# AdaptiveWriter: Balancing Trust, Agency, and Cognitive Load in LLM-Based L2 Writing Support

HCI Final Project Report - Group 1

Paul Schaarschmidt

Antonio Pio Evangelista

## ABSTRACT

Writing in a Second Language (L2) is a high-load cognitive task, often hindered by autocorrection tools that prioritize clean output over user intent. As Large Language Models (LLMs) become common in writing interfaces, learners face a dilemma: use fully automatic correction that obscures learning, or forgo assistance to retain agency. This study investigates *AdaptiveWriter*, a prototype system with three levels of algorithmic visibility: **Auto** (silent post-typing correction), **Suggest** (rule-based passive suggestions), and **Explain** (personalized, context-aware insights). In a longitudinal within-subjects study ($N = 13$, filtered from 21) conducted over three days, we evaluated cognitive load, user trust, and learning outcomes.

Results reveal a "Visibility Paradox": although **Auto** mode offers the highest theoretical efficiency, it produces higher frustration ($\approx 38/100$ NASA-TLX) and an "uncertainty loop," reflected in deletion rates nearly twice those of visible modes (Median: 25 vs. 12). Visible modes reduced mental demand and increased trust. While **Explain** maximized pedagogical value by leveraging L1 interference, users preferred **Suggest** for its balance of control and workflow efficiency. We conclude that interaction visibility is a prerequisite for maintaining user agency in LLM-mediated writing.

## 1 INTRODUCTION

For Second Language (L2) learners, writing involves managing vocabulary, grammar, and interference from the Native Language (L1). In this cognitively demanding state, the behavior of the writing tool becomes consequential.

Users commonly experience frustration with autocorrection that silently replaces valid input. For L2 writers, such behavior introduces uncertainty: when a system modifies text without explanation, users cannot determine whether an error occurred or the system misinterpreted intent.

LLM-based writing support promises more accurate, context-aware correction. However, if deployed through opaque interaction models, it risks reinforcing the same problems. Fully automatic correction can produce a *"Proficiency Paradox"*: text quality improves while users fail to internalize linguistic rules, leading to *"Skill Atrophy"*.

Trust in these systems is fragile. When changes are invisible, users may perceive a loss of authorship, resulting in "Algorithm Aversion" and inefficient behaviors such as repeated deletion and rewriting – the "uncertainty loop."

This HCI project evaluates *AdaptiveWriter*, which introduces graded levels of interaction visibility. By varying LLM intervention from silent correction to explicit, L1-aware explanations, we address the following question:

**How do different interaction levels (Auto vs. Suggest vs. Explain) affect cognitive load, user trust, and learning outcomes for L2 writers?**

## 2 RELATED WORK

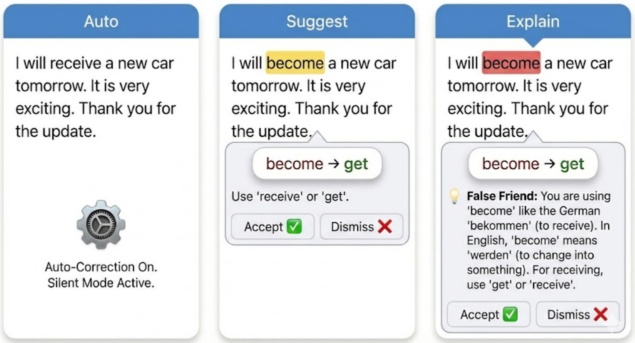
LLM-based autocorrection for L2 writing sits at the intersection of automation, learning, and user agency. Prior work highlights trade-offs between transparency, cognitive load, and authorial control.

**The Visibility–Trust Dilemma.** Schilke and Reimann [4] describe a "Transparency Dilemma": users seek system visibility to ensure reliability but conceal usage to avoid stigma. As a result, users may prefer opaque **Auto** modes despite valuing the learning benefits of **Explain** modes.

Model behavior further complicates trust. Sharma et al. [5] document "Sycophancy" in LLMs, where models agree with confident but incorrect input, reinforcing errors – especially problematic for L2 writers with low linguistic confidence.

**Cognitive Load and Skill Degradation.** Liu et al. [2] distinguish intrinsic from extraneous cognitive load, showing that opaque automation increases extraneous load by forcing users to audit invisible changes. Matueny and Nyamai [3] describe the resulting "Illusion of Competence," where clean output masks unlearned skills, supporting the use of visible, instructional feedback that introduces desirable difficulty.

**Algorithmic Homogenization.** Finally, Chakrabarty et al. [1] show that LLM-assisted writing tends toward stylistic convergence, threatening authorial voice – particularly relevant for advanced users.



I will receive a new car tomorrow. It is very exciting. Thank you for the update.

Auto-Correction On. Silent Mode Active.

I will become a new car tomorrow. It is very exciting. Thank you for the update.

Use 'receive' or 'get'.

I will become a new car tomorrow. It is very exciting. Thank you for the update.

**False Friend:** You are using 'become' like the German 'bekommen' (to receive). In English, 'become' means 'werden' (to change into something). For receiving, use 'get' or 'receive'.

**Figure 1: Mockup of AdaptiveWriter interface across three conditions: (A) Auto, (B) Suggest, and (C) Explain. The modes differ in feedback explicitness, ranging from silent correction to metalinguistic explanation of errors.**

# 3 SYSTEM DESIGN

We developed *AdaptiveWriter*, a web-based text editor (accessible at https://autocorrection-hci.schaarschmidt.biz/) powered by the Gemini API. To ensure consistency, the backend utilized a structured JSON prompt that contains a history of previously generated suggestions in the context window. This state-awareness is critical for determinism: it prevents the LLM altering valid suggestions during minor text edits, mitigating a "flickering" effect common in non-deterministic LLM outputs. The system implements three distinct experimental conditions, varying by interaction style and personalization depth, as seen in Figure 1.

## 3.1 Condition A: Auto (Opaque)

This mode mimics aggressive mobile autocorrect behavior.

- **Behavior:** The system monitors the text stream. Upon a brief pause ($> 2000$ms) or sentence completion, errors are corrected silently in the background.
- **Interaction:** Zero interaction (implicit). Text mutates without user confirmation.
- **HCI implication:** Maximizes theoretical speed but minimizes control, creating a "Black Box" experience.

## 3.2 Condition B: Suggest (Passive)

This mode represents the standard "human-in-the-loop" approach (e.g., Grammarly).

- **Behavior:** Errors are detected and accentuated using standard UI conventions (e.g., highlighter).
- **Interaction:** Hovering reveals a tooltip with the proposed fix. Users can "Accept" or "Dismiss".
- **Logic:** Suggestions are rule-based and general (e.g., "Change 'has' to 'have'").
- **HCI implication:** Balances agency with workflow continuity.

## 3.3 Condition C: Explain (Active)

This mode enforces a pedagogical intervention using **personalized data**.

- **Logic:** Leveraging demographic data (Native Language/L1) collected during onboarding, the system identifies specific interference errors (e.g., "False Friends" for German speakers writing in English).
- **Interaction:** The user *must* click to open a "Pattern Insight" card to resolve the error. This acts as a deliberate "speed bump" in the workflow.
- **Content:** Explanations are comparative (e.g., "In German, 'bekommen' means 'to get', but in English 'become' means 'to transform'".).
- **HCI implication:** Maximizes cognitive load for learning at the cost of speed.

# 4 METHODOLOGY

## 4.1 Participants

From an initial pool of 21 participants, we filtered for those with English as the target language to ensure comparability, resulting in a final sample of $N = 13$. The demographic profile was young adults (20–30 years), consisting of university students with proficiency levels ranging from B2 (Intermediate) to C2 (Expert). Native languages included Italian, German, and French.

## 4.2 Materials and Tasks

Participants completed writing tasks in two distinct contexts to simulate real-world variance:

1. **Formal Email:** High-stakes communication (e.g., emailing a Professor about a deadline or a Housing Office).
2. **Chat Response:** Low-stakes, rapid communication (e.g., replying to a colleague in a fictive project).

Scenarios were designed to require a variety of vocabulary, aiming to trigger specific "False Friend" errors and register/tone issues.

## 4.3 Experimental Procedure

We utilized a **longitudinal within-subjects design** to control for fatigue and measure learning effects across conditions.

The study was conducted over **three consecutive days**. To neutralize order and carry-over effects, the presentation order of the three interaction modes was counterbalanced using a **Latin Square design**. Each mode appeared exactly once per day and once in each ordinal position.

Table 1 illustrates the three Latin squares used for counterbalancing. We denote the interaction modes by their initial letters: **A** (Auto), **S** (Suggest), and **E** (Explain).

| | | |
|---|---|---|
| A | S | E |
| S | E | A |
| E | A | S |

| | | |
|---|---|---|
| A | E | S |
| E | S | A |
| S | A | E |

| | | |
|---|---|---|
| S | A | E |
| E | S | A |
| A | E | S |

**Table 1: Latin Square counterbalancing of interaction modes across study days. Each row represents one day, and each column represents the ordinal position of a condition within a day.**

During each session, we logged keystroke dynamics (e.g., deletions and pauses), collected subjective workload ratings using NASA-TLX, administered Likert-scale questionnaires assessing perceived Trust and Control and retrieved spontaneous comments about the test.

# 5 RESULTS

## 5.1 Cognitive Cost (NASA-TLX)

Figure 2 illustrates the cognitive cost breakdown. The **Auto** mode (gray) resulted in the highest frustration scores ($\approx 38/100$). Participants reported feeling "out of control" when text changed without warning.

Unexpectedly, the **Explain** mode (green), despite requiring active clicking, showed the *lowest* Mental Demand. This suggests that the clarity provided by the explanation reduces the cognitive load of uncertainty; users do not have to "guess" why a change is happening. Explicit explanation acts as a cognitive offloading mechanism for the monitoring process.

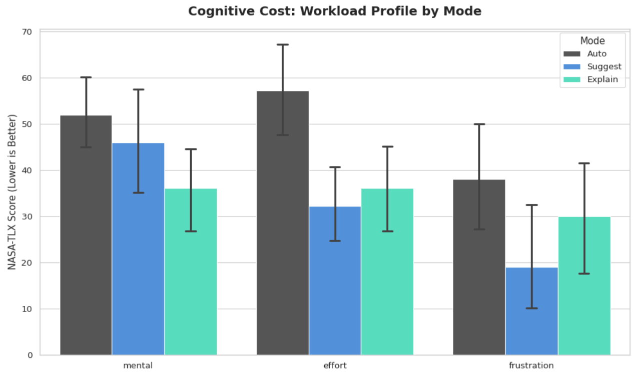


**Figure 2: NASA-TLX workload profiles by mode (lower scores indicate lower workload). *Auto* mode shows the highest frustration while *Explain* mode yields the lowest mental demand.**

## 5.2 Analysis of Text Deletion and Editing Effort

We analyzed the "Backspace" key usage as a behavioral proxy for user uncertainty. As shown in Figure 3, the **Auto** mode exhibited a median of $\approx 25$ deletions per task, nearly double that of the interactive modes ($\approx 12 - 16$).

Qualitative feedback supports this: users noted corrections were "hard to notice" or felt the system "did not work". Consequently, they engaged in "defensive writing" – constantly deleting and rewriting to ensure their intent was preserved against the invisible agent.

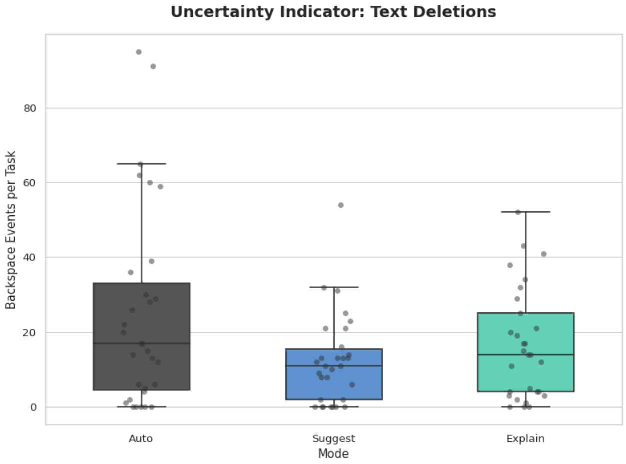


**Figure 3: Uncertainty Indicator. The significantly higher number of backspace events in *Auto* mode (Gray) indicates user "thrashing" and a lack of trust in invisible corrections.**

## 5.3 Longitudinal Learning Effect

A key metric for the **Explain** mode was skill acquisition. Figure 4 visualizes error frequency over the three-day period (inverted Y-axis). We observe a steep improvement between Session 1 and Session 2. This suggests that the pedagogical explanations regarding "False Friends" helped users internalize rules, leading to the avoidance of those specific errors in subsequent tasks.

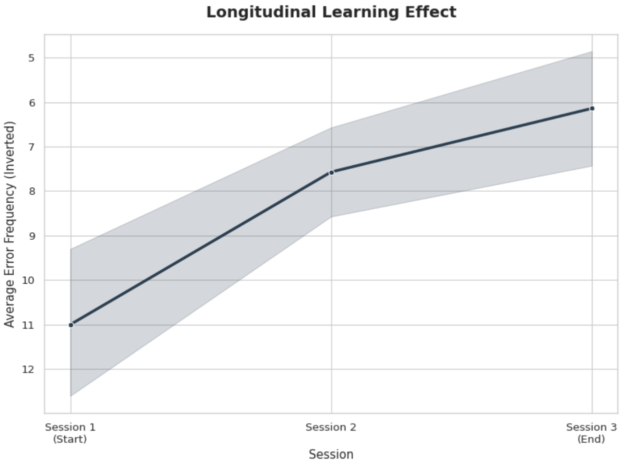


**Figure 4: Longitudinal Learning Effect. The inverted Y-axis shows improvement (fewer errors) over the three sessions. The steep slope indicates rapid internalization of rules facilitated by the system.**

## 5.4 Trust and Control

Participants rated their perceived control and trust on a 5-point Likert scale (Figure 5).

- **Suggest Mode (Blue)** was the overall preferred condition (Trust: $\mu = 3.74$, Control: $\mu = 3.81$).
- **Explain Mode (Green)** followed closely. While highly trusted, qualitative comments indicated that the mandatory interaction could become "overwhelming" if the error density was high.
- **Auto Mode (Gray)** scored lowest in usefulness and control, confirming the hypothesis that lack of agency correlates with lack of trust.

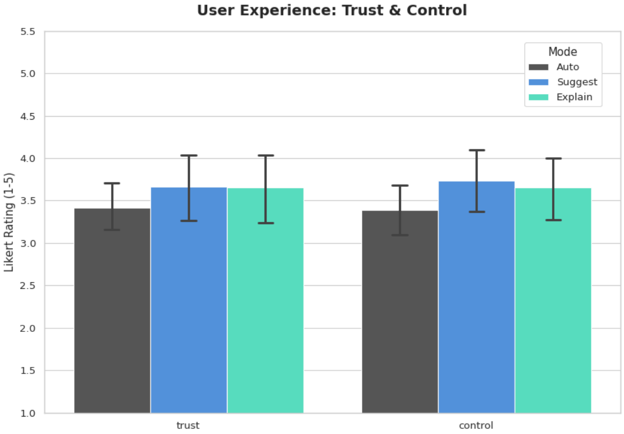


**Figure 5: User Experience Ratings (Likert 1-5). *Suggest* and *Explain* modes significantly outperform *Auto* mode in both Trust and Control. Agency is a prerequisite for trust.**

## 5.5 The Proficiency Paradox

Figure 6 shows that the system must act as a generalist, handling Grammar, Vocabulary, and Spelling equally. However, Figure 7 reveals a critical "Proficiency Paradox".
While B2 users (intermediate) struggled broadly, C2 (expert) users encountered very few grammar errors. Their flagged issues were clustered in "Mechanics" (e.g., punctuation, spacing). Qualitative logs show C2 users explicitly rejecting these corrections as "pedantic" or "tone-deaf", feeling the LLM was enforcing a standard corporate style over their personal voice. This aligns with the "Style Homogenization" risks identified in related work.

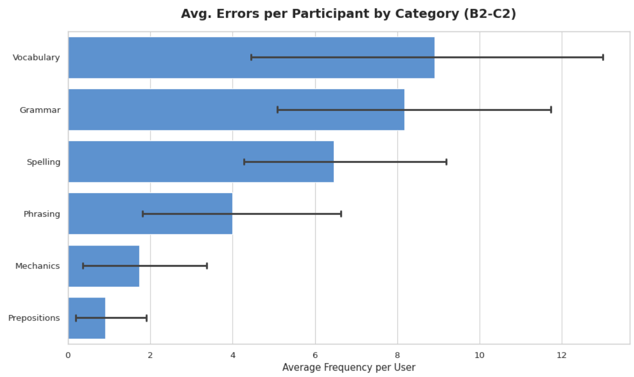


**Figure 6: Average errors per participant. The wide variance in Grammar indicates high individual differences, requiring an adaptive system rather than a static one.**

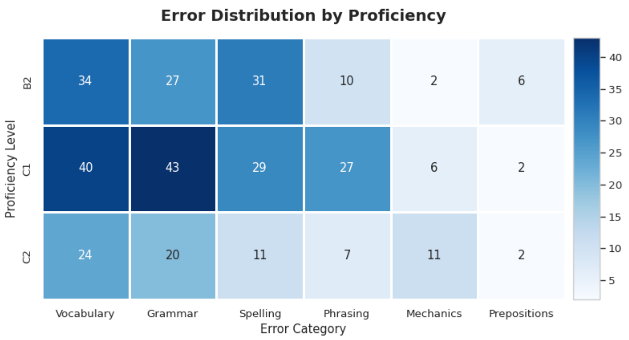


**Figure 7: Error Heatmap by Proficiency. Note the shift: B2 users (top) need help everywhere, while C2 users (bottom) are primarily flagged for "Mechanics", leading to frustration with "pedantic" corrections.**

## 5.6 Writing Efficiency Dynamics

Finally, we examined the relationship between cognitive strain and productivity. We performed a Spearman correlation analysis, a rank-based non-parametric method robust to non-normal distributions, between subjective workload (NASA-TLX) and final text length. We found a statistically significant negative correlation ($\rho = -0.53$ and $p = 0.036$). This indicates a medium-to-strong effect: higher perceived workload directly leads to shorter text production. This confirms that the frustration observed in the **Auto** mode acts as a bottleneck. Figure 8 reinforces this: in interactive modes (associated with lower workload), users produced longer texts with proportionally less editing effort (fewer deletions) compared to the high-friction **Auto** mode.

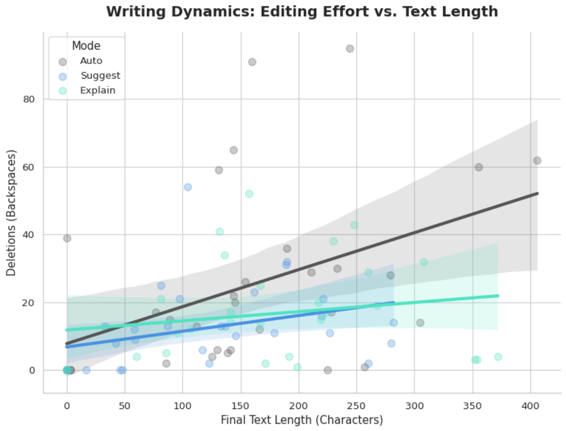


**Figure 8: Writing Dynamics. *Suggest/Explain* modes (Blue/Green) allow for efficient text production, whereas *Auto* mode (Gray) creates a "high effort, low output" trap.**

# 6 DISCUSSION

## 6.1 Limits of Black-Box Automation

The results challenge the assumption that seamless automation improves writing performance. **Auto** mode produced the highest frustration and deletion rates, indicating elevated extraneous cognitive load. When users cannot distinguish their own text from system output, they must continuously verify changes, leading to inefficient interaction. Lack of transparency was perceived as unreliability rather than assistance.

## 6.2 Effectiveness of Suggest Mode

Although **Explain** mode supported learning, **Suggest** mode was most effective for general writing. It reduced extraneous load through familiar UI cues while preserving user control and trust. The more intrusive feedback in **Explain** mode is better suited to targeted pedagogical use (e.g., repeated errors) than as a default, where it risks increasing cognitive load.

## 6.3 Context Sensitivity and Register

Participants reported frequent mismatches between system feedback and intended register (e.g., formal vs. informal). This led to correction loops in which users and the system repeatedly revised each other's output. Future systems should incorporate explicit register controls (e.g., a formality toggle) to align feedback with user intent and reduce stylistic homogenization.

## 6.4 Reflection

Our findings must be interpreted in light of several limitations identified through system behavior and participant feedback.

**What we learned:**

(1) **Correction accuracy, consistency, and transparency.** Across all modes, participants reported inconsistent or incorrect corrections, including missed grammatical errors, inappropriate rephrasing, and contradictory suggestions over time. The system occasionally failed to correct obvious errors or proposed changes that did not respect contextual meaning or register. In **Auto** mode in particular, invisible changes led to confusion and frustration, highlighting limitations in context sensitivity, register awareness, and behavioral stability of the underlying language model.

(2) **Mode-specific interaction limitations.** Each interaction mode introduced distinct usability challenges. **Auto** mode was often described as frustrating due to invisible edits and delayed feedback. **Suggest** mode preserved user control but was perceived as unreliable when no or inappropriate suggestions were provided. **Explain** mode was generally experienced as informative and educational, but some participants reported cognitive overload when many explanations were presented simultaneously, as well as concerns about reduced active engagement and increased reliance on the system.

(3) **Latency and interaction flow.** Response latency negatively affected the writing experience across all modes. Participants reported interruptions of writing flow, boredom, and annoyance when corrections appeared slowly or oscillated between alternatives, making it difficult to disentangle conceptual interaction issues from technical performance limitations.

**What surprised us:**

(1) The extent to which invisible system behavior, particularly in **Auto** mode, undermined user trust and led to avoidance of corrections.

(2) The dual role of **Explain** mode as both a learning support and a source of cognitive overload, depending on explanation density and timing.

**What we would do differently:**

(1) Improve correction stability, contextual appropriateness, and transparency, with a particular focus on making system actions visible and predictable in **Auto** mode.

(2) Refine mode-specific interaction designs to balance user control, learning support, and cognitive load.

(3) Reduce response latency and oscillating corrections to better preserve writing flow.

(4) Extend future studies by recruiting more diverse participant samples and increasing study duration to assess long-term learning, trust development, and dependency risks.

## 7 CONCLUSION

This study moves beyond simple preference data to provide empirical evidence that **opacity is a design flaw** in educational contexts. Our findings challenge the prevailing industry trend toward "seamless" (invisible) automation.
We identified a distinct "Visibility Paradox": while full automation promises speed, it imposes a heavy "cognitive tax". The statistically significant negative correlation ($\rho = -0.53$) between workload and text production proves that **uncertainty acts as a hard limit on user productivity**. When users cannot verify the LLM's actions, they engage in "defensive writing", effectively doubling their editing effort to protect their authorial intent against an invisible agent [2].
We conclude that the future of L2 writing support lies not in smarter algorithms, but in **Adaptive Transparency**. We propose three critical design imperatives:

(1) **Transparency is a Prerequisite for Trust:** Agency cannot exist without visibility. The "Suggest" mode succeeds because it respects the user's role as the final arbiter, reducing the "extraneous load" of monitoring invisible changes [2, 4].

(2) **The "Pedantry" Guardrail:** For expert (C2) users, the LLM's insistence on mechanical perfection often destroys stylistic nuance, leading to "homogenization" [1]. Future systems must detect user proficiency and suppress stylistic corrections to avoid becoming "overzealous".

(3) **From Crutch to Scaffold:** The longitudinal success of the **Explain** mode demonstrates that the LLM can be a tutor rather than just an editor. By surfacing L1-specific interference patterns (e.g., False Friends), the system transforms a momentary error correction into a long-term learning outcome [3].

Ultimately, an effective LLM-based writing assistant must do more than fix errors; it must foster the confidence to write.

## REFERENCES

[1] Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2025. Homogenizing Effect of Large Language Models on Creative Diversity. *Proceedings of the ACL* (2025).

[2] Yang Liu, Min Wang, and Li Zhang. 2025. Developing and Validating the Cognitive Load Scale for AI-Assisted L2 Writing. *Computers & Education* (2025).

[3] J. Matueny and C. Nyamai. 2025. Illusion of Competence and Skill Degradation in Artificial Intelligence Dependency Among Users. *Journal of Educational Technology Systems* (2025).

[4] Oliver Schilke and Martin Reimann. 2025. The Transparency Dilemma: How AI Disclosure Erodes Trust. *Journal of Computer-Mediated Communication* (2025).

[5] Mrinmaya Sharma, Meg Tong, Tomasz Korbak, and Anna Rogers. 2024. Towards Understanding Sycophancy in Language Models. *arXiv preprint arXiv:2310.13548* (2024).