as effective as using a smaller, highly-targeted dataset, directly reflecting the desired outcome—in our case, minimal edits.

Metrics for Evaluating Pedagogical GEC

Evaluating GEC requires balancing accuracy, fluency, and utility. While human evaluation is the gold standard, it is slow, expensive, and subject to “rater effects” leading to inconsistent scoring (Kim et al., 2019). This creates a need for scalable automatic metrics. The Scribendi Score (Islam & Magnani, 2021), for instance, combines Perplexity with traditional metrics and has been shown to strongly correlate (Pearson $r > 0.8$) with human judgments.

However, the validity of these metrics is a topic of intense academic discussion. A recent study by Goto et al. (2025) warns of a “reliability crisis” for reference-free metrics. They demonstrate the Scribendi score (2020) being “hacked” by adversarial systems, intentionally designed to get high scores. These systems can produce useless outputs (e.g., replacing the input with a single, fluent sentence like “The weather is mild.”) but still receive top scores, effectively undermining the metric’s reliability. This finding, along with critiques from Kobayashi (2024) argues that traditional n-gram metrics are “relatively poor at evaluating fluently corrected sentences with many edits” produced by

modern LLMs. This unresolved debate strongly justifies the need for a *profile* of metrics, as used in this study—combining reference-adherence (GLEU), precision-focused ($F_{0.5}$), and fluency-control (PPL)—rather than relying on a single, flawed score. For a Metric Glossary with pedagogical implications, and examples, see Appendix A.

Edit-Based Metrics (ERRor ANnotation Toolkit and M^2)

To evaluate the precision of corrections, we use edit-based metrics, which compare a model’s proposed edits (insertions, deletions, substitutions) against a gold-standard reference. The foundational metric in this area is the M^2 (MaxMatch) scorer, which calculates the precision, recall, and F-score of edits (Dahlmeier & Ng, 2012).

For this study, we employ the ERRANT toolkit (Bryant, Felice, & Briscoe, 2017), an advanced implementation of this approach. ERRANT’s primary advantage, and the reason it is essential for our analysis, is traceability. Unlike metrics providing a single, opaque score, ERRANT uses linguistic rules to categorize every single edit (e.g., ‘R:NOUN’ for a noun replacement, ‘M:VERB’ for a missing verb).

This traceability has become the standard for diagnosing how GEC models fail. For example, analyses using ERRANT reveal GPT-3.5 struggles with verb errors and prepositions (Katinskaia & Yangarber, 2024) or performs poorly on word order despite fixing spelling (Fang et al., 2023).

The tool is not without its critics. It has known limitations, such as misclassifying ambiguous edits into a generic “OTHER” category (Korre & Pavlopoulos, 2020) or using vague tags (Song et al., 2023). However, the very existence of this academic debate—which is only possible by analyzing ERRANT’s detailed logs—underscores its unique analytical strength. This tool allows us to move beyond if a correction is right and investigate why a model made it.

In this study, we leverage ERRANT quantitatively, to generate the Total Edit Count (as a direct proxy for minimality) and the $F_{0.5}$ score. We selected the $F_{0.5}$ score—a harmonic mean of precision and recall weighting precision twice as heavily ($\beta = 0.5$)—because it is philosophically aligned with our “minimal-edit” goal, as it severely penalizes the “over-correction” (low precision) we aim to avoid. We also use it qualitatively, to provide the rich diagnostic data needed to interpret our statistical findings.

Fluency- and Similarity-Oriented Metrics

We use GLEU (Napoles et al., 2016), which is adapted from the BLEU metric used in machine translation. It works by evaluating n-gram (sequences of words) overlaps between the model’s output and the corrected reference. Unlike BLEU, GLEU is sentence-level and gives credit for n-grams appearing in *either* the original text *or* the reference. For this reason, a small, targeted change preserving most of the original word order gets a higher score than a complete reformulation, making it an excellent proxy for measuring adherence to a minimal-edit style.

Reference-Less Metrics

Traditional metrics depend on gold-standard reference corrections, but newer approaches do not. Perplexity measures how “predictable” or “natural” a sentence is to a pre-trained language model (usually GPT-2). The model assigns a probability to the sentence; a lower PPL indicates the sentence is more fluent and less “surprising.” We use PPL as a standalone control variable to identify overly aggressive “fluency-oriented” models (which may have *low* PPL but *low* GLEU) and to ensure our preferred “minimal-edit” model does not produce awkward, disfluent text. The use of PPL as part of a *profile*

of metrics follows recommendations from scholars who advocate for combining reference-based and reference-less metrics for a more complete evaluation (Choshen & Abend, 2018; Katinskaia & Yangarber, 2024).

A more recent development is the use of LLMs themselves as the analysis tool. This moves beyond *reference-less* metrics (like PPL) and toward *LLM-generated* metrics. This approach is supported by Kobayashi et al. (2024), who demonstrated LLMs (specifically GPT-4) as “state-of-the-art evaluators” for GEC, achieving higher correlation with human judgments than all existing methods, especially when prompted to focus on fluency. Dillon (2025), for example, demonstrated that a general-purpose LLM can be prompted to analyze student chat transcripts and reliably extract custom metrics such as *question-type frequency*, *sentence complexity*, and *thematic vocabulary usage*. This suggests educators are no longer limited to pre-defined metrics like $F_{0.5}$ or GLEU and can instead generate feedback criteria specific to a given task.

Situating the Current Study: Standard vs. Pedagogical Corpora

This study develops a new corpus of Korean EFL learner writing, a methodological choice made to facilitate our goals. Our approach is informed by, and defined in contrast to, the field’s standard GEC benchmarks, which often embody different philosophies:

The FCE (Yannakoudakis et al., 2011) and W&I+LOCNESS (Bryant et al., 2019) corpora: These are the standard, large-scale benchmarks for GEC. They are general-purpose and essential for training broad-domain models. However, they are not specifically curated for a *minimal-edit* philosophy and are the source of the “general” training data (e.g., in Katinskaia & Yangarber, 2024) we critique as being less effective than targeted, fine-tuning.

The JFLEG corpus (Napoles et al., 2017): This corpus was *specifically designed to evaluate and reward fluency-oriented rewrites*, not just discrete error correction.

Therefore, our study’s use of a new, minimal-edit-focused learner corpus, evaluated with precision-focused metrics ($F_{0.5}$) penalizing overcorrection, represents a deliberate methodological departure from the fluency-first approach embodied by datasets like JFLEG.

Research Questions and Operationalization

The literature reviewed above highlights three challenges: LLMs default toward fluency-oriented rewrites even when minimal edits are preferred; prompting alone provides inconsistent model behavior control; and the field lacks empirical evidence regarding the minimal quantity of data needed for effective fine-tuning. Addressing these gaps requires two research questions and a principled operationalization of “minimal-edit” GEC.

In this study, the minimal-edit construct is operationalized as a multidimensional performance profile rather than a single metric. We evaluate model configurations using four complementary lenses capturing precision, conservatism, and the preservation of learner voice. Precision is represented through the ERRANT $F_{0.5}$ score, which utilizes automated M2 references and prefers precision over recall, penalizing more for unnecessary interventions than for missing minor mistakes. Minimality is indexed via GLEU, adapted here as a similarity score between the original and corrected text; this rewards n-gram overlap and localized, surgical edits over extensive rewrites.

Furthermore, perplexity serves as an indicator of fluency overcorrection. By monitoring the marginal effect of the model on PPL, we determine if a configuration respects the baseline linguistic properties

of the learner’s input or imposes artificial fluency through excessive intervention. Finally, ERRANT diagnostic labels provide a qualitative check, categorizing edits into specific linguistic types (e.g., VERB:TENSE, PREP) to ensure the model’s “edit budget” aligns with pedagogical targets. Together, these metrics provide an empirically tractable and theoretically grounded measure of pedagogical fidelity.

On this basis, the study addresses two research questions:

- RQ1:** Which LLM configuration—prompting strategies versus task-specific fine-tuning—better produces a pedagogically sound, minimal-edit style of GEC?
- RQ2:** What is the optimal fine-tuning sample size needed to achieve high-quality GEC aligned with pedagogical goals, and what are the trade-offs between correction quality and data quantity?

These questions motivate the study’s two-part design: a $2 \times 2 \times 2$ factorial experiment comparing prompting and fine-tuning configurations for RQ1, followed by a comparative analysis of fine-tuning sample sizes for RQ2. This structure links the theoretical and motivations of the literature to the following empirical analyses.

Methodology

The present study evaluated two models, GPT-4o and a fine-tuned GPT-4o model (FT), on an essay-correction task. It then evaluated the quantity of data needed for fine tuning.

Dataset and Sampling

Data for this study was collected from 1,156 Korean learners between August 2017 and March 2025 through an online writing workshop, which produced a longitudinal corpus of 48,000 short essays (averaging 60–80 words per entry). The instructional design of this workshop emphasized Dynamic Written Corrective Feedback (Hartshorn, 2023), where instructors provided frequent (maximum 18 per month) “surgical,” minimal-edit corrections. This approach was intended to maximize Individual Processing (Yan, 2024), allowing learners to focus on specific errors in an asynchronous, self-directed learning environment. Learners self-selected topics based on immediate communicative needs, ranging from professional correspondence to personal reflections and commentary on current affairs.

Table 1 presents participant statistics by engagement duration, showing long-term users contributed significantly more essays ($M = 137.8$, $SD = 198.1$) and remained active for longer periods ($M = 18.7$ months, $SD = 17.0$) compared to short- and mid-term users.

Table 1. *Participant statistics*

User Category	N.	Total Essays		Monthly Essays		Usage Time	
		Mean	SD	Mean	SD	Mean	SD
Short-term	71	9.01	5.49	9.01	5.49	1	0
Mid-term	38	30.58	14.07	11.05	3.85	2.76	0.76
Long-term	59	137.8	198.07	8.02	5.33	18.73	16.97

The data in table 1 reflect significant linguistic autonomy and sustained commitment, with 65 users active for over two years and 23 for more than four years. This longitudinal engagement provided a diverse range of interlanguage development and error patterns for the fine-tuning process. To exclude AI-assisted writing, a random sample of 1,000 essays (maximum four per learner) written before November 2022 was selected. Essays containing any Korean text were replaced, the dataset was preprocessed by removing excess whitespace.

Fine-Tuning Procedure

A GPT-4o model was iteratively fine-tuned on batches of human-corrected essays. The first iteration was trained on 759 samples. By July 2024, the dataset had grown to 1,864 essays, and a stronger model was produced; this rose to 2,242 essays by September 2024. By December 2024, the dataset reached 2,617 essays. Annotators manually corrected new texts and then used the model to correct them. After ~2,000 examples, performance plateaued with manual comparison agreement over 90%. It also maintained high generalization across unseen topics because the training data originated from spontaneous writing diaries rather than restricted prompts, so the result is a topic-agnostic model; it effectively applies minimal-edit principles to diverse topics and writing genres.

Prompting the Model

A simple prompt and a complex prompt were designed. The complex prompt was engineered to provide chain-of-thought reasoning and built on rationales for clear, grammar-focused feedback (Leki, 1991; Schmidt, 2012). It instructed the model to: only fix clearly ungrammatical or confusing sentences; avoid altering acceptable but unusual constructions; not replace words with synonyms unless the meaning was incorrect; and bracket optional elements. This prompt elicited targeted, minimal corrections, preserving learner voice. The 1-shot condition included sample text with its corresponding correction. The prompts and 1-shot example can be found in Appendix B.

Experimental Conditions

For RQ1, the 1,000 sample essays were corrected under eight conditions in a 2x2x2 factorial design: Model (4o vs. FT), Prompt (Simple vs. Complex), and Example (None vs. 1-shot). Essays were processed individually via the OpenAI API to prevent serial biases (Zhao et al., 2021).

For RQ2, the 1,000 sample essays were evaluated against six models fine-tuned on incremental data subsets (M1-M6). To standardize the experiment the complex prompt was used for all instructions. 4o-complex-none was reused as the baseline model (M1). Fine-tuned models were trained from scratch on subsets of the 2617 human annotated samples. Fine tuning data was increased in 500 sample increments for these models, M2 (500 samples), M3 (1,000 samples), M4 (1,500 samples), M5 (2,000 samples). Lastly, M6 represents the original model FT Complex (2,617 samples) M6 was the reference standard for $F_{0.5}$ analysis.

Quantitative Metrics and Analysis

Corrections were evaluated using three quantitative metrics generated using python libraries, (example code snippets can be seen in Appendix C):

1. GLEU: Assesses n-gram overlap with the corrected text.
2. $F_{0.5}$ Score (from ERRANT): A precision-weighted F-score penalizing over-correction.
3. Perplexity: Estimated by GPT-2 to gauge fluency.

A $2 \times 2 \times 2$ three-way LMM was conducted for RQ1, and a one-way LMM was conducted for RQ2, using JASP. For the LMMs, models included ‘essay id’ as a random effects grouping factor, and model terms were tested using the Satterthwaite test method.

Qualitative Analysis

To explain the quantitative results, a qualitative analysis was conducted using the ERRANT toolkit. This was particularly crucial for investigating the significant Model * Prompt interaction on perplexity.

First, ERRANT’s edit type counts were generated to compare the frequency of edit categories (e.g., “R:NOUN”, “U:CONJ”, “OTHER”) between the 4o-Complex and FT-Complex models.

Second, ERRANT’s parallel -ex function was used to produce an “atomic” breakdown of individual edits. This output preserved the original text, categorized the edit type, and showed the specific correction span. These atomic outputs for 4o-Complex and FT-Complex were merged into a single comparison spreadsheet, facilitating a direct, side-by-side review of how each model handled the same error, which was essential for diagnosing *why* the 4o-Complex model’s fluency was degraded. A similar qualitative review was conducted to compare the outputs of M2 (the most minimally aligned) and M6 (the most precise) to understand the nature of the trade-off between text similarity (GLEU) and fluency (PPL).

Results

To address the research questions, two separate LMM analyses were conducted. RQ1 was addressed with a $2 \times 2 \times 2$ three-way LMM to analyze the main effects and interactions of *Model* (GPT-4o vs. Fine-tuned), *Prompt* (Simple vs. Complex), and *Example* (None vs. 1-shot). RQ2 was addressed with a one-way LMM to compare the performance of models fine-tuned on different sample sizes (M1-M6).

The dependent variables selected to operationalize the target “minimal-edit” profile were GLEU (similarity between original and corrected essays), $F_{0.5}$ Score (a precision-weighted metric penalizing over-correction), and Perplexity (a control for fluency). Total Edit Counts were also compared as a direct proxy for minimality.

RQ1: Model, Prompt, and Example Configuration

This analysis investigates the main and interaction effects of Model, Prompt, and Example configuration on the performance metrics. Descriptive statistics; Estimated Marginal Means (EMMs) and Standard Errors (SE) from the LMM analysis for all eight experimental conditions are presented in Table 2.

A preliminary inspection reveals two key patterns. First, the FT-Complex-None condition demonstrated the strongest performance on the target “minimal-edit” profile. Second, a notable and counter-intuitive pattern is visible in the Perplexity data. For the base 4o model, switching from a Simple prompt ($M = 21.98$) to the Complex prompt ($M = 29.60$) resulted in a sharp increase in perplexity, indicating a significant loss of fluency. In contrast, the FT model’s fluency remained stable and low. This large discrepancy highlights the Model $\times$ Prompt interaction, analyzed in Section 4.1.1.3.

LMM Main Effects and Interactions

To assess the robustness of the inferential results, model fit was evaluated for each dependent variable using marginal and conditional R^2 (R^2_m and R^2_c). The models for GLEU ($R^2_m = .44$, $R^2_c = .88$) and

Table 2. Descriptive statistics for GLEU, $F_{0.5}$, and PPL by condition

Model	Prompt	Example	GLEU (M, SE)	$F_{0.5}$ (M, SE)	PPL (M, SE)	Total Edits
4o	Simple	None	44.29 (0.389)	0.220 (0.005)	21.98 (0.332)	36,284
		1-shot	53.43 (0.389)	0.302 (0.005)	21.60 (0.332)	30,094
	Complex	None	71.00 (0.389)	0.497 (0.005)	29.60 (0.332)	17,813
		1-shot	70.49 (0.389)	0.529 (0.005)	26.44 (0.332)	18,704
FT	Simple	None	72.22 (0.389)	0.700 (0.005)	24.76 (0.332)	17,760
		1-shot	74.14 (0.389)	0.755 (0.005)	23.46 (0.332)	16,384
FT	Complex	None	75.42 (0.389)	N/A	24.67 (0.332)	15,181
		1-shot	75.72 (0.389)	0.844 (0.005)	24.52 (0.332)	15,045

Note: FT-Complex-None (bolded) served as the reference standard for $F_{0.5}$ calculations. M and SE values are from the LMM EMMs output. Total Edits are from the original analysis.

$F_{0.5}$ ($R^2_m = .75$, $R^2_c = .85$) demonstrated substantial explanatory power, with the experimental factors (Model, Complexity, Exemplar) accounting for a large proportion of the variance in fidelity and edit alignment.

In contrast, the perplexity model showed a high conditional R^2_c (.88) but a low marginal R^2_m (.05). This divergence indicates the overall model is statistically stable, but fluency is driven primarily by the inherent linguistic properties of the original sentences rather than the model configurations. This suggests the models adhered to the minimal-editing constraint, leaving the baseline fluency of the learner's input largely intact.

GLEU (Adherence to Reference)

The LMM analysis of GLEU scores revealed significant main effects for Model, Prompt, and Example. All two-way and three-way interactions were also significant ($p < .001$). The LMM ANOVA Summary is presented in Table 3. The Model factor exerted the strongest influence on reference adherence ($\eta^2 = 0.66$), followed by Prompt complexity ($\eta^2 = 0.57$). Based on the means in Table 2, the fine-tuned (FT) model ($M = 74.37$, $SE = 0.39$) outperformed the base (4o) model ($M = 59.80$, $SE = 0.39$), $F(1, 6993.00) = 13294.7$, $p < .001$.

Table 3. LMM ANOVA summary for GLEU score

Effect	df*	F	p	η^2
Model	1, 6993.00	13294.7	< .001	0.66
Prompt	1, 6993.00	9222.5	< .001	0.57
Example	1, 6993.00	460.1	< .001	0.06
Model * Prompt	1, 6993.00	5944.2	< .001	0.46
Model * Example	1, 6993.00	159.9	< .001	0.02
Prompt * Example	1, 6993.00	497.8	< .001	0.07
Model * Prompt * Example	1, 6993.00	251.6	< .001	0.03

**Note:* Denominator degrees of freedom were calculated using the Satterthwaite approximation. Variations in df reflect adjustments for unequal variance across model conditions to maintain statistical stringency.

F_{0.5} Score (Precision)

The analysis of the $F_{0.5}$ score showed significant main effects for *Model* and *Prompt*, but the main effect for *Example* was not significant ($p = .231$, $\eta^2 = <.01$). All interactions were significant ($p < .001$). Effect sizes demonstrate Model choice was the primary determinant of precision ($\eta^2 = 0.81$), followed by Prompt complexity ($\eta^2 = 0.52$). The LMM ANOVA Summary is presented in Table 4. The F-values highlight *Model* ($F(1, 28995.09) = 6887.19$) as the critical factor, with the FT model (avg. $M = 0.825$) dramatically outperforming the base 4o model (avg. $M = 0.39$).

Table 4. LMM ANOVA summary for $F_{0.5}$ score

Effect	df	F	p	η^2
Model	1, 28995.09	6887.19	< .001	0.81
Prompt	1, 6876.81	7531.74	< .001	0.52
Example	1, 6875.52	1.44	.231	<.01
Model * Prompt	1, 6876.81	127.51	< .001	0.02
Model * Example	1, 6875.51	438.48	< .001	0.06
Prompt * Example	1, 6874.51	644.14	< .001	0.09
Model * Prompt * Example	1, 6874.51	246.82	< .001	0.03

Perplexity (Fluency Control)

The analysis of Perplexity revealed significant main effects for Model, Prompt, and Example, with all interactions also reaching significance ($p < .001$). The LMM ANOVA Summary is presented in Table 5. While the main effect for Prompt complexity was substantial ($\eta^2 = 0.18$), it is best understood through the large Model $\times$ Prompt interaction ($F(1, 6993.00) = 1157.37$, $p < .001$, $\eta^2 = 0.14$). This interaction confirms the pattern observed in Table 2: for the base 4o model, the complex prompt significantly increased perplexity (worse fluency), while for the FT model, fluency remained stable.

Table 5. LMM ANOVA summary for perplexity

Effect	df	F	p	η^2
Model	1, 6993.00	43.01	< .001	0.01
Prompt	1, 6993.00	1581.17	< .001	0.18
Example	1, 6993.00	218.17	< .001	0.03
Model * Prompt	1, 6993.00	1157.37	< .001	0.14
Model * Example	1, 6993.00	38.41	< .001	0.01
Prompt * Example	1, 6993.00	23.36	< .001	<.01
Model * Prompt * Example	1, 6993.00	136.00	< .001	0.02

Qualitative Analysis of PPL Interaction

To understand the counter-intuitive finding of higher perplexity for the 4o-Complex model (Table 2: $M = 29.60$) despite its higher accuracy, a qualitative analysis of ERRANT edit types was conducted.

Table 6. *ERRANT edit type frequencies and paired t-test (4o-complex vs. FT-complex)*

Edit Category	4oC (M)	FTC (M)	N (4oC – FTC)	t	p	d
R:OTHER	1.87	0.68	1187	17.46	<.001	0.55
R:NOUN	1.15	0.31	842	18.57	<.001	0.59
R:ORTH	0.60	0.29	314	10.03	<.001	0.32
U:CONJ	0.17	0.03	143	7.92	<.001	0.25
M:VERB	0.21	0.08	134	6.57	<.001	0.21

Note: R = Replacement, U = Unnecessary, M = Missing. Only top 5 differing categories shown.

Paired T-tests showed five edit types had significant differences. Table 6 compares the edit frequencies between the 4o-Complex (4oC) and FT-Complex (FTC) models.

Table 6 indicates the 4o-Complex model made significantly more edits, particularly in the “Replacement:Other” (R:OTHER) and “Replacement:Noun” (R:NOUN) categories. Qualitative inspection showed the 4o-Complex model’s higher perplexity was driven by ambiguous and unnecessary “R:OTHER” edits, often resulting from the over-application of bracketed insertions. These intrusive, stylistic edits reduced n-gram similarity (lowering GLEU) and introduced unnatural phrasing, increasing perplexity. Example edits appear in Appendix D.

RQ2: Fine-Tuning Sample Size

This analysis uses a one-way LMM to compare the performance of models fine-tuned on increasing dataset sizes (M1 to M6). M1 is the baseline (GPT-4o Complex prompt, 0 samples). Descriptive statistics for the models are in Table 7.

Descriptive statistics in Table 7 show a clear pattern of diminishing returns. The largest performance jump occurs between M1 (baseline) and M2 (500 samples). M2 achieved the highest mean GLEU score (M = 77.21), while M6 achieved the lowest PPL (M = 24.67).

Table 7. *Descriptive statistics for M1-M6 (GLEU, $F_{0.5}$, and PPL)*

Model	N Samples	GLEU (M, SE)	$F_{0.5}$ (M, SE)	PPL (M, SE)	Total Edits
M1	0	71.00 (0.388)	0.493 (0.005)	29.60 (0.375)	17,813
M2	500	77.21 (0.388)	0.719 (0.005)	25.49 (0.375)	13,791
M3	1,000	76.23 (0.388)	0.732 (0.005)	25.13 (0.375)	14,385
M4	1,500	76.44 (0.388)	0.750 (0.005)	25.05 (0.375)	14,418
M5	2,000	75.36 (0.388)	0.741 (0.005)	24.72 (0.375)	15,066
M6	2,617	75.42 (0.388)	N/A	24.67 (0.375)	15,181

Note: M and SE values are from the LMM EMMs output. Total Edits are from the original analysis. M6 (FT-Complex-None) served as the reference standard.

GLEU

The LMM revealed a significant effect of sample size on GLEU scores ($F(5, 4995.00) = 358.9$, $p < .001$), with the model factor explaining a significant portion of the variance ($\eta^2 = 0.26$). The model

Table 8. *EMMs for GLEU*

Model	Estimate	SE	95% CI Lower	95% CI Upper	p
M1	71.00	0.388	70.24	71.76	< .001
M2	77.21	0.388	76.45	77.97	< .001
M3	76.23	0.388	75.47	76.99	< .001
M4	76.44	0.388	75.68	77.20	< .001
M5	75.36	0.388	74.60	76.12	< .001
M6	75.42	0.388	74.66	76.18	< .001

achieved high overall reliability ($R^2_c = 0.912$), although the fixed effect of sample size accounted for a smaller proportion of total variance ($R^2_m = 0.026$). Table 8 shows performance peaked at M2 and plateaued thereafter.

$F_{0.5}$ Score

A linear mixed-effects model revealed a significant effect of sample size on $F_{0.5}$ scores, $F(5, 4915.96) = 1789.0$, $p < .001$, with sample size explaining a substantial proportion of variance ($\eta^2 = 0.65$). The full model accounted for a large share of the total variance ($R^2_c = 0.673$), while the fixed effect of sample size alone explained nearly half of the variance ($R^2_m = 0.496$). In Table 9, precision increased monotonically with additional training data.

Table 9. *EMMs for $F_{0.5}$ score*

Model	Estimate	SE	95% CI Lower	95% CI Upper	p
M1	0.493	0.005	0.484	0.502	< .001
M2	0.719	0.005	0.710	0.728	< .001
M3	0.732	0.005	0.722	0.741	< .001
M4	0.750	0.005	0.741	0.760	< .001
M5	0.741	0.005	0.732	0.750	< .001
M6	N/A	N/A	N/A	N/A	N/A

Note: M6 is the reference model.

Perplexity

A significant effect of sample size on perplexity, $F(5, 4995.00) = 336.8$, $p < .001$, was found with sample size accounting for a meaningful proportion of variance ($\eta^2 = 0.25$). The full model explained a large share of the total variance ($R^2_c = 0.926$), whereas the fixed effect of sample size alone accounted for a small proportion ($R^2_m = 0.021$). In Table 10, M1 exhibits the highest perplexity by a substantial margin. Perplexity decreased sharply from M1 to M2 and M3, plateauing afterwards, with models M3 through M6 showing similarly low perplexity scores.

Table 10. *EMMs for perplexity*

Model	Estimate	SE	95% CI Lower	95% CI Upper	p
M1	29.60	0.375	28.87	30.34	< .001
M2	25.49	0.375	24.76	26.23	< .001
M3	25.13	0.375	24.39	25.86	< .001
M4	25.05	0.375	24.31	25.78	< .001
M5	24.72	0.375	23.99	25.46	< .001
M6	24.67	0.375	23.93	25.40	< .001

Summary of RQ2 Trade-Off

The LMM results show a clear trade-off between the metrics. M2 (500 samples) achieved the highest GLEU score ($M = 77.21$), suggesting it was the most aligned with the strict minimal-edit reference. Conversely, M6 (2,617 samples) achieved the best $F_{0.5}$ score ($M = 1.000$) and the best perplexity score ($M = 24.67$) but had a lower GLEU score than M2. This suggests M6 adopted a more comprehensive, fluency-oriented profile diverging slightly from the strict minimal-edit reference but achieved higher precision and fluency. However, the performance gap between M2 and M5 was minimal ($F_{0.5}$ difference approx 0.07; GLEU difference approx 1.85 points; negligible difference in perplexity).

A qualitative comparison of M2 and M6 identified 432 cases where M2 achieved higher GLEU scores but exhibited higher perplexity, a pattern indicative of under-correction. For example, in the sentence ‘*approval was complete a couple of weeks ago*,’ M2 made verb tense adjustment but left ‘*was complete*’ unchanged, while M6 correctly used ‘*was completed*.’ Similarly, M2 failed to correct the article error in ‘*a example photograph*,’ whereas M6 revised it to ‘*an example photograph*.’ These instances demonstrate M2’s high reference alignment occasionally allowing errors to persist.

Discussion

This study’s primary objective was to identify the LLM configuration producing minimal-edit GEC. By operationalizing this goal using a suite of metrics ($F_{0.5}$ for precision, Total Edits for minimality, GLEU for text similarity, and PPL for fluency), the $2 \times 2 \times 2$ LMM results for RQ1 provide an answer.

RQ1: Fine-Tuning Outperforms Prompting

The results demonstrated task-specific fine-tuning as decisively more effective than prompt engineering for overcoming the “fluency bias” inherent in LLMs. While the base model consistently over-edited—prioritizing stylistic polishing over the requested minimal corrections—fine-tuning shifted the model’s output distribution towards minimal corrective edits, internalizing the minimalism constraint rather than relying on prompt-level activation.

Prompt-only approaches failed most notably as fluency degraded when complex instructions were introduced. This pattern indicates a trade-off effect: as instruction complexity increases, the model’s ability to maintain both strict instruction adherence and output fluency decreases. Fine-tuning resolves this tension by baking task requirements into the model’s weights, ensuring specialized behavior—such as the specific use of annotation brackets—does not come at the cost of fluency. Furthermore, the diminishing returns observed when adding exemplars to fine-tuned configurations suggests increased instructional consistency and decreased output variance rendering manual examples redundant.

RQ2: The “Sweet Spot” of Fine-Tuning

Investigation into sample sizes (Table 7) reveals a pragmatic “sweet spot”. Finding a minimal dataset ($N = 500$) provides substantial improvement in model behavior confirming data quality and target-alignment are more critical than corpus volume for specialized GEC. This aligns with Eguchi and Kyle (2024), who achieved strong performance with a modest number of labeled spans, but contrasts with Katinskaia and Yangarber (2024), who failed to achieve comparable results despite using a significantly larger but less targeted corpus. These results demonstrate high-quality data making high-precision, fine-tuning feasible even in low-resource educational contexts.

Operationalizing Pedagogical Goals: A Generalizable Framework

The finding that task-specific fine-tuning ($N = 500$) decisively outperformed complex prompting is significant not just for GEC, but as a proposed method for aligning AI tools with any L2 teaching objective. This methodology validates fine-tuning as an accessible and robust strategy for educators, because it solves the problem of instability across runs, “stochastic wobble” plaguing LLM interactions. While complex prompting can successfully operationalize analysis goals, the resulting model behavior is unstable, requiring intensive session management to prevent model drift and maintain consistency across inputs. Fine-tuning, conversely, locks in the desired behavior, providing a reliable foundation for tool creation.

This methodology is transferable because any abstract goal can be translated into a measurable output. Beyond GEC, this framework accommodates goals such as operationalizing vocabulary acquisition into a “count of thematic set items” or translating interactional competence into “follow-on question counts” and “diversity of question types,” as demonstrated in Dillon (2025). Once the output is operationalized, with a corresponding metric profile, educators can leverage the LLM-as-Evaluator hypothesis validated in this study to test these custom metrics, establishing qualitative analytic criteria specific to task rubrics.

Crucially, the iterative fine-tuning process is highly accessible, validating the use of a small, high-quality dataset ($N = 500$). Teachers can efficiently curate this data by first using a prompted LLM to generate initial corrected or analyzed outputs. Incorporating human post-editing into fine-tuning cycles is highly efficient as LLMs can provide rationales for their coding or correction decisions. Human-in-the-loop validation therefore becomes a rapid process of correcting errors rather than annotating manually (Östling et al. 2023; Dillon, 2025.) This digital literacy makes the creation of a stable, pedagogically aligned model feasible for individual teachers or small departments, with educators becoming system designers capable of customizing tools for class-level aggregation and analytics.

Instructional design implications

The stability gained through fine-tuning also opens new avenues for instructional design. Reliable, minimal-edit feedback enables iterative drafting cycles and supports flipped or self-paced grammar instruction, where students focus class time on meaning and revision strategies rather than surface correctness. Moreover, fine-grained edit classifications can help teachers highlight recurrent error types and direct learners’ attention to targeted before-and-after examples, supporting metalinguistic “noticing” (Schmidt, 2012) and uptake (Lyster & Ranta, 1997).

Because fine-tuned models behave conservatively, they are particularly well-suited for novice and intermediate learners who benefit from minimal, interpretable corrections. This alignment is critical for self-directed study; as Yan (2024) demonstrates, solo learners lack the interpretive support required to navigate complex, fluency-oriented rewrites without cognitive overload.

In teacher-scaffolded (TS) contexts, learners can use a standard LLM to intentionally generate fluency-oriented feedback. The instructor can bridge the gap between the AI's complex suggestions and the learner's current proficiency, transforming native-like rewrites into durable long-term learning gains.

Limitations and Future Directions

This study successfully used a factorial analysis to validate a minimal-edit GEC configuration. The use of $F_{0.5}$, Total Edits, Perplexity, and GLEU in tandem provided a multi-dimensional profile of “minimality,” directly linking the study's theoretical framework to its empirical results.

This study's primary limitation is its reliance on automatic metrics. While we used a profile of metrics, as recommended by Choshen and Abend (2018) and Katinskaia and Yangerber (2024), and triangulated with qualitative data, these metrics do not measure learner uptake or perceived usefulness. However, evaluation metrics such as GLEU, PPL, or ERRANT edit counts may be translated into simple, interpretable indicators (e.g., percentage improvement across drafts), and become accessible as student-facing feedback tools.

However, this methodological validation is the precursor to a larger, longitudinal study addressing this gap by applying our validated “metric profile” to the entire 48,000-essay corpus. Using a Generalized Additive Mixed Model (GAMM), this future research will track the grammatical development of the 1,156 learners over time. The goal is to determine if their development, as measured by our metric profile, follows the conventional SLA trajectory of steep initial improvement followed by a plateau. This will also allow for a comparison of the developmental paths of the short-, mid-, and long-term participants identified in Table 1, providing empirical data on how grammatical accuracy and fluency evolve with sustained engagement.

Beyond the immediate findings, this study provides preliminary evidence for the LLM-as-evaluator hypothesis (Dillon, 2025). As a secondary methodological test, we investigated whether GPT-4o could replicate our primary Python-based metrics. When provided with the 1,000 original and corrected essays, the model's internal GLEU calculations achieved a near-perfect Pearson correlation ($r = 0.99$) with our standard programmatic analysis. This capability suggests that future work could successfully transition from specialized software to LLM-driven diagnostics. Specifically, comparing the qualitative output of ERRANT against prompted LLMs could determine if models can reliably generate “atomic edits” (Song et al., 2023) at scale. If so, this would further empower educators to develop custom pedagogical analytical tools without requiring specialized programming expertise.

Conclusion

This study sought to identify an effective and accessible method for aligning an LLM with a specific objective—in this case, minimal-edit GEC. Using a $2 \times 2 \times 2$ LMM design and a suite of empirically grounded metrics, we showed task-specific fine-tuning consistently outperforming prompt engineering for achieving high-precision, pedagogically aligned behavior.

Fine-tuned models demonstrated superior performance across all target metrics. Even a small, high-quality dataset ($N = 500$) produced dramatic gains, validating the feasibility of fine-tuning for educational contexts with limited annotated data. The comparison across fine-tuning sample sizes revealed a practical continuum: smaller datasets yield more conservative, reference-aligned corrections, while larger datasets improve fluency and coverage. This offers educators meaningful flexibility depending on learner proficiency and instructional goals.

Beyond the specific GEC task, the broader contribution of this work is methodological. This study demonstrates pedagogical goals being operationalized into measurable outputs with LLM behavior reliably shaped through targeted data rather than increasingly complex prompts. Because LLMs can also function as evaluators—replicating metrics such as GLEU with high fidelity—educators can create custom analytic tools aligned with their own rubrics and learning outcomes.

As generative AI becomes further integrated into education, this approach empowers teachers to act as system designers who set the parameters for how LLMs support instruction. Fine-tuned models offer stability and alignment for large-scale feedback, while prompted, out-of-the-box models remain useful for advanced learners seeking fluency-oriented revision. With these complementary tools, instructors can match model behavior to different needs and learner levels.

This study offers a data-driven, replicable framework for aligning LLMs with pedagogical intent. By combining targeted fine-tuning with LLM-driven analytics, this methodology provides a foundational pathway for developing learner-centered AI tools. While the current evidence is situated within the context of GEC, the findings suggest a scalable approach that could be extended to a variety of L2 classroom tasks, offering a precedent for future research into broader pedagogical applications.

Use of Generative AI

This manuscript made use of Google Gemini to assist in paraphrasing and improving the clarity of text. No AI tools were used for the generation of research hypotheses or the writing of conclusions. The final manuscript was authored and reviewed by the listed authors without additional AI input.

Funding

This work was supported by the Daegu Catholic University Research Fund.

References

- Brockett, C., Dolan, B., & Gamon, M. (2006). Correcting ESL errors: Aligning an SMT system with the problem. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 865–872). Association for Computational Linguistics. <https://aclanthology.org/P06-1032.pdf>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* 33 (pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6b-fcb4967418bfb8ac142f64a-Paper.pdf
- Bryant, C. (2023). Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3), 643–701. <https://direct.mit.edu/coli/article/49/3/643/115846>
- Bryant, C., Felice, M., & Briscoe, E. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 793–805). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P17-1074>
- Bryant, C., Ng, H. T., Ranta, S., & Hann, N. A. (2019). The BEA-2019 shared task on grammatical error correction. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 52–75). Association for Computational Linguistics. <https://aclanthology.org/W19-4406/>

- Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing and research*. Cambridge University Press. <https://www.cambridge.org/core/books/computer-applications-in-second-language-acquisition/9546AF8B1D0B3E89B2B639B92949B310>
- Choshen, L., & Abend, O. (2018). Automatic metric validation for grammatical error correction. arXiv. <https://arxiv.org/abs/1804.11225>
- Coyne, S., Sakaguchi, K., Galvan-Sosa, D., Zock, M., & Inui, K. (2023). *Analyzing the performance of GPT-3.5 and GPT-4 in grammatical error correction*. arXiv. <https://arxiv.org/abs/2303.14342>
- Dahlmeier, D., & Ng, H. T. (2012). Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 568–572). Association for Computational Linguistics. <https://aclanthology.org/N12-1067.pdf>
- Dale, R., & Kilgariff, A. (2011). Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation* (pp. 242–249). Association for Computational Linguistics. <https://aclanthology.org/W11-2838.pdf>
- Davis, C., Caines, A., Andersen, Ø., Taslimipour, S., Yannakoudakis, H., Yuan, Z., & Buttery, P. (2024). *Prompting open-source and commercial language models for grammatical error correction of English learner text*. arXiv. <https://arxiv.org/abs/2401.07702>
- Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (Vol. 1, pp. 416–436). SAGE. <https://doi.org/10.4135/9781446249215.n21>
- Dillon, T. (2025). Automated Analysis of ESL Interaction Tasks Using ChatGPT. *Korean Journal of English Language and Linguistics*, 25(0), 955–981. <https://doi.org/10.15738/kjell.25..202507.955>
- Eguchi, M., & Kyle, K. (2024). Building custom NLP tools to annotate discourse-functional features for second language writing research: A tutorial. *Research Methods in Applied Linguistics*, 3(3), 100153. <https://doi.org/10.1016/j.rmal.2024.100153>
- Ekanayaka, W. I., & Ellis, R. (2024). Engaging with written corrective feedback. *Instructed Second Language Acquisition*, 8(1), 41–63. <https://doi.org/10.1558/isla.26982>
- Ellis, R. (2010). A framework for investigating oral and written corrective feedback. *Studies in Second Language Acquisition*, 32(2), 337–349. <https://www.cambridge.org/core/journals/studies-in-second-language-acquisition/article/abs/epilogue/22F1837189E1965228E7CAA D904B3E16>
- Fang, T., Yang, S., Lan, K., Wong, D. F., Hu, J., Chao, L. S., & Zhang, Y. (2023). *Is ChatGPT a highly fluent grammatical error correction system? A comprehensive evaluation*. arXiv. <https://arxiv.org/abs/2304.01746>
- Ferris, D. R. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81–104). Cambridge University Press. <https://doi.org/10.1017/CBO9781139524742>
- Goto, T., Sakai, Y., & Watanabe, T. (2025) Reliability crisis of reference-free metrics for grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 24913–24926). <https://aclanthology.org/anthology-files/anthology-files/pdf/findings/2025.findings-emnlp.1356.pdf>
- Gozali, I., Wijaya, A. R. T., Lie, A., Cahyono, B. Y., & Suryati, N. (2024). Leveraging the potential of ChatGPT as an automated writing evaluation (AWE) tool: Students' feedback literacy development and AWE tools integration framework. *The JALT CALL Journal*, 20(1), 1–22. <https://www.castledown.com.au/journals/jaltcall/article/download/1200/233>

- Hartshorn, K. J. (2008). *The effects of manageable corrective feedback on ESL writing accuracy* [Doctoral dissertation, Brigham Young University]. BYU ScholarsArchive. <https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=2521&context=etd>
- Hartshorn, K. J., Rice, S. H., Eckstein, G., & Evans, N. W. (2023). Dynamic written corrective feedback frequency and its effects on ESL writing fluency, accuracy, and complexity. *Foreign Language Studies*, 1(2). <https://api.eurokd.com/Uploads/Article/737/frsl.2023.01.02.pdf>
- Horwitz, E. K. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics*, 21, 112–126. <https://doi.org/10.1017/s0267190501000071>
- Islam, M. A., & Magnani, E. (2021). Is this the end of the gold standard? A straightforward reference-less grammatical error correction metric. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 3009–3015). Association for Computational Linguistics. <https://aclanthology.org/2021.emnlp-main.239.pdf>
- Katinskaia, A., & Yangarber, R. (2024). GPT-3.5 for Grammatical Error Correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. (pp. 7831–7843). ELRA and ICCL. <https://aclanthology.org/2024.lrec-main.692>
- Kim, H. J., Lee, J., & You, H. J. (2019). Analysis of rater effect in the evaluation of second language grammatical knowledge in the context of writing: Application of a generalized linear model. *Linguistic Research*, 36(Special Edition), 49–73. http://isli.khu.ac.kr/journal/content/data/36_S/2.pdf
- Kim, M., & Lu, X. (2024). Exploring the potential of using ChatGPT for rhetorical move-step analysis: The impact of prompt refinement, few-shot learning, and fine-tuning. *Journal of English for Academic Purposes*, 71, 101422. <https://doi.org/10.1016/j.jeap.2024.101422>
- Kobayashi, M., Mita, M., & Komachi, M. (2024). Large language models are state-of-the-art evaluator for grammatical error correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp.68–79). Association for Computational Linguistics. <https://aclanthology.org/2024.bea-1.6/>
- Korre, K., & Pavlopoulos, J. (2020, December). Errant: Assessing and improving grammatical error type classification. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 85–89). Association for Computational Linguistics. <https://aclanthology.org/2020.latechclfl-1.10.pdf>
- Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon. http://www.sdkrashen.com/content/books/principles_and_practice.pdf
- Lantolf, J. P. & Thorne, S. L. Sociocultural theory and the genesis of second language development. Oxford University Press, 2006. *Applied Linguistics*, 28(3), September 2007, (pp. 477–480). <https://doi.org/10.1093/applin/amm027>
- Lee, S., Cai, Y., Meng, D., Wang, Z., & Wu, Y. (2024). *Unleashing large language models' proficiency in zero-shot essay scoring*. arXiv. <https://arxiv.org/abs/2404.04941>
- Leki, I. (1991). The preferences of ESL students for error correction in college-level writing classes. *Foreign Language Annals*, 24(3), 203–218. <https://doi.org/10.1111/j.1944-9720.1991.tb00464.x>
- Leng, K. T. P. (2014). An analysis of written feedback on ESL students' writing. *Procedia – Social and Behavioral Sciences*, 123, 389–397. https://www.sciencedirect.com/science/article/pii/S187704281401475X/pdf?md5=077413605e39cba70dce32d4684acb6d&pid=1-s2.0-S187704281401475X-main.pdf&_valck=1
- Li, J., Gui, L., Zhou, Y., West, D., Aloisi, C., & He, Y. (2023). *Distilling ChatGPT for explainable automated student answer assessment*. arXiv. <https://arxiv.org/abs/2305.12962>
- Lin, S. (2024). Evaluating LLMs' grammatical error correction performance in learner Chinese. *PLOS ONE*, 19(10), e0312881. <https://doi.org/10.1371/journal.pone.0312881>

- Lin, Y., & Lin, Y. (2020). The enlightenment of affective filter hypothesis and risk-taking on English learning. *Studies in Literature and Language*, 20(2), 51–57. <http://cscanada.net/index.php/sll/article/view/11629>
- Loem, M., Kaneko, M., Takase, S., & Okazaki, N. (2023). *Exploring effectiveness of GPT-3 in grammatical error correction: Performance and controllability in prompt-based methods*. arXiv. <https://arxiv.org/abs/2305.18156>
- Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19(1), 37–66. <https://escholarship.mcgill.ca/downloads/0p096b851>
- Mayfield, E., & Black, A. W. (2020). Should you fine-tune BERT for automated essay scoring? In *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 151–162). Association for Computational Linguistics. <https://aclanthology.org/2020.bea-1.15/>
- Myers, S. A. (2003). Reassessing the “proofreading trap”: ESL tutoring and writing instruction. *The Writing Center Journal*, 24(1), 51–70. <https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1544&context=wcj>
- Napoles, C., Sakaguchi, K., Post, M., & Tetreault, J. (2016). *GLEU without tuning*. arXiv preprint arXiv:1605.02592. <https://arxiv.org/pdf/1605.02592>
- Napoles, C., Sakaguchi, K., & Tetreault, J. (2017). JFLEG: A fluency-extended GUG corpus for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 229–234). Association for Computational Linguistics. <https://doi.org/10.18653/v1/E17-2037>
- Nguyen, P. T., Nuss, B., Dressler, R., & Ovens, K. (2025). A small-scale evaluation of large language models used for grammatical error correction in a German children’s literature corpus: A comparative study. *Applied Sciences*, 15(5), 2076–3417 <https://www.mdpi.com/2076-3417/15/5/2476>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., ... & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv. <https://arxiv.org/abs/2203.02155>
- Östling, R., Gillholm, K., Kurfalı, M., Mattson, M., & Wirén, M. (2023). *Evaluation of really good grammatical error correction*. arXiv. <https://arxiv.org/abs/2308.08982>
- Peng, C., Yang, X. I., Smith, K. E., Yu, Z., Chen, A., Bian, J., & Wu, Y. (2024). Model tuning or prompt tuning? A study of large language models for clinical concept and relation extraction. *Journal of Biomedical Informatics*, 153, 104630. <https://doi.org/10.1016/j.jbi.2024.104630>
- Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students’ writing skills. *Cogent Education*, 11(1), 2410101. <https://doi.org/10.1080/2331186x.2024.2410101>
- Reinders, H., & White, C. (2011). Special Issue Commentary: Learner autonomy and new learning environments. *Language Learning & Technology*, 15(3), 1–7. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44254/1/15_03_commentary.pdf
- Schmidt, R. (2012). Attention, awareness, and individual differences in language learning. In W. M. Robinson (Ed.), *Perspectives on individual characteristics and foreign language education* (pp. 27–50). De Gruyter Mouton. <https://nflrc.hawaii.edu/pdfs/schmidt%20attention,%20awareness,%20and%20individual%20differences.pdf>
- Scribendi Inc. (2020). *Comparing BERT and GPT-2 as language models to score the grammatical correctness of a sentence*. Scribendi.ai. <https://www.scribendi.ai/comparing-bert-and-gpt-2-as-language-models-to-score-the-grammatical-correctness-of-a-sentence/>
- Shadiev, R., & Feng, Y. (2024). Using automated corrective feedback tools in language learning: A review study. *Interactive Learning Environments*, 32(6), 2538–2566. <https://eric.ed.gov/?id=EJ1439559>

- Slamet, J. (2024). Potential of ChatGPT as a digital language learning assistant: EFL teachers' and students' perceptions. *Discover Artificial Intelligence*, 4(1), 46. <https://link.springer.com/content/pdf/10.1007/s44163-024-00143-2.pdf>
- Song, Y., Krishna, K., Bhatt, R., Gimpel, K., & Iyyer, M. (2023). *Gee! Grammar error explanation with large language models*. arXiv. <https://arxiv.org/abs/2311.09517>
- Staruch, R., Graliński, F., & Dzienisiewicz, D. (2024). *Adapting LLMs for minimal-edit grammatical error correction*. arXiv. <https://arxiv.org/pdf/2506.13148>
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., ... & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>
- Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence*, 7, 100270. <https://doi.org/10.1016/j.caeai.2024.100270>
- Tungka, N. F. (2024). Enhancing writing skills of EFL learners through automated feedback: An empirical investigation. *Jurnal Ilmiah Spectral*, 10(1), 028–042. https://www.researchgate.net/publication/378919898_enhancing_writing_skills_of_efl_learners_through_automated_feedback_an_empirical_investigation
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press. <https://www.hup.harvard.edu/books/9780674576292>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 35 (pp. 24824–24837). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
- Wu, H., Wang, W., Wan, Y., Jiao, W., & Lyu, M. (2023). ChatGPT or Grammarly? *Evaluating ChatGPT on grammatical error correction benchmark*. arXiv. <https://arxiv.org/abs/2303.13648>
- Xia, W., Mao, S., & Zheng, C. (2024). *Empirical study of large language models as automated essay scoring tools in English composition: Taking TOEFL Independent Writing Task for example*. arXiv. <https://arxiv.org/abs/2401.03401>
- Yan, D. (2024). Collaborative processing of ChatGPT-generated feedback: Effects on L2 writing task improvement and learning. *Language Learning & Technology*, 28(1), 1–19. <https://doi.org/10.64152/10125/73597>
- Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 576–584). Association for Computational Linguistics. <https://aclanthology.org/2023.bea-1.49>
- Yannakoudakis, H., Briscoe, E., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 180–189). Association for Computational Linguistics. <https://aclanthology.org/P11-1019/>
- Yuan, Z., & Briscoe, T. (2016). Grammatical error correction using neural machine translation. In *Proceedings of the 2016 conference of the North American Chapter of the Association for computational linguistics: Human language technologies* (pp. 380–386). <https://aclanthology.org/N16-1042.pdf>
- Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 12697–12706). PMLR. <https://proceedings.mlr.press/v139/zhao21c.html>

Appendix

Appendices are hosted on GitHub:

<https://github.com/Professor77/Evaluating-Prompting-and-Fine-Tuning-Strategies-for-GPT-Based-Grammatical-Error-Correction/blob/main/Appendices.md>