NEVÜ
Nevşehir Hacı Bektaş Veli Üniversitesi Sosyal Bilimler Enstitüsü dergisi

Sosyal Bilimlerde Yapay Zeka:
Kuram, Uygulama ve
Gelecek Perspektifleri

Vol.15
SPECIAL ISSUE 2025

EVALUATION OF OPTICAL CHARACTER RECOGNITION ACCURACY OF CLAUDE OPUS 4 IN HANDWRITTEN STUDENT COMPOSITIONS

ÖĞRENCİLERİN EL YAZISI KOMPOZİSYONLARINDA CLAUDE OPUS 4'ÜN OPTİK KARAKTER TANIMA DOĞRULUĞUNUN DEĞERLENDİRİLMESİ

Yusuf Emre YEŞİLYURT
Burdur Mehmet Akif Ersoy University
Education Faculty, Department of Foreign Languages Education
yeyesilyurtnew@gmail.com
0000-0002-8520-1359

Geliş Tarihi:
22.08.2025

Kabul Tarihi:
16.11.2025

Yayın Tarihi:
04.12.2025

Keywords
Handwriting Recognition,
Automated Writing Evaluation,
Multimodal Large Language Models
Teacher–AI Collaboration

Anahtar Kelimeler
El Yazısı Tanıma,
Otomatik Yazma Değerlendirmesi,
Çok Kipli Büyük Dil Modelleri,
Öğretmen–YZ İş Birliği

ABSTRACT

This study evaluates Claude Opus 4 for optical character recognition (OCR) on 30 handwritten English as a Foreign Language (EFL) essay by testing if its fidelity is adequate for AI-supported assessment. For each essay, a human baseline and Claude's verbatim output were obtained via the web interface. Accuracy was analyzed at character, word, and sentence-coherence levels. The findings revealed high character accuracy (over 98%) and strong word accuracy (over 96%), though sentence coherence was more variable. While word accuracy showed a significant correlation with coherence, character accuracy did not. A 95% word-accuracy threshold was found effective in separating usable from unreliable transcripts. The most common errors were substitutions, punctuation/capitalization, and deletions, while rarer hallucinated insertions (text not present in the original) and semantic distortions disproportionately harmed meaning. The study recommends spot-checking outputs with ≥95% word accuracy and fully re-transcribing those below, suggesting that an "AI + Teacher" workflow remains essential for high-stakes use.

ÖZ

Bu çalışma, İngilizceyi Yabancı Dil Olarak (EFL) öğrenenlerin el yazılarıyla yazdıkları 30 kompozisyon üzerinde Claude Opus 4'ün optik karakter tanıma (OCR) performansını ve bu performansın ortaya koyduğu ürünün yapay zekâ destekli değerlendirmeye uygunluğunu değerlendirmeyi amaçlamıştır. Sınıfta yazılan kompozisyonlar fotoğraflanmış ve web arayüzü üzerinden işlenmiştir. Doğruluk; insan referans metinlerine kıyasla karakter, kelime ve cümle tutarlılığı düzeylerinde değerlendirilmiştir. Sonuçlar yüksek karakter (%98 üzeri) ve kelime (%96 üzeri) doğruluğu göstermiş, ancak cümle tutarlılığı değişkenlik göstermiştir. Kelime doğruluğu tutarlılık ile anlamlı bir ilişki gösterirken, karakter doğruluğu göstermemiştir. %95'lik bir kelime doğruluğu eşiğinin, kullanılabilir transkriptleri ayırt etmede etkili olduğu bulunmuştur. Yaygın hatalar arasında yer değiştirme, noktalama/büyük harf ve silme yer alırken; daha nadir görülen halüsinasyonlar ve anlamsal bozulmalar anlamı orantısız şekilde etkilemiştir. Çalışma, "Yapay Zekâ (YZ) + Öğretmen" iş akışının gerekliliğini koruduğunu öne sürmekte, %95 kelime doğruluğunun üzerindeki çıktıların nokta atışı kontrol edilmesini ve bu oranın altındakilerin yeniden transkribe edilmesini önermektedir.

DOI: https://doi.org/10.30783/nevsosbilen.1770463
Atıf/Cite as: Yeşilyurt, Y. E. (2024). Evaluation of Optical Character Recognition Accuracy of Claude Opus 4 in Handwritten Student Compositions. *Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi, Sosyal Bilimlerde Yapay Zekâ: Kuram, Uygulama ve Gelecek Perspektifleri Özel Sayısı*, 142-156.

Introduction

Writing assessment remains a cornerstone of English as a Foreign Language (EFL) instruction, where students' written output serves as a key indicator of both academic performance and language development (Schaefer, 2008; Karatay & Karatay, 2024). However, traditional assessment practices in EFL contexts face persistent challenges. Manual grading is time-consuming, often delayed, and subject to inconsistency due to rater subjectivity, varying interpretations of rubrics, and learner proficiency gaps (Yu & Yang, 2021). Even when feedback is delivered, students may resist or fail to engage meaningfully with it (Han & Hyland, 2015; Shen & Chong, 2023). These limitations not only burden educators but also reduce the pedagogical value of writing tasks.

In response, the educational field has increasingly explored the integration of Artificial Intelligence (AI) into writing assessment. Automated Writing Evaluation (AWE) systems such as *e-rater* (ETS) or *Criterion* have been developed to deliver scalable, consistent, and timely feedback (Karatay & Karatay, 2024). More recently, advanced multimodal AI models like OpenAI's GPT, Anthropic's Claude and Google's Gemini have further expanded possibilities for AI-supported feedback and evaluation. These models can interpret not only typed input but also visual data—including handwriting, which offers new pathways for processing analog student work (Kim et al., 2025).

Yet, a significant practical barrier remains: the dominance of handwritten responses**,** especially in time-constrained, in-class essay tasks and high-stakes exams. In countries like Türkiye and many others worldwide, pen-and-paper exams persist due to infrastructure limitations, exam security concerns, or pedagogical traditions (Chan, 2023). This creates a fundamental bottleneck—AI tools designed for digital text input cannot interact with handwritten scripts unless an accurate Optical Character Recognition (OCR) process is employed (Ramani et al., 2024).

OCR is thus the critical first step in enabling automated feedback and/or scoring for handwritten essays. However, its accuracy varies with handwriting legibility, script style, and image quality, and even modest transcription errors can undermine the reliability of downstream AI analysis (AlKendi et al., 2024; Hamdi et al., 2023; Michail et al., 2025). Emerging evidence suggests that advanced multimodal AI models such as Claude 3.5 may outperform traditional OCR engines in certain contexts, especially with modern, cursive handwriting (Kim et al., 2025). Yet, empirical evaluations of Claude's OCR capabilities on real-world educational data remain scarce. Addressing this gap, Yeşilyurt and Sevli (2024) conducted a preliminary investigation using ChatGPT-4o's OCR and reported high character- and word-level accuracy, alongside minor sentence-level issues in specific handwriting cases.

This preliminary study addresses this gap by evaluating the OCR performance of Claude on handwritten English essays written by university-level EFL students. Following the preliminary investigation of ChatGPT-4o's OCR capabilities (Yeşilyurt & Sevli, 2024), Anthropic's Claude was selected for this study as it represents a primary, state-of-the-art competitor with distinct architecture and advanced vision capabilities. A focused evaluation of its performance is thus a necessary step in understanding the practical utility of current-generation LLMs for this task. By focusing on the foundational process of converting analog text into machine-readable form, this research lays the groundwork for future AI-assisted feedback systems. Accurate OCR is not merely a technical convenience but a prerequisite for the equitable and scalable integration of AI into real-world writing assessment.

Methodology

Research Design

This study employs a quantitative document analysis design to evaluate the OCR performance of Claude Opus 4 in transcribing handwritten English essays written by EFL university students. Evaluations were conducted via the provider's web interface. No manual control over generation hyperparameters (e.g., temperature, top-p) was available, so default service settings applied to all runs. The focus is on assessing transcription accuracy across three linguistic levels—character, word, and sentence—by comparing Claude's output against human-transcribed ground-truth data. The focus on these three linguistic levels was deliberate. Character-level accuracy serves as the baseline for granular technical fidelity. Word-level accuracy measures the preservation of lexical units, which is essential for meaning. Finally, sentence-level coherence evaluates the output's semantic integrity, which is the most critical factor for determining if a transcript is usable for downstream assessment and feedback tasks (Aljishi et al., 2024).

Data Source

The data corpus consisted of 30 handwritten English essays, each produced by an undergraduate student enrolled in an EFL academic writing course at a public university in Türkiye. All participants were at the intermediate level of English proficiency, based on their CEFR-aligned course placements (Council of Europe, 2020). Each essay was written as part of a timed, in-class argumentative writing task, simulating real-world assessment conditions where digital entry was not permitted. The students were informed that their anonymized essays might be used for educational research purposes, and ethical approval was granted by the institution's research ethics board.

OCR System

Capture and System (Claude Opus 4)

Claude Opus 4 (multimodal LLM with vision) was accessed through the official web interface. Because web deployments may change over time, results reflect the service state during the study period under default provider settings (no exposed controls for decoding randomness or style).

Source pages were photographed using a Samsung Galaxy S24+ rear camera under daylight conditions in an office setting. The Samsung Galaxy S24+ was utilized as it represents a high-end, widely available smartphone, simulating the likely capture device a contemporary educator might employ for such tasks. Images were taken with the default camera app and default exposure/processing. There was no manual adjustment of resolution and no post-processing (e.g., denoising, sharpening, deskewing, or cropping) prior to upload. Photos were uploaded at the native device output (file format/size as produced by the phone) without alteration. As a capture limitation, hand-held smartphone photography in ambient light can introduce variability (e.g., slight motion blur, focus/exposure fluctuations, and compression artifacts), which may affect OCR fidelity. To maintain this ecological validity, no mitigation efforts (e.g., taking multiple shots and selecting the clearest) were applied. The goal was to test the system against a *typical* capture, not an *optimized* one, reflecting likely real-world use by an educator.

To facilitate replication, we retained the original image files (with EXIF metadata) and the timestamped outputs from the web interface. Replication should therefore use the same unaltered images and the service's web interface under default settings.

Transcription Procedure

The analysis involved two transcription layers:

Human Baseline Transcription: Each essay was manually transcribed by its author. They tried to preserve all original punctuation, spacing, grammar, and lexical idiosyncrasies. These transcriptions served as the reference standard.

Claude OCR Transcription: Each phone-captured handwritten essay image was processed via the same web interface using Claude Opus 4, a multimodal AI model capable of image-to-text transcription. The model was prompted to transcribe the content verbatim, without correcting or interpreting the text. Ten essay images were uploaded simultaneously in a single multi-image prompt. The exact prompt was: "Transcribe the attached 10 handwritten essays as they are. Do not correct or modify any part of the text." This prompt was selected as a direct, zero-shot instruction, specifically designed to mitigate the model's tendency to auto-correct or interpret the text. By explicitly forbidding modification, we aimed to force a verbatim transcription, which is essential for an OCR fidelity analysis. Text returned by the model was copied verbatim (UTF-8), preserving line breaks where provided by the service. Both transcriptions were saved as plain-text files and systematically aligned for comparison.

Accuracy Metrics and Error Analysis

The Claude OCR outputs were evaluated against the human baseline using three accuracy metrics: Character-Level Accuracy: Calculated by comparing the total number of matching characters between the human and Claude versions. Accuracy percentage was computed as: Character Accuracy = Number of Correct Characters / Total Characters in Human Version × 100.

Word-Level Accuracy: Tokens were compared using whitespace delimiters. Mismatches due to substitutions, deletions, or insertions were recorded. Word-level accuracy was calculated similarly: Word Accuracy = Number of Correct Words / Total Words in Human Version × 100.

Sentence-Level Coherence: A qualitative metric evaluating whether OCR outputs preserved the semantic integrity and readability of the original sentences. Each sentence was manually reviewed and coded as either "intact" or "distorted" (due to OCR-induced semantic shift, hallucination, or omission). The sentence coherence score was computed as the percentage of intact sentences in each essay.

In addition to descriptive accuracy rates, two complementary statistical procedures were employed to better understand the relationship between surface-level transcription fidelity and overall semantic coherence. First, Pearson correlation coefficients were calculated between sentence coherence scores and both character- and word-level accuracy to assess predictive relationships. Second, a threshold analysis was conducted by grouping essays into two categories based on word accuracy (≥ 95% vs. < 95%) and comparing their average sentence coherence scores. These procedures aimed to evaluate whether high word-level accuracy could be used as a practical indicator of transcriptions that were semantically usable for downstream assessment purposes. All analyses were conducted in Python 3.11 using pandas 2.2 (McKinney, 2010; Van Rossum & Drake, 2009); visualization used matplotlib (Hunter, 2007).

Error Categorization

To enable a fine-grained analysis of Claude's OCR performance, the errors were systematically categorized based on their nature and linguistic impact. Substitution errors refer to instances where an incorrect word or character replaces the original, as in *flaw* → *flow*. Insertions involve the addition of extraneous elements not found in the original text, such as the model rendering *"use it"* as *"use I"*. Conversely, deletions denote the omission of content that was originally present. Another noteworthy category, hallucinations, includes fabricated or nonsensical text segments that bear no traceable connection to the input. Moreover, semantic distortions encompass errors that preserve grammatical structure while altering the intended meaning of a sentence. Lastly, punctuation and capitalization errors include inconsistencies in spacing, punctuation marks, or letter casing, which may impair textual clarity and readability. Each essay in the dataset was manually coded for both the frequency and type of these errors to facilitate a nuanced evaluation of the model's transcription capabilities. Table 1 below summarizes the error types and provides illustrative descriptions for each type of OCR-related issue.

Table 1. Error Types

Error Type	Description
Substitution	Incorrect words or characters replacing the original (e.g., flaw → flow)
Insertion	Extraneous words or characters not present in the original (e.g., 'use I' instead of 'use it')
Deletion	Missing content that was present in the original
Hallucination	Fabricated or nonsensical text not traceable to the original input
Semantic Distortion	Errors that alter sentence meaning while maintaining structural plausibility
Punctuation and Capitalization	Changes to spacing, punctuation marks, or letter casing that could affect readability

The classifications in Table 1 served as the basis for subsequent quantitative error distribution analysis, enabling consistent cross-essay comparisons. They informed the structured annotation process used throughout the analysis.

Tools and Analysis Environment

Text alignment and accuracy computations were conducted using Python 3.11 with difflib and pandas (McKinney, 2010; Van Rossum & Drake, 2009). Sentence-level coherence judgments were made manually by the lead researcher and cross-validated on a random subset by a second rater to ensure consistency. Descriptive statistics were computed using standard spreadsheet functions and visualized using matplotlib (Hunter, 2007).

Results

This section presents the findings of Claude Opus 4's OCR performance in transcribing 30 handwritten English essays produced by EFL university students. The results are organized around three levels of analysis: quantitative accuracy metrics, threshold-based reliability, and error type impact.

Descriptive Accuracy Metrics

Claude's OCR performance was evaluated across three linguistic levels: character, word, and sentence. As summarized in Table 2 and visualized in Figure 1, the model demonstrated consistently high character-level accuracy (M = 98.51%, SD = 0.94), with most essays exceeding 97%. Word-level accuracy also remained strong (M = 96.48%, SD = 2.55), though more variable than character-level performance. In contrast, sentence-level coherence, measured through manual holistic evaluation, showed broader fluctuation (M = 77.65%, SD = 10.76), suggesting that even minor word-level deviations could affect semantic integrity.

Table 2. Accuracy Rates Across Levels

Essay No.	Character Accuracy (%)	Word Accuracy (%)	Sentence Coherence (%)
1	99.61	98.04	83.33
2	97.54	94.51	76.47
3	97.54	93.88	72.22
4	97.58	93.82	71.43
5	99.64	98.94	91.67
6	98.85	96.88	71.43
7	99.34	98.05	84.6
8	99.82	100	100
9	99.61	98.02	87.5
10	99.14	97.31	60
11	98.68	94.73	58.33
12	99.5	97.92	77.78
13	99.42	97.95	72.7
14	99.46	97.71	62.5
15	99.43	97.21	70
16	99.31	97.98	85.71
17	99.74	99.04	85.71
18	99.85	98.66	66.67
19	99.23	97.65	57.14
20	98.62	97.3	71.43
21	98.08	97.12	85.71
22	98.09	97.51	85.71
23	97.27	97.48	85.71
24	96.9	97.62	85.71
25	97.09	97.37	85.71
26	96.51	96.67	85.71

27	98.79	93.57	73.91
28	98.03	88.48	65
29	98.68	92.4	75
30	99.57	98.13	100

Table 2 summarizes descriptive OCR accuracy across linguistic levels. To visualize these patterns across essays, Figure 1 presents the distribution of accuracy rates at the character, word, and sentence levels.

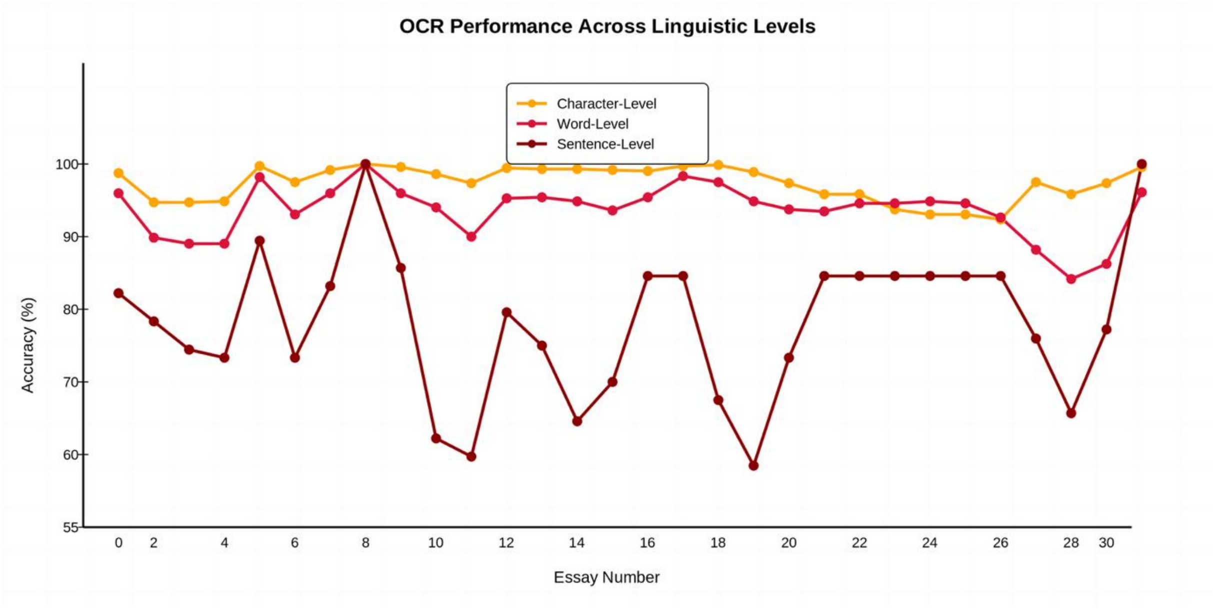

Figure 1. OCR Accuracy Performance of Claude Opus 4 Across Character-, Word-, and Sentence-Level

Table 2 and Figure 1 summarize descriptive OCR accuracy at the character-, word-, and sentence-levels: character accuracy is consistently high (M = 98.51%, SD = 0.94; most essays ≥ 97%), word accuracy is also strong but more variable (M = 96.48%, SD = 2.55), and sentence-level coherence shows the widest dispersion (M = 77.65%, SD = 10.76), underscoring that small word errors can disproportionately affect meaning.

Correlation Analyses

To determine whether surface-level OCR accuracy predicts meaningful transcription quality, Pearson correlations were computed between sentence coherence and both character and word accuracy. As shown in Table 3, a moderate, statistically significant correlation was found between word accuracy and sentence coherence (r = .44, p = .015), indicating that higher word-level fidelity generally supports coherent output. However, no meaningful relationship was observed between character accuracy and sentence coherence (r = –.02, p = .922), suggesting that character-level precision alone is insufficient to guarantee semantic usability.

Table 3. Correlations Between OCR Accuracy and Sentence Coherence

Correlation Pair	Pearson r	p-value
Word Accuracy ↔ Sentence Coherence	0.438485724	0.015358575
Character Accuracy ↔ Sentence Coherence	-0.018603462	0.922270649

Table 3 shows a moderate, statistically significant positive association between word accuracy and sentence coherence (r = .44, p = .015). By contrast, character accuracy is essentially unrelated to coherence (r = −.02, p

= .922). Thus, word-level fidelity—not character-level precision—better predicts whether OCR output is semantically usable.

Threshold Analysis

To establish a practical benchmark for safe automation, sentence coherence scores were compared between essays with word accuracy ≥ 95% and those below that threshold. As shown in Figure 2 and Table 4, essays with higher word-level accuracy yielded a mean coherence score of 80.11%, compared to 70.34% for those below the threshold. This supports the use of a 95% word accuracy cut-off as a viable indicator of assessment-ready transcriptions.

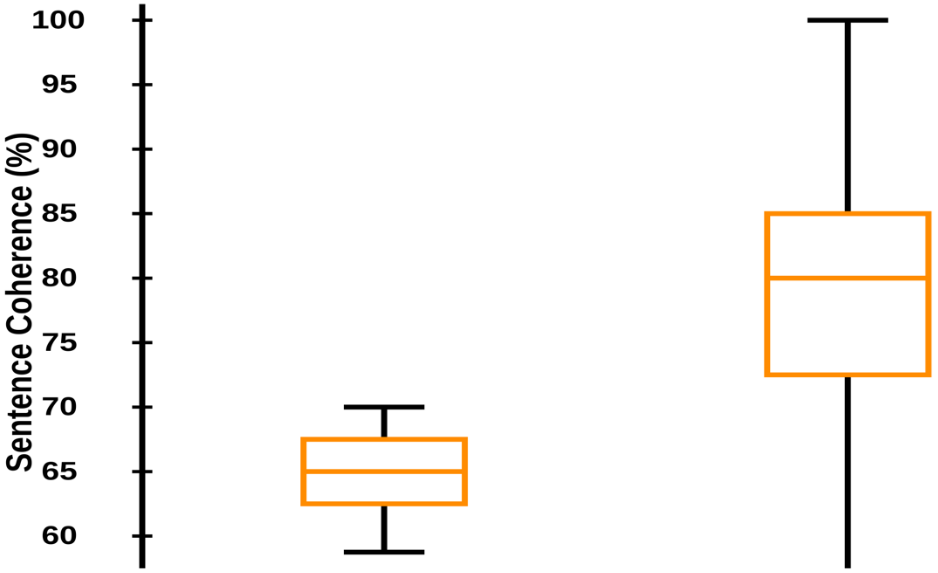

Figure 2. Sentence Coherence by OCR Word-Accuracy Threshold

Figure 2 visualizes the threshold effect: essays with ≥95% word accuracy cluster at higher coherence levels (around ~80%), with limited overlap from the lower group. The <95% group centers near ~70%, indicating a clear performance separation. This pattern supports ≥95% as a practical cut-off for triaging OCR outputs.

Table 4. Sentence Coherence by Word-Level Accuracy

Word Accuracy Group	Number of Essays (n)	Mean Sentence Coherence (%)	Standard Deviation (SD)
≥ 95%	22	80.11	9.65
< 95%	8	70.34	9.75

Table 4 quantifies the threshold pattern: coherence averages 80.11% (SD = 9.65, n = 22) for ≥95% versus 70.34% (SD = 9.75, n = 8) for <95%. The ~10-point gap is pedagogically meaningful for downstream scoring and feedback. Accordingly, ≥95% word accuracy is a workable operational rule for when OCR can be relied upon with minimal human correction.

Error Type Frequency and Impact

A manual review of Claude's transcriptions revealed six recurring error types, as visualized in Figure 3. Substitution errors (e.g., "flaw" → "flow") were the most frequent (n = 89), followed by punctuation/capitalization issues (n = 43) and deletions (n = 31). While semantic hallucinations and distortions were relatively rare (n = 11 and n = 18, respectively), their impact on coherence was disproportionately high. Essays containing such errors typically exhibited marked breakdowns in sentence structure, often rendering parts of the text ungradable without human intervention.

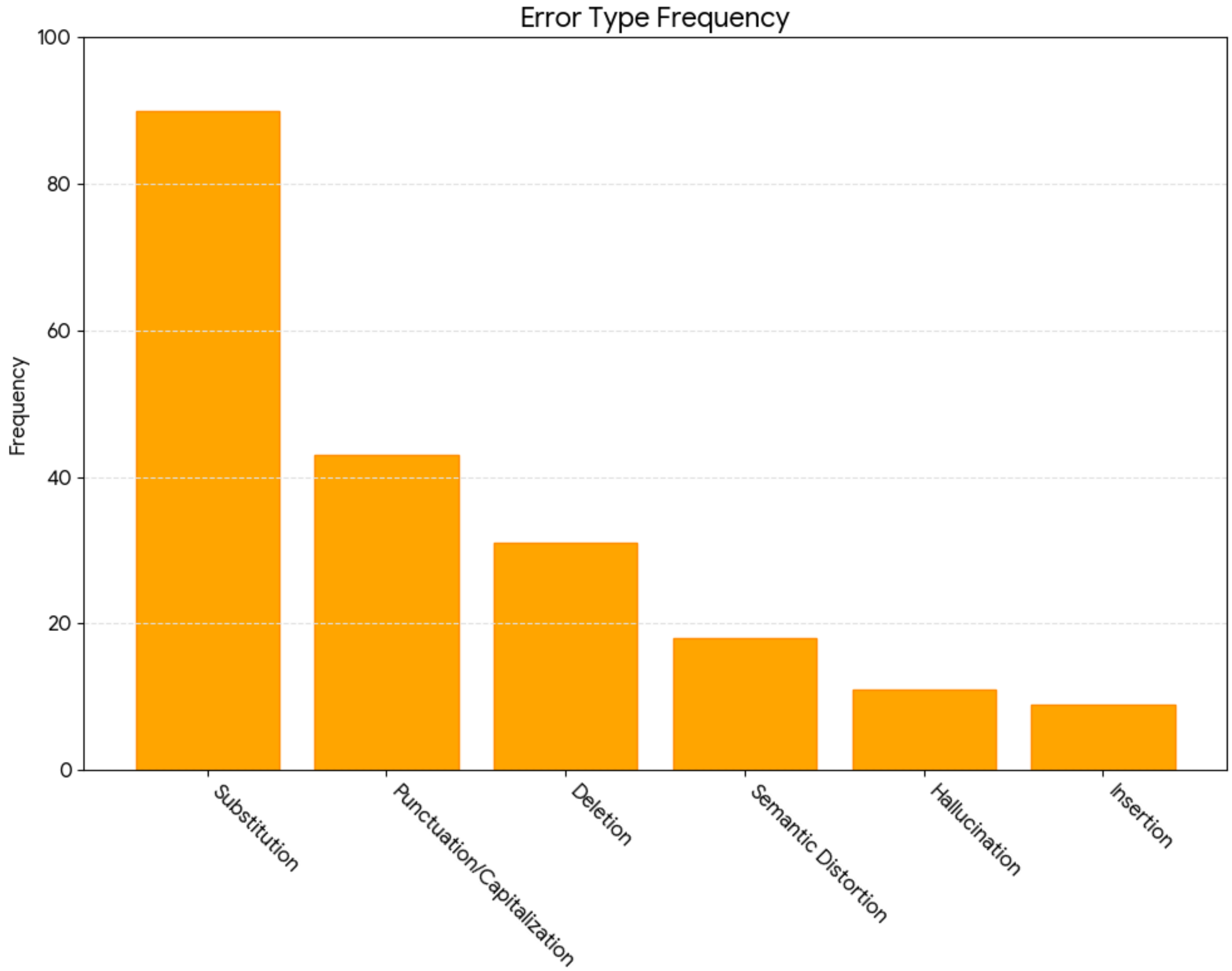

Figure 3. Distribution of OCR Error Types in Claude Opus 4 Outputs

Figure 3 shows substitutions as most frequent, followed by punctuation/capitalization issues and deletions. Although hallucinations and semantic distortions are rarer, they align with the largest coherence losses and often necessitate human review. Practically, quality control should prioritize detecting these high-impact categories while batch-correcting common surface errors.

Discussion and Implications

Interpretation and Implications for Practice

The results of this study indicate that Claude's OCR engine can accurately transcribe handwritten EFL essays, effectively bridging the analog-to-digital gap in writing assessment. Claude achieved high character- and word-level transcription accuracy, corroborating earlier evidence that advanced multimodal AI models may outperform traditional OCR systems on contemporary handwriting. Complementary to Yeşilyurt & Sevli (2024) on ChatGPT-4o, our findings provide Claude-specific OCR evidence on authentic, in-class EFL essays, thereby broadening model-level evidence for AI-supported assessment. This finding aligns with Yeşilyurt and Sevli's (2024) preliminary study of ChatGPT's OCR, which also reported high character- and word-level accuracy rates recognition rates apart from minor errors in certain handwriting styles. By successfully converting pen-and-paper essays into machine-readable text, our study addresses the "fundamental bottleneck" identified in prior research: AI-powered evaluation tools cannot interact with handwritten scripts unless an OCR step is applied (Kim et al., 2025). In contexts where in-class and high-stakes exams are still predominantly handwritten (e.g., Türkiye), this is a critical advancement. An effective AI-driven OCR paves the way for deploying automated writing evaluation in such settings, enabling consistent, immediate feedback even for analog student work. Furthermore, automating the transcription process can help mitigate human rater inconsistencies and subjectivity in scoring, since the text that the AI evaluates is standardized. In short, our findings demonstrate that integrating AI-based OCR can make EFL writing assessment more scalable and equitable by opening AI's benefits to traditionally paper-based classrooms.

Another key implication of this study is the potential to significantly reduce teacher workload in writing assessment. Combining OCR with AI evaluation consolidates what were once multiple labor-intensive tasks—deciphering handwriting, assigning scores, and providing feedback into a single streamlined process. This automation offers substantial time savings. Empirical studies have shown that (AWE) tools can cut the time teachers spend on providing feedback by a factor of two to three without diminishing the amount or quality of feedback given (Han & Sari, 2022; Wilson & Czik, 2016). In our context, Claude's ability to both transcribe and generate preliminary assessments means that instructors are relieved from the chore of deciphering illegible scripts and performing tedious error corrections. Indeed, AWE systems significantly reduce the effort required for personalized feedback delivery (Han & Li, 2024), allowing educators to reallocate their time. Instead of expending hours on mechanical marking, teachers can focus on higher-order concerns such as content, argumentation, and organization—feedback areas that benefit most from human expertise. This shift can elevate the pedagogical value of writing tasks, as instructors spend more time discussing global writing issues (e.g., idea development, coherence) rather than marking basic language errors. Moreover, lightening the grading burden may encourage teachers to assign writing tasks more frequently. Research suggests that when faced with overwhelming marking loads, teachers often limit the number of writing assignments or forego detailed feedback to avoid overwork (Karatay & Karatay, 2024). By alleviating these pressures, AI-supported assessment can foster a more continuous practice–feedback cycle: instructors may feel empowered to integrate more frequent short essays or drafts, knowing that an AI can handle initial evaluations swiftly. Increased writing practice accompanied by prompt feedback is expected to benefit student learning, as it closes the gap between performance and targeted expectations in a timely manner (Biber et al., 2011).

From the student perspective, the implications are equally promising. If teachers implement more regular writing exercises with quicker turnaround on feedback, students gain additional opportunities to improve their writing skills. Prior studies have found that immediate, AI-generated feedback can spur students to engage in more revisions and iterative improvements to their texts (Guo & Wang, 2023; Chen et al., 2009). For example, AWE tools have been shown to increase the number of revisions students perform and enhance certain aspects of writing quality, especially in mechanics and grammar (Wilson & Czik, 2016). Timely feedback is known to strengthen the formative learning process, as it allows learners to act on corrections while the material is still fresh (Biber et al., 2011). By converting handwritten work to digital text on the fly, Claude's OCR enables this rapid feedback loop even in traditionally feedback-sparse contexts such as handwritten exams. Furthermore, consistent automated scoring and comments can give students a clearer sense of their writing performance free from the variability of human markers, which may improve their confidence and uptake of feedback. Of course, human oversight and guidance remain crucial to ensure students interpret and apply AI feedback correctly. But overall, a system that "vastly accelerates the practice–feedback loop" holds considerable potential for improving writing outcomes (Kellogg et al., 2010). It could particularly help large EFL classes where individualized teacher feedback for every draft is impractical; with AI handling initial critiques, students can receive at least some formative input on each piece of writing, rather than waiting weeks for a grade.

Crucially, our findings underscore the importance of an "AI + Teacher" collaborative model in writing assessment. While AI tools like Claude can efficiently handle transcription and generate preliminary evaluations, they are not a wholesale replacement for expert human judgment. Recent research emphasizes that the best results arise when AI's analytical strengths are paired with the teacher's pedagogical and contextual knowledge (Steiss et al., 2024; Han & Li, 2024). Han and Li (2024) describe how ChatGPT can mediate the feedback process, allowing teachers to provide detailed comments at scale, yet the teacher still must customize and moderate this feedback to fit their students' needs. In our scenario, a teacher could use Claude's OCR output and automated feedback as a starting point, then validate, adjust, and expand upon that feedback for each student. This approach leverages AI to cover routine tasks and frees the instructor to add nuanced, personalized insights that AI might miss (for instance, addressing subtle content issues or motivational aspects). The value of maintaining teacher involvement is supported by Steiss et al. (2024), who found that well-trained teachers still provide higher-quality feedback than ChatGPT alone—especially in terms of tailored advice—but they also note that AI feedback is extremely useful when instructors are stretched thin or must review many drafts quickly. Similarly, Guo and Wang (2023) observed that ChatGPT can deliver a greater volume of specific feedback across content, organization, and language, but teachers excel at giving adaptive, student-centered comments. Therefore, the ideal implementation is a synergistic one: AI systems handle the heavy lifting of transcription and error detection, and perhaps suggest corrections, while teachers oversee the process, ensure accuracy, and provide the pedagogical interpretation of the feedback. This synergy not only safeguards quality but also keeps the human element in the feedback loop, which is essential for student motivation and trust. Notably, students in an AI-supported feedback environment have been shown to incorporate more of the feedback into their revisions, indicating deeper engagement with the comments when the feedback is mediated by a teacher figure (Barrot, 2023). Maintaining teacher involvement helps prevent students from feeling that impersonal machine feedback is the final word; instead, the AI output becomes a springboard for teacher–student conferences or follow-up explanations, thus enriching the learning experience. In short, our work implies that teacher training and feedback literacy will be vital—instructors need guidance on interpreting AI-generated evaluations and integrating them into practice. With clear guidelines and support, teachers can maximize the benefits of tools like Claude while avoiding potential pitfalls such as over-reliance on formulaic AI comments or uncritical trust in OCR results.

Operationally, our results support using a ≥95% word-accuracy threshold as a triage rule for OCR outputs (see Results: Threshold Analysis). Transcriptions meeting or exceeding this level can be spot-checked—focusing on flagged spans—rather than fully re-transcribed. Outputs below 95% should be re-transcribed in full before any downstream automated analysis or feedback.

Taken together, the present study provides an encouraging proof-of-concept: with accurate OCR capabilities, AI can be harnessed to streamline writing assessment, reduce instructors' workloads, and possibly enhance the feedback landscape in EFL writing instruction. As OCR and language models continue to improve, we envision AI-powered assessment tools becoming a supportive partner for educators—not replacing teacher judgment, but extending it to achieve more timely, consistent, and formative evaluation of student writing. This evolution holds the promise of reinvigorating the teaching of writing by freeing up teachers' time for richer instructional interactions and giving learners more frequent, high-quality feedback on their journey to writing proficiency.

Limitations

This study is preliminary and context-bound. The dataset comprises authentic handwritten essays drawn from a single institutional setting, which constrains the generalizability of the findings. Handwriting legibility and capture conditions (e.g., device, resolution, lighting) were not systematically manipulated; as such, observed fidelity may vary under different imaging parameters and scripts. Ground-truthing relied on human transcriptions with limited inter-rater checks on a subset of pages, which may bound the precision of error estimates. We did not benchmark against classical OCR engines or alternative LLM-based OCR systems under identical conditions. Such a comparative analysis, while a critical next step, was considered beyond the scope of this *preliminary* study which was focused on first establishing a detailed fidelity baseline for Claude Opus 4. Therefore, comparative performance remains uncertain. Finally, the focus was on English essays; results may not transfer to other languages or orthographies without additional validation.

Directions for Future Research

Future research should broaden populations (multi-site samples across age groups, proficiency bands, and handwriting styles) and diversify orthographies (e.g., Turkish cursive and other scripts) to test robustness. Systematic experiments that vary imaging conditions can clarify sensitivity to input quality. Head-to-head benchmarks with established OCR tools and other LLMs are needed to identify cost–accuracy trade-offs. In addition, classroom-based and preferably longitudinal studies should investigate the real-world impact of OCR-mediated AWE on students' revision quality, writing development, and motivation over time, so that the practical consequences of adopting such systems can be documented beyond technical fidelity. Integrating OCR with end-to-end AI-assisted writing pipelines (automated analysis, formative feedback, and revision tracking) would allow evaluation of downstream effects on feedback timeliness, revision behavior, and writing outcomes. Human-in-the-loop safeguards—such as confidence flags and triage of low-fidelity spans—should be designed and tested for reliability and teacher workload. Finally, qualitative work with instructors and students can illuminate trust, uptake, and feedback literacy in AI-mediated assessment contexts

Conclusion

This study evaluated Claude (Opus 4) for OCR transcription of 30 handwritten EFL essays. The system achieved high character- and word-level fidelity (M_character ≈ 98.5%, M_word ≈ 96.5%), while sentence-level coherence varied across scripts. Word accuracy moderately predicted coherent sentences (r ≈ .44), and a practical threshold of ≥95% word accuracy emerged as a useful indicator of transcription that is reliable enough for downstream analysis. Although rare, hallucinations and semantic distortions carried disproportionate impact on coherence and therefore warrant human review in high-stakes settings. Taken together, the findings indicate that AI-based OCR can remove a key bottleneck in paper-based writing assessment, enabling faster feedback cycles and reducing instructor workload. We argue for an "AI + Teacher" model in which OCR and initial analytics are automated, while teachers validate, contextualize, and extend feedback. With careful safeguards (confidence flags, review of low-fidelity passages) and attention to ethics and equity, integrating OCR into EFL assessment can make feedback timelier, more consistent, and more scalable—without displacing expert human judgment.

Declaration of Conflicting Interests

The author(s) declare that there are no actual or potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Declaration of AI Use

During the preparation of this work, the author(s) used ChatGPT exclusively for language editing and proofreading purposes to improve readability. The author(s) reviewed and edited the output as needed and take(s) full responsibility for the content of the publication.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Statements of Publication Ethics

We hereby declare that the study involves no unethical issues and that research and publication ethics were observed carefully throughout the process.

Researchers' Contribution Rate

The study was conducted and reported with equal collaboration among the author(s) (equal contribution).

Ethics Committee Approval Information

Ethics committee approval for this study was obtained from the Non-Interventional Clinical Research Ethics Committee (GOKA), Burdur Mehmet Akif Ersoy University, at the meeting dated 06 August 2025 (Meeting No: 2025/8) with Decision No: GO 2025/1916.

References

Aljishi, F., Mughaus, R., Luqman, H., & Parvez, M. T. (2024). A comparative study of four handwritten text recognition models in Arabic script. *Ingénierie des Systèmes d'Information, 29*(6), 2243–2250. https://doi.org/10.18280/isi.290614

AlKendi, W., Gechter, F., Heyberger, L., & Guyeux, C. (2024). Advancements and challenges in handwritten text recognition: A comprehensive survey. *Journal of Imaging, 10*(1), 18. https://doi.org/10.3390/jimaging10010018

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

Biber, D., Nekrasova, T., & Horn, B. (2011). The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. *ETS Research Report Series*, *2011*(1), i-99. https://doi.org/10.1002/j.2333-8504.2011.tb02241.x

Chan, C. K. Y. (2023). A systematic review – handwritten examinations are becoming outdated, is it time to change to typed examinations in our assessment policy? *Assessment & Evaluation in Higher Education*, *48*(8), 1385–1401. https://doi.org/10.1080/02602938.2023.2219422

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment—Companion volume.* Council of Europe Publishing

Guo, K., Wang, D. To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies* 29, 8435–8463 (2024). https://doi.org/10.1007/s10639-023-12146-0

Hamdi, A., Linhares Pontes, E., Sidere, N., Coustaty, M., & Doucet, A. (2023). In-depth analysis of the impact of OCR errors on named entity recognition and linking. *Natural Language Engineering, 29*(2), 425–448. https://doi.org/10.1017/S1351324922000110

Han, J., & Li, M. (2024). *Exploring ChatGPT-supported teacher feedback in the EFL context.* System, 126, 103502. https://doi.org/10.1016/j.system.2024.103502

Han, Y., & Hyland, F. (2015). Exploring learner engagement with written corrective feedback in a Chinese tertiary EFL classroom. *Journal of Second Language Writing, 30*, 31–44. https://doi.org/10.1016/j.jslw.2015.08.002

Han, T., & Sarı, E. (2022). *An investigation on the use of automated feedback in Turkish EFL students' writing classes.* Computer Assisted Language Learning. Advance online publication. https://doi.org/10.1080/09588221.2022.2067179

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Karatay, Y., & Karatay, L. (2024). Automated writing evaluation use in second language classrooms: A research synthesis. *System, 123*, 103332. https://doi.org/10.1016/j.system.2024.103332

Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does Automated Feedback Help Students Learn to Write? *Journal of Educational Computing Research*, *42*(2), 173-196. https://doi.org/10.2190/EC.42.2.c

Kim, S., Baudru, J., Ryckbosch, W., Bersini, H., & Ginis, V. (2025). Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical records. *arXiv preprint.* https://arxiv.org/abs/2501.11623

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference (SciPy 2010)* (pp. 56–61)

Michail, A., Opitz, J., Wang, Y., Meister, R., Sennrich, R., & Clematide, S. (2025). Cheap character noise for OCR-robust multilingual embeddings. *Findings of the Association for Computational Linguistics: ACL 2025*, 11705–11716

Ramani, K., Maheswari, G. U., Krishna, K. P., Meghashyam, S. V., Komirisetty, V. P. K., & Duraiswamy, Y. (2024). Automatic grading of answer sheets using machine learning techniques. In K. R. Madhavi, P. Subba Rao, J. Avanija, I. L. Manikyamba, & B. Unhelkar (Eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)* (pp. 275–284). Atlantis Press. https://doi.org/10.2991/978-94-6463-471-6_27

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493. https://doi.org/10.1177/0265532208094273

Shen, R., & Chong, S. W. (2023). Learner engagement with written corrective feedback in ESL and EFL contexts: A qualitative research synthesis using a perception-based framework. *Assessment & Evaluation in Higher Education, 48*(3), 276–290. https://doi.org/10.1080/02602938.2022.2072468

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). *Comparing the quality of human and ChatGPT feedback of students' writing.* Learning and Instruction, 91, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Van Rossum, G., & Drake, F. L., Jr. (2009). *Python 3 reference manual.* CreateSpace

Wilson, J., & Czik, A. (2016). *Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality.* Computers & Education, 100, 94–109. https://doi.org/10.1016/j.compedu.2016.05.004

Yeşilyurt, Y. E., & Sevli, O. (2024). Evaluating the accuracy of optical character recognition in ChatGPT for handwritten student essays. In İ. Erpay (Ed.), *Proceedings of the 2nd BİLSEL International Kibyra Scientific Researches Congress* (pp. 195–201). Astana Publications.

Yu, R., & Yang, L. (2021). ESL/EFL learners' responses to teacher written feedback: Reviewing a recent decade of empirical studies. *Frontiers in Psychology, 12*, 735101. https://doi.org/10.3389/fpsyg.2021.735101

GENİŞLETİLMİŞ ÖZET

Bu çalışma, üniversite düzeyindeki İngilizceyi yabancı dil olarak öğrenen (EFL) öğrencilerin el yazısıyla ürettiği kompozisyonlarda Anthropic Claude Opus 4'ün optik karakter tanıma (OCR) performansını inceleyen ön bulguları sunmaktadır. Yazma değerlendirmesinin önemli bir kısmı hâlen kâğıt-kalem ortamında yürütüldüğünden, analog çıktının güvenilir biçimde sayısallaştırılması, yapay zekâ destekli geri bildirim ve puanlamanın ön koşuludur. El yazısı çözümlerinde küçük yazım veya tanıma hatalarının bile anlam bütünlüğünü bozabildiği bilindiğinden, bu çalışmanın temel amacı, Claude'un karakter, sözcük ve cümle düzeylerinde sunduğu doğruluğun, aşağı akıştaki değerlendirme süreçleri için "yeterince iyi" olup olmadığını nicel göstergelerle sınamaktır. Çalışma aynı zamanda uygulayıcılar için pratik bir eşik önerisi sunmayı ve öğretmenle yapay zekânın birlikte çalıştığı bir iş akışının nasıl yapılandırılabileceğine ilişkin kanıt üretmeyi hedefler.

Araştırma nicel belge çözümlemesi deseninde yürütülmüştür. Veri kaynağını, Türkiye'de bir devlet üniversitesinde İngilizce Yazma Becerileri 2 dersine kayıtlı, CEFR orta düzey (B1/B2) İngilizce dil yeterliliğine sahip 30 lisans öğrencisinin, sınıf içinde süreli koşullarda yazdığı tartışmacı kompozisyonlar oluşturur. Tüm öğrenciler çalışmalarının anonimleştirilmiş kopyalarının araştırma amaçlı kullanılabileceği konusunda bilgilendirilmiş, kurumun etik kurul onayı alınmıştır. Kompozisyon sayfaları, gündüz ışığında ve ofis ortamında Samsung Galaxy S24+ arka kamerasıyla, varsayılan kamera ayarlarıyla ve çözünürlük değiştirilmeden fotoğraflanmıştır. Görseller sağlayıcının web arayüzüne değiştirilmeden yüklenmiş; Claude'e tek bir çoklu-girdi oturumunda on görsel birlikte gönderilmiştir. Kullanılan talimat, "Transcribe the attached 10 handwritten essays as they are. Do not correct or modify any part of the text." biçimindedir. Her kompozisyon için iki metin elde edilmiştir: öğrencinin kendi el yazısının kendisi tarafından transcribe edilmiş hali ve Claude'un birebir OCR çıktısı. Çıktılar UTF-8 olarak satır sonları korunarak kaydedilmiş, herhangi bir düzeltme/normalizasyon yapılmamıştır.

Değerlendirme üç düzeyde yürütülmüştür. Karakter düzeyi doğruluk, insan transkripsiyonu ile Claude çıktısındaki eşleşen karakter sayısının tüm karakterlere oranıyla; sözcük düzeyi doğruluk, boşlukla ayrılmış belirteçler temelinde doğru sözcük oranıyla; cümle düzeyi tutarlılık ise her cümlenin anlam ve okunaklılık açısından "sağlam" ya da "bozulmuş" olarak kodlanmasıyla hesaplanmıştır. Ek olarak iki analiz uygulanmıştır: (i) cümle tutarlılığı ile karakter/sözcük doğruluğu arasındaki Pearson korelasyonları ve (ii) eşik analizi (sözcük doğruluğu %95 ve üzeri olanlarla altı olanların karşılaştırılması). Hata türlerini ayrıntılı görmek için yer değiştirme, ekleme, silme, halüsinasyon, anlamsal bozulma ve noktalama/büyük-harf olmak üzere altı kategoriden oluşan bir tipoloji kullanılmış; her kompozisyon bu tipolojiye göre kodlanmıştır. Analizler Python 3.11 üzerinde pandas ve görselleştirmede matplotlib ile yürütülmüştür.

Bulgular, Claude'un karakter düzeyinde yüksek doğruluk sergilediğini (Ort. = 98.51, SS = 0.94) göstermektedir. Sözcük düzeyi doğruluk da yüksek olmakla birlikte (Ort. = 96.48, SS = 2.55) karakter düzeyine göre daha değişkendir. Cümle düzeyi tutarlılık ise daha geniş bir dağılım sunmakta (Ort. = 77.65, SS = 10.76), bu da yüzeydeki küçük sözcük hatalarının anlam bütünlüğünü etkileyebildiğine işaret etmektedir. İlişkisel çözümlemeler, sözcük doğruluğu ile cümle tutarlılığı arasında orta düzeyde ve anlamlı bir ilişki bulunduğunu ortaya koymuştur (r = .44, p = .015); buna karşılık karakter doğruluğu ile cümle tutarlılığı arasında anlamlı bir ilişki saptanmamıştır (r = −.02, p = .922). Eşik analizinde, %95 ve üzeri sözcük doğruluğuna sahip kompozisyon için ortalama cümle tutarlılığı %80.11, eşik altı kompozisyonlar için ise %70.34 olarak bulunmuştur. Bu sonuç, pratikte %95 sözcük doğruluğunu değerlendirmeye uygun metinleri ayırt etmekte kullanılabilir bir gösterge olarak desteklemektedir.

Hata tipolojisi, yer değiştirme (n ≈ 89), noktalama/büyük-harf (n ≈ 43) ve silme (n ≈ 31) hatalarının en sık görüldüğünü; anlamsal bozulma (n ≈ 18) ve halüsinasyonların (n ≈ 11) daha nadir olmakla birlikte cümle tutarlılığını orantısız biçimde zedelediğini göstermiştir. Bu iki nadir hata, tekil örneklerde bütün paragrafların notlandırılamaz hâle gelmesine neden olabileceğinden, özellikle yüksek riskli/yüksek önemlilikteki kullanımlarda insan denetimini gerektirmektedir. Bulgular toplamda, Claude'un OCR işlemini çok büyük ölçüde güvenilir biçimde gerçekleştirdiğini; ancak az sayıdaki yüksek etkili hata nedeniyle hedefli insan kontrolü olmadan tam otomasyonun riskler taşıdığını göstermektedir.

Bu bağlamda çalışma, uygulayıcılara operasyonel bir karar kuralı önermektedir: %95 ve üzeri sözcük doğruluğuna ulaşan OCR çıktıları yerinde/odaklı kontrol ile (sadece işaretlenen düşük güven aralıkları) doğrulanabilir; %95'in altındaki çıktılar ise aşağı akıştaki otomatik analiz ya da geri bildirimden önce tam yeniden yazım gerektirir. Böyle bir "Yapay Zekâ + Öğretmen" düzeni, öğretmenin pedagojik sezgisini dışlamadan, yoğun rutinleri otomatikleştirerek geri bildirim döngüsünü hızlandırır. Özellikle büyük sınıflarda veya kısa sürede çok sayıda yazının değerlendirildiği durumlarda, öğrencilerin daha sık yazma ve daha hızlı geri bildirim alma olanağı doğar; bu da alanyazında vurgulanan uygulama-geribildirim döngüsünü güçlendirir. Adalet ve tutarlılık açısından, el yazısının dijital metne standart biçimde aktarılması değerlendirmenin karşılaştırılabilirliğini artırır; aynı zamanda öğretmen iş yükünde kayda değer bir azalma sağlar.

Çalışmanın sınırlılıkları, tek kurumdan elde edilmiş 30 yazıya dayanması; görüntülemenin elde akıllı telefonla ve doğal ışıkta yapılması nedeniyle olası bulanıklık/odak/sıkıştırma değişkenlikleri içermesi; yalnızca İngilizce metinlerin incelenmiş olması; Claude Opus 4'ün web arayüzündeki varsayılan ayarlarla çalıştırılması ve karşılaştırmalı (klasik OCR ya da diğer LLM tabanlı çözümlerle başa baş) testlerin yapılmamış olmasıdır. Bu nedenlerle genellenebilirlik sınırlıdır ve bulgular modelin incelenen sürüm dönemine özgüdür. Bununla birlikte, yöntem şeffaf biçimde raporlandığından (girdi koşulları, talimat, çoklu-girdi kullanımı ve ölçütler), araştırma kolaylıkla yinelenebilir.

Sonuç olarak çalışma, Claude Opus 4'ün el yazısı İngilizce kompozisyonlarda yüksek karakter ve sözcük düzeyi doğruluğa ulaştığını; ancak anlamsal bütünlüğün metinler arasında değişkenlik gösterdiğini ortaya koyarak alan yazısına kanıta dayalı bir eşik ve iş akışı ilkesi kazandırmaktadır. Gelecek çalışmaların farklı yaş/başarı düzeylerinden ve farklı yazı sistemlerinden (ör. Türkçe yazı çeşitleri ve eğik yazı) daha büyük örneklemlerle yürütülmesi; görüntüleme koşullarının sistematik olarak değiştirilmesi, klasik OCR ve diğer büyük dil modeli tabanlı çözümlerle başa baş karşılaştırmalar yapılması ve OCR-AWE-geri bildirim zincirinin öğrenci öğrenme çıktıları üzerindeki etkisinin deneysel tasarımlarla sınanması önerilir. Ayrıca öğretmen ve öğrencilerin teknolojiye güven, benimseme ve geri bildirim okuryazarlığı boyutlarının nitel ve karma yöntemlerle incelenmesi, okul ortamlarına aktarımı güçlendirecektir. Bütün bu adımlarla, OCR doğruluğunun belirlenen %95 eşik rehberi eşliğinde izlenmesi, kâğıt temelli yazma değerlendirmesinde tıkanıklığı azaltarak daha zamanlı, tutarlı ve ölçeklenebilir bir değerlendirme ekosistemine katkı sunacaktır.