Comprehensive Evaluation of Grammatical Error Correction Systems: Including and Beyond Reference-Based Metrics

Anonymous ACL submission

Abstract

This study addresses three current limitations in Grammatical Error Correction (GEC): the absence of comprehensive evaluation of the newest Large Language Models (LLMs), the reliance on single evaluation metrics for comparative analysis, and the underestimation of system performance by reference-based metrics. We address these limitations first by fine-tuning state-of-the-art LLMs (GPT-4o, LLaMA 3.3 70B) and incorporating these along with zero-shot DeepSeek V3 in an ensemble, which outperformed previous GEC systems in multiple reference-based metrics. We also present the first comprehensive GEC system comparison, evaluating performance across multiple sequence tagging, sequence-to-sequence, and LLM-based approaches using both reference-based and reference-free metrics. Finally, using LLM-as-a-Judge with human validation, we demonstrate that 73.76% of fine-tuned GPT-4o's corrections which did not match the gold reference are either equally valid grammatically or preferred over the gold reference, revealing that reference-based metrics significantly underestimate GEC system performance.

1 Introduction

Effective Grammatical Error Correction (GEC) systems would ideally not only achieve high accuracy but also align closely with human correction patterns. This alignment is particularly crucial in language learning context, where both under-correction (minimal edits) and over-correction can impede effective learning. In order to address this challenge, multiple GEC systems have been developed, which typically fall into three main categories: (1) Sequence Tagging (Omelianchuk et al., 2020); (2) Sequence-to-Sequence (Seq2seq) models, which translate incorrect sentences to correct ones (Yuan and Briscoe, 2016); and (3) Large Language Models (LLMs), which generate corrections through conditional text generation (Omelianchuk et al., 2024; Davis et al., 2024).

However, each approach has inherent limitations as sequence tagging struggles with long-range dependencies (Omelianchuk et al., 2020) and seq2seq models struggle with over-correction (Omelianchuk et al., 2024). These challenges have remained difficult to solve because GEC systems must balance making enough corrections (recall) without making unnecessary ones (precision), while also accounting for grammar's contextual nature. While newer LLM architectures might potentially address some of these limitations, comprehensive evaluations of the newest generation of LLMs (e.g., LLaMA 3.3 70B (Grattafiori et al., 2024), GPT-4o (OpenAI, 2024), and DeepSeek v3 671B (DeepSeek-AI et al., 2025b)) are notably absent. Previous studies (Omelianchuk et al., 2024) on LLMs evaluated older models (e.g., LLaMA-2), leaving practitioners without current data to help inform deployment decisions.

Beyond these challenges, most research relies on a single evaluation metric, typically ERRANT $F_{0.5}$ (Bryant et al., 2017), which evaluates correction quality on individual edit level (Omelianchuk et al., 2024). But there is scarcity of research evaluating different types of GEC systems on higher-level dimensions like fluency or meaning preservation after correction, which can be retrieved using other reference based metrics like GLEU (Napoles et al., 2016) and PT-ERRANT (Gong et al., 2022).

Additionally, reference-based metrics severely penalize systems that generate valid alternative corrections simply because they differ from the provided reference. This limitation is further propelled by LLMs as they introduce stylistic enhancements that frequently diverge from gold references by generating alternative but equally valid corrections (Bryant et al., 2023). Alternatively, the creation of comprehensive datasets containing all possible valid corrections for each error remains practically infeasible due to the combinatorial explosion of grammatical alternatives, contextual variations, and stylistic preferences. To address this, reference-free metrics, like IMPARA (Maeda et al., 2022), are developed to directly assess semantic and grammatical acceptability without references, but they often exhibit bias towards minimal edits and struggle with comprehensive corrections that make multiple changes.

To address these challenges, first, we conducted experiments with latest LLMs (LLaMA 3.3 70B, GPT-4o, and DeepSeek v3) in both zero-shot and fine-tuned settings. Second, we developed an ensemble architecture by combining our best models using sentence-level majority voting, with an innovative fallback strategy that selects corrections with the highest n-gram overlap across candidate corrections when there is no majority. Third, instead of using one metric for evaluation, we present the first multi-metric comparative analysis across multiple sequence tagging, seq2seq, and LLM-based GEC approaches, utilizing both reference-based and reference free metrics. This helps us capture different dimensions of correction quality, including edit accuracy and fluency. Finally, to address the fundamental reference-based evaluation problem, we conduct LLM-as-a-Judge evaluation with human validation to properly estimate our system's performance[1]. Our key contributions are:

- A fine-tuned GPT-4o model established a new state-of-the-art for individual GEC systems, an achievement further validated by our top-ranking performance in XXX Shared Task[2].
- A majority voting ensemble with n-gram overlap fallback that further advances the state-of-the-art on ERRANT $F_{0.5}$ and PT-ERRANT.
- First extensive GEC system comparison with reference-based and reference-free metrics.
- A hybrid evaluation framework combining LLM-as-a-Judge with human evaluation, revealing that 73.76% of our model's corrections that differ from gold standards are actually equally valid or preferred, demonstrating limitations in reference-based metrics.

[1] All code, fine-tuned models and annotated data will be made available at xxx

[2] The shared task link has been anonymized for review

2 Related Work

This section examines the evolution of GEC systems from sequence tagging to LLMs, their ensemble methods and evaluation challenges.

2.1 Individual GEC Systems

Sequence tagging GEC approaches like GECToR (Omelianchuk et al., 2020) use pre-trained encoders (BERT, RoBERTa, XLNet) to predict specific edit operations for each token from a large vocabulary of transformation tags, with larger version of these encoders giving better performance (Tarnavskyi et al., 2022). However these approaches struggle with complex, interconnected errors which requires broader contextual understanding.

On the other hand, Seq2seq models approach GEC as a translation task from incorrect to correct sentences. These systems, like Neural Machine Translation has evolved over time from attention-enhanced RNNs (Yuan and Briscoe, 2016) to CNNs (Chollampatt and Ng, 2018) and eventually to Transformer-based models (T5) (Rothe et al., 2021). They can capture complex error patterns which overcame the shortcomings of previous Seq2seq systems (Flachs et al., 2019), but they suffered from computational demands and over-correction tendencies.

Distinct from previous methods, LLMs approach GEC through conditional text generation, leveraging knowledge acquired during pre-training. In zero-shot settings, models like GPT-3.5 excel at error detection but exhibit low precision due to unnecessary corrections (Fang et al., 2023), with GPT-3.5, GPT-4, and LLaMA-2 variants all achieving ERRANT $F_{0.5}$ scores below 50 on the BEA-dev set (Omelianchuk et al., 2024). Fine-tuning dramatically improves performance, with LLaMA-7B and 13B models reaching competitive $F_{0.5}$ scores of 55.4 and 56.4 respectively which is comparable to specialized seq2seq and sequence tagging systems on the same benchmark (Omelianchuk et al., 2024).

2.2 Ensembles of GEC Systems

While these individual models show promising results, ensemble approaches consistently outperform them. Qorib et al. (2022) introduced Edit-based System Combination (ESC) using logistic regression on extracted edit features to select corrections, while Tarnavskyi et al. (2022) extended this with span-level majority voting. Qorib and Ng (2023) developed GRECO, a DeBERTA-based grammaticality scorer for re-ranking grammatical corrections but their systems are optimized to perform well in specific datasets, rather than a single generalizable model. Building on these, Omelianchuk et al. (2024) analyzed various ensembling techniques (GRECO, GPT-based ranking, and second-order ensembles), achieving state-of-the-art performance (62.9 $F_{0.5}$ on BEA-dev) through a simple edit-span level majority voting ensemble, without any dataset specific optimization.

2.3 Evaluation of GEC Systems

Reference-based metrics like ERRANT, GLEU and PT-ERRANT evaluate corrections by comparing them with gold references. But these metrics underestimate system performance by failing to account for the diversity of valid corrections. Rozovskaya and Roth (2021) demonstrated that system scores improved by 20–40 points when references were adjusted to accept any valid correction. Reference-free metrics like IMPARA (Maeda et al., 2022) attempt to overcome these limitations by directly measuring semantic and grammatical acceptability of corrections without requiring gold standards. However, they are biased toward minimal edits and have difficulty in scoring comprehensive corrections (Bryant et al., 2023). Human evaluation remains the ideal option but is often impractical at scale. To bridge this gap, LLM-as-a-Judge approaches (Gu et al., 2025) have emerged as promising alternatives, as they evaluate correction quality with inter-annotator agreement comparable to human evaluators. However, using a single LLM for evaluation risks introducing biases that could compromise assessment objectivity and reliability.

3 Methodology

To address these challenges, we experiment with newest LLMs and multiple ensemble architectures, conduct the first comprehensive multi-metric GEC system comparison, and utilize a hybrid LLM-human framework to validate corrections.

3.1 Dataset

For our experiments, we utilized the W&I+LOCNESS dataset from the BEA 2019 Shared Task on GEC (Bryant et al., 2019; Granger, 1998), organized by CEFR levels: A (beginner), B (intermediate), C (advanced), and N (native). We fine-tuned models on the ABC train partition (combining beginner, intermediate, and advanced texts) and used the ABCN development set (including native texts) as our test set.

While several established GEC benchmark datasets exist, including CoNLL-14 (Ng et al., 2014), JFLEG (Napoles et al., 2017), and FCE (Yannakoudakis et al., 2011), our evaluation is intentionally limited to W&I+LOCNESS due to practical constraints. Comprehensive multi-dataset evaluations significantly increase computational demands when evaluating multiple LLM variants, while our hybrid LLM-human evaluation framework introduces substantial cost considerations that would multiply across test sets. This is why we allocated resources towards more thorough model comparisons on a single standardized benchmark.

3.2 Model Selection and Fine-tuning

We evaluated three leading LLMs representing diverse architectures and accessibility paradigms: GPT-4o (commercial, OpenAI), LLaMA-3.3-70B-Instruct (open-source, Meta), and DeepSeek-V3-671B (Mixture-of-Experts architecture activating only 37B of 671B parameters per token).

The W&I LOCNESS dataset contains original sentences annotated using ERRANT (ERRor ANnotation Toolkit) (Bryant et al., 2017), which standardizes error annotation by aligning source-corrected text pairs, extracting edits, and classifying them into specific edit types. We parsed these ERRANT annotations to extract the specified edit operations and applied them to create gold references for fine-tuning.

We first evaluated all three models in a zero-shot setting using the prompt in Figure 2 in the Appendix, to establish a baseline for each model's inherent GEC capabilities without additional training. Following this, we fine-tuned two of the three models (GPT-4o[3] and LLaMA 3.3) on the BEA training data for two epochs with cross-entropy loss function and using the same prompt template as in zero-shot inference. DeepSeek-V3 was maintained in its zero-shot configuration due to the prohibitive computational costs of fine-tuning a 671B-parameter model and its superior baseline performance, making it valuable for ensemble integration without fine-tuning.

3.3 Ensemble System

We experimented with four ensemble systems combining the outputs from our two fine-tuned mod-

[3]GPT-4o fine-tuned using OpenAI API: https://openai.com/index/gpt-4o-fine-tuning/

els (GPT-4o and LLaMA-3.3) and the zero-shot DeepSeek-V3 model. All our ensemble variants use a majority voting mechanism as their primary decision rule: when two or three systems agree on a correction, the agreed correction is applied. However, the ensembles differ in their fallback strategy for cases where no agreement exists among the three models:

Best Model Fallback: Our baseline ensemble relies on the best model's correction as fallback when no majority exists. This approach is based on the assumption that the best model is most likely to produce the optimal correction when consensus cannot be reached.

Qwen Fallback: This ensemble uses Qwen-2.5-7B-Instruct (Qwen et al., 2025) as a meta-model to select among candidate corrections when models disagree, using the prompt in Figure 3 in Appendix. This approach leverages Qwen's language understanding to make informed grammatical judgments when models disagree, rather than relying on simple statistical measures.

Perplexity Fallback: When models disagree, this ensemble computes the perplexity score of each candidate correction using base Qwen-7B[4] and selects the correction with the lowest perplexity. This approach is based on the fundamental nature of LLMs, which are pre-trained primarily on grammatically correct text, causing them to assign higher probabilities (lower perplexity) to corrections that follow grammatical patterns they have encountered during pre-training.

N-gram Fallback: Our final ensemble resolves disagreements by selecting the correction with the highest n-gram overlap with other candidate corrections. This approach operates on the principle that correct edits often share common subsequences across different models' outputs, even when full agreement isn't reached.

3.4 Automated Metrics Evaluation

For comprehensive evaluation of GEC systems of different categories, we use three reference-based metrics (ERRANT $F_{0.5}$, GLEU and PT-ERRANT) to assess correction quality against gold standards. ERRANT $F_{0.5}$ focuses on precision, ensuring our system makes accurate corrections without making unnecessary edits. GLEU assesses overall correction quality through n-gram matching with reference corrections, indicating how well the model produces naturally fluent corrections that align with human judgment (Equations provided in Appendix A.1). PT-ERRANT uses pre-trained models to evaluate phrase-level and structural modifications with semantic understanding, specifically showing a model's ability to preserve intended meaning while making corrections. Alongside reference-based metrics, we also incorporate reference-free metrics like IMPARA, which evaluates correction quality without gold standards.

[4]https://huggingface.co/Qwen/Qwen-7B

3.5 LLM-as-a-Judge and Human Evaluation

To move beyond purely quantitative metrics in assessing grammatical corrections, we introduce a hybrid evaluation framework that combines LLM-as-a-Judge with targeted human evaluation. This is a core contribution of our work as it addresses critical limitations in automated GEC evaluation approaches while validating the reliability of our systems and quality of its produced corrections.

Our hybrid evaluation protocol first employs two state-of-the-art LLMs, Claude 3.7 Sonnet[5] and DeepSeek-R1 as primary judges to assess whether corrections from our best-performing fine-tuned model are preferred over gold references. We selected these specific models based on Claude's demonstrated high inter-annotator agreement with human evaluators (Zheng et al., 2023) and DeepSeek-R1's powerful reasoning capabilities (DeepSeek-AI et al., 2025a).

For each edit, the LLM judges categorize comparisons into one of three categories: (1) The gold reference is preferred, (2) The model's correction is preferred, or (3) Both corrections are equally grammatically valid (even if syntactically different). When both models reach consensus on a preference, their determination is considered final, as LLMs are known to have similar inter-annotator agreement to humans (Gu et al., 2025). In cases where the LLM judges disagree, two qualified human GEC evaluators apply the same three category framework to resolve discrepancies.

This evaluation framework offers two key advantages: (1) Substantially reduced time and resources compared to comprehensive human assessment, as human judgment is invoked only for contested cases, making the evaluation approach highly scalable; (2) Improved reliability through two-model consensus mechanisms that mitigate individual LLM biases.

[5]https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf

4 Results and Analysis

As demonstrated in Table 1, our approach achieves state-of-the-art performance on all 3 reference-based metrics (ERRANT $F_{0.5}$, GLEU, and PT-ERRANT), surpassing existing sequence tagging, seq2seq, and LLM-based GEC approaches.

4.1 Analysis of Individual Models

Upon initial examination, Table 1 reveals that fine-tuned GPT-4o achieved the highest overall performance among all individual models tested. Additionally, it demonstrates that fine-tuning gives substantial performance gains, with GPT-4o improving by 22.07 points in ERRANT $F_{0.5}$ and LLaMA by 24.94 points. This significant gap confirms that fine-tuning remains essential for competitive GEC performance, even as base model sizes of LLMs continue to increase.

Among zero-shot models, DeepSeek outperforms both GPT-4o and LLaMA across all metrics, likely due to its mixture-of-experts (MoE) architecture. Unlike its competitors' dense transformer design, DeepSeek selectively activates specialized sub-networks for each token, possibly enabling it to better handle diverse linguistic patterns, while also making it faster in inference for GEC applications. On the other hand, GPT-4o is outperformed by both open-source LLMs in zero-shot setting, particularly DeepSeek. Furthermore, after fine-tuning, GPT-4o's modest advantage over LLaMA demonstrates the competitiveness of open-source models, which offer greater accessibility and transparency to the broader community.

4.2 Analysis of Ensemble Systems

The Majority Voting with N-gram Fallback (where N=3[6]) ensemble outperforms all individual models and other ensemble approaches, achieving the highest ERRANT $F_{0.5}$ score of 0.6623 and PT-ERRANT score of 0.7122. This suggests it is best at effectively balancing grammatical correctness with semantic meaning preservation. Furthermore, this fallback approach outperforms simply defaulting to GPT-4o when models disagree, proving that selecting corrections based on maximum subsequence agreement between candidate outputs yields better results than relying solely on the strongest individual model. However, fine-tuned GPT-4o still maintains the highest GLEU score (0.8400), indicating its corrections are the most fluent among all the GEC systems we tested.

The Qwen and Perplexity Fallback ensembles underperform compared to N-gram Fallback because they make decisions based on external criteria (LLM judgments or fluency scores) rather than analyzing patterns of overlap among the candidate corrections. Their superior performance on IMPARA, a metric known to favor minimal edits (explained further in Section 4.3), suggests these external evaluation criteria inherently prioritize conservative corrections over comprehensive ones even if comprehensive corrections are necessary.

4.3 Comparative Analysis of Existing Systems

Our N-gram Fallback Ensemble model advances the state-of-the-art in GEC by surpassing all existing systems on all three reference-based metrics (ERRANT $F_{0.5}$, PT-ERRANT, and GLEU), with even our individual fine-tuned GPT-4o model outperforming all existing systems on ERRANT $F_{0.5}$ and GLEU. This strong performance across these metrics indicates our approach makes precise and fluent corrections, with minimal unnecessary edits, while preserving intended meaning. This highlights both the effectiveness of our ensemble strategy and remarkable capabilities of our individual models.

Despite these impressive results with fine-tuned models, Table 1 reveals state-of-the-art LLMs, in its base form, under-perform compared to established seq2seq and sequence tagging approaches, and even fine-tuned significantly smaller LLMs, across all reference-based metrics. However, our fine-tuned models outperform LLaMA 2 variants by around 10 percentage points in ERRANT, reflecting architecture advances in newer LLMs. Compared to Seq2Seq models, we see larger gaps in ERRANT and PT-ERRANT but closer GLEU scores, suggesting decoder-only architectures excel at precise corrections while encoder-decoder models still maintain competitive fluency. The largest performance gap exists against sequence tagging systems, demonstrating that token-level edit prediction through sequence labeling is less effective than large-scale pre-training with fine-tuning for comprehensive GEC. While our systems achieve strong IMPARA scores, we emphasize more on other metrics due to IMPARA's documented bias toward minimal edits and tendency to penalize comprehensive corrections.

Our systems' performance was further validated in the Shared Task, where our ensemble with GPT-

[6] N=3 produced best performance among N={2,3,4}. Higher N values will make it biased towards minimal edits

Model	ERRANT $F_{0.5}$	GLEU	PT-ERRANT	IMPARA
Our Individual Models				
Fine-Tuned GPT-4o	0.6599	**0.8400**	0.7064	0.7768
Fine-Tuned LLaMA 3.3	0.6420	0.8281	0.6842	0.7705
Base DeepSeek	0.4926	0.7677	0.5666	0.7754
Base GPT-4o	0.4592	0.7420	0.5484	0.7564
Base LLaMA 3.3	0.4826	0.7345	0.4973	0.7122
Our Ensemble Systems				
Majority Voting + GPT-4o Fallback	0.6607	0.8392	0.7039	0.7746
Majority Voting + Qwen Fallback	0.6249	0.8312	0.6905	0.7855
Majority Voting + Perplexity Fallback	0.6251	0.8304	0.6879	0.7858
Majority Voting + N-gram Fallback	**0.6623**	0.8347	**0.7122**	0.7749
LLMs (Decoder Only Transformers)				
Fine-Tuned LLaMA 2 7B (Omelianchuk et al., 2024)	0.5530	0.7985	0.6157	0.7529
Fine-Tuned LLaMA 2 13B (Omelianchuk et al., 2024)	0.5640	0.8027	0.6399	0.7554
Seq2Seq Models (Encoder-Decoder Transformer)				
Fine-Tuned T5 11B (Omelianchuk et al., 2024)	0.5860	0.8231	0.6656	0.7629
Fine-Tuned FLAN 20B (Omelianchuk et al., 2024)	0.5770	0.8149	0.6630	0.7582
Sequence Tagging Systems				
GeCTOR (XLNet) (Omelianchuk et al., 2020)	0.5630	0.7687	0.6248	0.7058
CTC-Copy (Zhang et al., 2023)	0.5270	0.7714	0.6096	0.7302
EditScorer (Sorokin, 2022)	0.5740	0.7565	0.6285	0.7072
Ensemble and Model Ranking Approaches				
Ensemble Best 7 (Omelianchuk et al., 2024)	0.6290	0.7854	0.7040	0.7153
Ensemble Best 3 (Omelianchuk et al., 2024)	0.6250	0.7907	0.7000	0.7216
GRECO Rank 7 (Omelianchuk et al., 2024)	0.6200	0.8032	0.7084	0.7366
GPT-4 Rank 3 (Omelianchuk et al., 2024)	0.5810	0.8270	0.6654	0.7753
Shared Task Competing Systems				
Sugiyama et al. (2025)	0.4283	0.7603	0.4761	**0.8171**
Gotō et al. (2025)	0.6189	0.7597	0.6483	0.6987

Table 1: Performance comparison of our models against GEC Systems on BEA-dev dataset. Existing systems results obtained from Omelianchuk et al. (2024) (https://github.com/grammarly/pillars-of-gec/tree/main/data/system_preds), except shared task systems which was provided by organizers. As stated by (Omelianchuk et al., 2024), Ensemble Best 7 includes all 7 models (Fine-T LLaMA 2 7B, Fine-Tuned LLaMA 2 13B, Fine-Tuned T5 11B, Fine-Tuned FLAN 20B, GeCTOR (XLNet), CTC-Copy, and EditScorer); Ensemble Best 3 contains top three: LLaMA-2-13B-Fine-Tuned, FLAN-20B, and LLaMA-2-7B-Fine-Tuned; GRECO Rank 7 uses quality estimation guided beam search to combine edits from 7 models; GPT-4 Rank 3 ranks outputs from best models of each type: LLaMA-2-13B-Fine-Tuned (LLMs), T5-11B (Seq2seq), and EditScorer (Sequence Tagging).

4o Fallback and Fine-tuned GPT-4o model ranked highest across all three reference-based evaluation metrics, as shown in Table 1 (We did not submit our best ensemble because it was implemented after the shared task concluded). Among other notable systems, Sugiyama et al. (2025) used zero-shot GPT-4o, while Gotō et al. (2025) implemented an ensemble approach using the GECToR framework, combining three fine-tuned encoders (RoBERTa-large, XLNet-large-cased, and DeBERTa-v1-large) with majority voting.

One of the objectives of the shared task was to examine GEC evaluation metric vulnerabilities, particularly how reference based metrics can unfairly penalize valid corrections that differ from the reference. And how reference free metrics like IMPARA will prioritize making minimal edits and penalize comprehensive corrections even if they are necessary. As seen in Table 2, first couple of corrections are penalized simply because it is different from the gold reference, even though both are equally grammatically correct. And these kinds of linguistic differences are common in prepositional (on/at) and verb form edits (are suffering/-suffer), as shown in the Table. This demonstrates the inherent limitations of relying exclusively on reference-based metrics for evaluating GEC system performance.

To address these limitations reference free metrics like IMPARA was developed. But our analysis, which was also supported by Sugiyama et al. (2025) and Gotō et al. (2025), reveals significant biases in IMPARA, which favors minimal edits while severely penalizing essential comprehensive corrections. As demonstrated in Table 2 sentences requiring multiple corrections receive extremely low scores regardless of quality and necessity of the correction. This limitation stems from IMPARA's

Original	Gold Reference	Ensemble Correction	ERRANT	GLEU	PT-ERRANT
I had a wonderful day yesterday because I was in the beach all the afternoon .	I had a wonderful day yesterday because I was on the beach all afternoon .	I had a wonderful day yesterday because I was at the beach all afternoon .	50.00	73.05	63.88
In our modern world , many people are suffered from stress that spring from life conditions .	In our modern world , many people are suffering from stress that springs from life conditions .	In our modern world , many people suffer from stress that springs from life conditions .	50.00	76.60	62.29

Original	Ensemble Correction	IMPARA
I take too much photos because you don't visit places like that everyday.	I take too many photos because you don't visit places like that every day.	0.0006
What if you don't have none of those requierments?	What if you don't have any of those requirements?	0.0116

Table 2: GEC examples with evaluation metrics. Here Ensemble is the Majority Voting with GPT-4o Fallback. Incorrect phrases in Red, their corresponding correction in Green

BERT-based architecture, which measures semantic similarity via vector distances. Multiple edits push original and corrected sentences too far apart in embedding space, causing IMPARA to penalize them. Thus, it can be argued that IMPARA proves unreliable as a GEC evaluation metric.

4.4 LLM-as-a-Judge and Human Evaluation

To address these metric limitations, we implemented a hybrid evaluation combining LLM-as-a-Judge with human assessment to compare our fine-tuned GPT-4o corrections against gold standards. Table 3 shows the two LLM judges (reaching consensus in 64.34% of cases) preferred our model's corrections (30.87%) over gold standards (19.77%), with 13.70% rated the corrections equally valid. Following human evaluation to resolve LLM disagreements, final results revealed that GPT-4o's corrections were preferred for 35.61% of edits, gold standards for 26.34%, while 38.15% of edits were judged equally grammatically valid.

So, in 73.76% of cases, where our model produced corrections which were different from gold standards, these are actually judged as either superior or equally valid compared to gold references. This further validates that our system not only outperforms existing GEC approaches on automated metrics, but also produces corrections that are frequently preferred over or considered the same as the gold standards themselves. This complementary evaluation brings forward an argument that while reference based metrics remain valuable for standardized comparison with existing systems, they should be supplemented with human judgment (assisted by LLM judges for scalability purposes) to comprehensively assess correction quality.

4.5 Edit Type Analysis

Figure 1 shows the distribution of edit types where GPT-4o corrections are different from gold standards, and their subsequent evaluation through LLM-as-a-Judge and human validation. The most common differences between GPT-4o corrections and gold standards are in replacement edits indicating our model often chooses different, yet valid, replacement strategies when correcting the same underlying grammatical issues. Presence of significant punctuation edits in GPT-4o corrections indicate our model often adds marks absent in gold corrections (M:PUNCT), preserves punctuation that gold standard removes (U:PUNCT), and selects alternative punctuation marks (R:PUNCT). Determiner errors (189 instances across unnecessary, replacement, and missing articles) reflect GPT-4o's distinct handling of English articles, a grammatical feature that allows multiple acceptable forms in any given context.

After these correction edits are evaluated by human evaluators and LLM judges, it can be seen from Figure 1 that our model's punctuation additions (M:PUNCT) are strongly preferred over gold standard (130 vs 9 instances), while gold standard reference is preferred when GPT-4o omitted some necessary punctuation (U:PUNCT). This suggests GPT-4o excels at identifying missing punctuation but occasionally makes incorrect omissions. Evaluators also preferred our model's orthography corrections (R:ORTH) and determiner choices, with determiner edits from our model consistently rated superior to or at least equal to gold standard alternatives.

Additionally, 38.15% of edits across all error types were judged equally valid, revealing reference-based metrics penalize alternatives sim-

Evaluation Stage	Gold Standard Preferred	GPT-4o Preferred	Both Equally Valid	Disagreement
LLM Consensus	342 (19.77%)	534 (30.87%)	237 (13.70%)	617 (35.66%)
After Human Resolution	454 (26.24%)	616 (35.61%)	660 (38.15%)	-

Table 3: LLM-as-a-Judge and Human Evaluation Results (In the BEA-dev dataset, there were 1,730 edits in total where fine-tuned GPT-4o produced different edit corrections compared to gold reference)

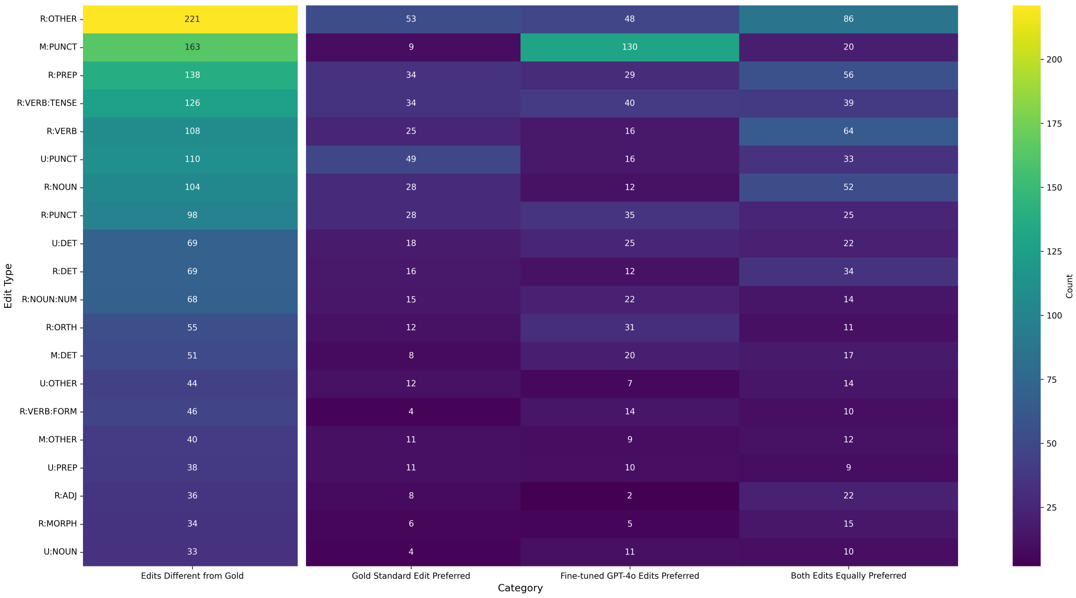

Figure 1: Edit Type Counts where fine-tuned GPT-4o produced different corrections compared to gold standard. Table 4 in the Appendix provides detailed descriptions of all edit type codes.

ply for differing from the gold reference. This limitation is most evident in replacement operations related to prepositions (R:PREP) and verbs (R:VERB), for example. This highlights the inherent flexibility of English preposition usage where multiple alternatives can be grammatically correct and also showcases how verb usage is highly context dependent, with multiple tense or form options often being grammatically acceptable.

This provides evidence that for certain grammatical features, especially punctuations, determinants, prepositions, and verb tense, there is bound to be multiple valid corrections, even if we strictly follow the approach of minimal edits. These findings substantiate the argument that reference based metrics exclusively does not provide a good estimation of GEC system performance, but rather it requires to be supported by human evaluation to obtain a more accurate representation of correction quality. Appendix A.5 further provides examples where GPT-4o edits are different from gold reference.

5 Conclusion

Our experiments reveal that fine-tuned LLMs significantly outperform traditional GEC approaches, with our fine-tuned GPT-4o model establishing a new state-of-the-art for individual systems, surpassing previous benchmarks across a couple of reference-based metrics. This performance advantage is further extended by our majority voting ensemble with N-gram overlap fallback, which achieves even higher scores on ERRANT $F_{0.5}$ and PT-ERRANT. However, reference-based metrics systematically underestimate system performance by penalizing legitimate alternatives, as demonstrated by our hybrid LLM-human evaluation framework, which reveals that 73.76% of corrections diverging from gold standards are judged equally valid or superior. Through systematic error type analysis, we provide empirical evidence that certain grammatical features, particularly punctuation, determiners, prepositions, and verb forms inherently support multiple valid corrections, further challenging the exclusive use of reference-based metrics for assessing GEC performance. These findings underscore the necessity of complementing automated metrics with human evaluation, potentially aided by LLMs for scalability, to accurately assess GEC system performance.

6 Limitation

While our approach demonstrates advancement in GEC performance, several important limitations should be considered when interpreting our results. The effectiveness of our fine-tuned models is contingent on the error type distribution in the training data. Despite achieving state-of-the-art performance, our models may excel at correcting error types well-represented in the ABC train set of W&I+LOCNESS dataset while potentially underperforming on less common error types.

One of the limitation is related to potential data contamination. LLMs may have encountered our test datasets during pre-training, artificially inflating performance metrics. We cannot fully eliminate this possibility without proper knowledge of the training data used during pre-training. Future work could address this by creating entirely new test sets with recent content or techniques to detect contamination effects.

While our zero-shot experiments included DeepSeek V3 (671B parameters), we could not host and fine-tune this model due to prohibitive infrastructure requirements. In contrast, we successfully fine-tuned GPT-4o via OpenAI's API without hosting the model locally, and adapted LLaMA 70B using parameter-efficient methods like LoRA. However, DeepSeek's superior zero-shot performance suggests it might have established an even higher benchmark if fine-tuned. This highlights an important trade-off between model size, accessibility, and performance in GEC research.

For our LLM-as-a-Judge approach with human verification, we limited our human evaluation to GPT-4o corrections due to resource constraints, as this model demonstrated the strongest overall performance in automated metrics.

Our evaluations were conducted on standardized academic datasets. Performance may vary in real-world applications with domain-specific writing styles, specialized terminology, or less common error patterns not represented in the evaluation data.

Furthermore, our research focuses exclusively on English grammatical error correction. The architectures, fine-tuning approaches, and evaluation frameworks may not directly transfer to other languages, particularly those with significantly different grammatical structures, morphological complexity, or writing systems. This limitation is particularly relevant given the global need for grammatical error correction across diverse languages.

Finally, while grammatical error correction systems primarily aim to assist users in improving their writing, several ethical considerations merit acknowledgment. Our evaluation framework and models may embed normative assumptions about "correct" grammar that could disadvantage speakers of non-standard English dialects. The Large Language Models employed in this study (GPT-4o, LLaMA 3.3, and DeepSeek V3) may perpetuate linguistic biases present in their training data, potentially resulting in corrections that privilege certain language varieties over others. Additionally, while our dual-LLM-as-a-Judge evaluation approach helps mitigate individual model biases, residual biases from each model may still separately influence which corrections are deemed "equally valid" or "preferred" compared to gold references. We also acknowledge that widespread deployment of automated GEC systems could influence language standardization in ways that require careful consideration by the research community.

7 Ethics Statement

This research utilizes established public benchmark datasets with appropriate consent and no additional personal data collection. We also recognize that our GEC systems may embed assumptions about "correct" grammar that could disadvantage non-standard English dialects, supported by our finding that 73.76% of model corrections differing from gold standards were equally valid or preferred highlights the inherent flexibility in grammatical correctness. Furthermore, the LLMs employed may perpetuate linguistic biases from their training data. Additionally, human evaluators for our LLM-as-a-Judge framework participated voluntarily with appropriate compensation, and all evaluator identities were anonymized in research records.

Our work intends to enhance educational practices by providing supplementary tools for grammar assessment rather than substituting expert human evaluation. The GEC systems developed here are designed primarily for instructional feedback and learning support, and we caution against their use in critical assessment scenarios without substantial human supervision and review.

References

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings* *of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 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)*, pages 793–805. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, page 1–59.

Shamil Chollampatt and Hwee Tou Ng. 2018. Neural quality estimation of grammatical error correction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2528–2539, Brussels, Belgium. Association for Computational Linguistics.

Christopher Davis, Andrew Caines, Øistein E. Andersen, Shiva Taslimipoor, Helen Yannakoudakis, Zheng Yuan, Christopher Bryant, Marek Rei, and Paula Buttery. 2024. Prompting open-source and commercial language models for grammatical error correction of English learner text. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11952–11967, Bangkok, Thailand. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *Preprint*, arXiv:2304.01746.

Simon Flachs, Ophélie Lacroix, and Anders Søgaard. 2019. Noisy channel for low resource grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 191–196, Florence, Italy. Association for Computational Linguistics.

Peiyuan Gong, Xuebo Liu, Heyan Huang, and Min Zhang. 2022. Revisiting grammatical error correction evaluation and beyond. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6891–6902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Takumi Gotō, Kōsuke Doi, Adam Nohejl, Justin Vasselli, Sei-shi Gōhara, Yūsuke Sakai, and Tarō Watanabe. 2025. Nice glittchers: Bunpō ayamari teisei bumon [grammatical error correction department]. In *XXX*. XXX. Original title contains Japanese text.

Sylviane Granger. 1998. The computer learner corpus: a versatile new source of data for sla research. In Sylviane Granger, editor, *Learner English on Computer*, pages 3–16. Routledge, London.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Koki Maeda, Masahiro Kaneko, and Naoaki Okazaki. 2022. Impara: Impact-based metric for gec using parallel data. In *International Conference on Computational Linguistics*.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2016. Gleu without tuning. *arXiv preprint arXiv:1605.02592*.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational*

Applications, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Korniienko, and Igor Samokhin. 2024. Pillars of grammatical error correction: Comprehensive inspection of contemporary approaches in the era of large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 17–33, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o System Card. `https://openai.com/index/gpt-4o-system-card/`. Accessed: 2025-04-18.

Muhammad Reza Qorib, Seung-Hoon Na, and Hwee Tou Ng. 2022. Frustratingly easy system combination for grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1964–1974, Seattle, United States. Association for Computational Linguistics.

Muhammad Reza Qorib and Hwee Tou Ng. 2023. System combination via quality estimation for grammatical error correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12746–12759, Singapore. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2021. How good (really) are grammatical error correction systems? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2686–2698, Online. Association for Computational Linguistics.

Alexey Sorokin. 2022. Improved grammatical error correction by ranking elementary edits. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11416–11429, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Seiji Sugiyama, Taku Morioka, Junya Takayama, and Tomoyuki Kajiwara. 2025. ehimetrick: Jidō hyōka hakku shared task tasuku bunpō ayamari teisei bumon [ehimetrick: Automated assessment hack shared task on grammatical error correction section]. In *XXX*. XXX. Original title contains Japanese text.

Maksym Tarnavskyi, Artem Chernodub, and Kostiantyn Omelianchuk. 2022. Ensembling and knowledge distilling of large sequence taggers for grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3842–3852, Dublin, Ireland. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 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*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Zheng Yuan and Ted Briscoe. 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*, pages 380–386, San Diego, California. Association for Computational Linguistics.

Yu Zhang, Yue Zhang, Leyang Cui, and Guohong Fu. 2023. Non-autoregressive text editing with copy-aware latent alignments. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7075–7085, Singapore. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Appendix

A.1 Equations of Automated Metrics

The equation for ERRANT $F_{0.5}$ is given below:

$$\text{ERRANT } F_{0.5} = \frac{1.25 \times \text{precision} \times \text{recall}}{0.25 \times \text{precision} + \text{recall}} \tag{1}$$

The equation for GLEU is given below:

$$\text{GLEU} = \min\left(1, \exp\left(1 - \frac{r}{c}\right)\right) \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2}$$

where r is reference length, c is candidate length, N is the maximum n-gram order, w_n are n-gram weights, and p_n is the modified n-gram precision.

A.2 Prompt for GEC Inference and Fine-tuning

Figure 2 shows the prompt used for fine-tuning and also for generating the corrections in inference.

Prompt

You are an English linguist and your task is to correct the grammatical and mechanical errors in English sentences.
Please make only necessary corrections to the extent that a sentence will be free from errors and comprehensible.
Do not alter word choices unnecessarily (e.g., replacing words with synonyms) or make stylistic improvements.
Also, the sentences are tokenized, which means punctuation marks are separated from the English words by spaces.
When returning the corrected sentences, please use the same tokenized format.
Please respond in the following JSON format:
{{
"corrected": "..."
}}
The original sentence is:
{original}

Figure 2: Inference and Fine-tune Prompt

A.3 Prompt for Qwen Meta-Model

Figure 3 shows the prompt used for the Qwen-2.5-7B-Instruct model to judge which correction out of the three model correction is better.

A.4 ERRANT Error Type Descriptions

Table 4 provides descriptions of the ERRANT edit types referenced in Figure 1.

A.5 Examples of Edit Differences

Tables 5 and 6 show examples of instances where the correction edits generated by GPT-4o model is different from the gold references, and the verdict and preference explanation of the LLM and human judges.

Prompt

Compare the sentences given below and tell me which one (A, B, or C) is the most grammatically correct version of the original given below.
Original: "{original_sentence}"
A: "{correction_a}"
B: "{correction_b}"
C: "{correction_c}"

Provide your response in JSON format as follows:
{
"best_option": "The letter of the best option (A, B, C, etc.)",
"reasoning": "A brief explanation of why this is the best option",
}

Figure 3: Prompt used for Qwen to judge which model correction is best

Edit Code	Description
M:PUNCT	Missing punctuation - punctuation present in the correction but absent in gold standard
U:PUNCT	Unnecessary punctuation - punctuation present in gold standard but omitted in correction
R:PUNCT	Replacement of punctuation - different punctuation used in correction compared to gold
R:PREP	Replacement of preposition - different preposition used in correction compared to gold
R:VERB	Replacement of verb - different verb used in correction compared to gold
R:VERB:TENSE	Replacement of verb tense - different tense of the same verb used in correction
R:VERB:FORM	Replacement of verb form - different form of the same verb used in correction
R:NOUN	Replacement of noun - different noun used in correction compared to gold
R:NOUN:NUM	Replacement of noun number - singular/plural variation of the same noun
U:DET	Unnecessary determiner - determiner present in gold standard but omitted in correction
R:DET	Replacement of determiner - different determiner used in correction compared to gold
M:DET	Missing determiner - determiner present in the correction but absent in gold standard
R:ORTH	Replacement of orthography - spelling or formatting differences
R:OTHER	Other replacements not falling into above categories
U:OTHER	Other unnecessary words not falling into above categories
M:OTHER	Other missing words not falling into above categories
R:ADJ	Replacement of adjective - different adjective used in correction compared to gold
U:PREP	Unnecessary preposition - preposition present in gold standard but omitted in correction
R:MORPH	Replacement of morphology - different word form used in correction compared to gold
U:NOUN	Unnecessary noun - noun present in gold standard but omitted in correction

Table 4: ERRANT Edit Type Descriptions

Original	Gold Reference	GPT-4o	Reasoning
Different Replacement Strategies (R:OTHER)			
I have taken that decision because...	I took that decision...	I made that decision...	Both 'took' and 'made' are grammatically valid. 'Made that decision' is a more idiomatic and standard collocation in English compared to 'took that decision'. Original "have taken" is grammatically correct as well.
I think the government will slow down the fee in public transportation, that could make people to use public transportation and caring the environment from CO2 emissions and prevent the global heating.	I think the government should reduce the fares on public transportation. That could make people use public transportation and protect the environment from CO2 emissions and prevent global warming.	I think the government will lower the fee for public transportation. That could make people use public transportation and care for the environment by reducing CO2 emissions and preventing global warming.	Both 'reduce' (Gold) and 'lower' (GPT-4o) are grammatically valid verbs to correct 'slow down'. While Gold changes 'fee' to 'fares' and GPT-4o retains 'fee', both successfully address the original error with different but equally appropriate lexical choices. The same is true for the use of "protect" and "care", and "prevent" and "preventing".
Punctuation Differences			
Adding punctuation marks (M:PUNCT)			
It can be passed on to humans through eating parts of the animal such as the brain which is affected...	It can be passed on to humans through eating parts of the animal such as the brain which is affected...	It can be passed on to humans through eating parts of the animal, such as the brain, which is affected...	GPT-4o adds commas to set off the non-restrictive example clause and relative clause, improving clarity and grammatical correctness.
Preserving punctuation (U:PUNCT)			
One month ago I ordered two batteries...	One month ago, I ordered two batteries...	One month ago I ordered two batteries...	Gold reference correctly has a comma after the introductory time phrase 'One month ago' to adhere to standard punctuation rules.
Alternative punctuation (R:PUNCT)			
When I was little I had tried a lot of sports, Handball, Swimming, Tennis, but...	When I was little, I tried a lot of sports; handball, swimming, tennis, but...	When I was little, I tried a lot of sports: handball, swimming, tennis, but...	A colon is grammatically preferred to introduce a list after 'sports', whereas a semicolon is incorrect in this context.

Table 5: Comparison of Gold Reference and GPT-4o Correction Edits [1 of 2]. "Reasoning" contains the reasoning provided by both Claude and DeepSeek LLM judges, along with human evaluators, which are holistically combined and presented here. Red shows edits in original sentences, while green shows its corresponding correction edits

Original	Gold Reference	GPT-4o	Reasoning
Orthography Corrections (R:ORTH)			
Nowadays everybody watches T.V.	Nowadays, everybody watches TV.	Nowadays, everybody watches T.V.	Gold Reference's correction of 'T.V.' to 'TV' without periods follows modern standard abbreviation conventions, while GPT-4o maintains the original format. This shows how orthographic preferences can vary, with the Gold standard version considered more contemporary.
Preposition Replacements (R:PREP)			
Finally, in the future, the commodity of people maybe is going to determinate the most use of the cars than the public transport.	Finally, in the future, the commodity of people is maybe going to determine the most use of the cars than the public transport.	Finally, in the future, the convenience for people maybe is going to determine the greater use of cars than public transport.	Both versions use different but valid prepositions: Gold Reference retains 'of people' while GPT-4o uses 'for people'. Neither choice is grammatically incorrect, demonstrating the inherent flexibility of English preposition usage where alternatives can be equally correct depending on intended meaning.
Verb Replacements (R:VERB)			
After her 18th birthday, Anna felt a sudden necessity of knowing what happened to her bilogical mother and why did she gave Anna away.	After her 18th birthday, Anna felt a sudden need to know what happened to her biological mother and why she gave Anna away.	After her 18th birthday, Anna felt a sudden need to know what had happened to her biological mother and why she had given Anna away.	Gold Reference uses simple past tense ('happened', 'gave') while GPT-4o uses past perfect ('had happened', 'had given'). Both are grammatically valid choices that reflect different temporal perspectives, with GPT-4o's version emphasizing sequence more explicitly. This demonstrates how multiple verb tense options can be acceptable in the same context.
Original	**Gold Reference**	**GPT-4o**	**Reasoning**
Determiner Differences			
The town has also improved the management and treatment of waste system...	The town has also improved the management and treatment of waste...	The town has also improved the management and treatment of the waste system...	Both corrections are grammatically valid. Gold Reference removes 'the', while GPT-4o adds 'the' before 'waste'. Both structures are grammatically correct.
This is the guestion for every one...	This is the question for everyone...	This is a question for everyone...	Both 'the question' and 'a question' are grammatically correct. The choice between 'the' and 'a' is stylistic, not grammatical.

Table 6: Comparison of Gold Reference and GPT-4o Correction Edits (continued) [2 of 2]