Evaluating Grammatical Error Correction Beyond Standard Metrics

Anonymous ACL submission

Abstract

Evaluation of Grammatical Error Correction (GEC) systems faces three critical limitations: the absence of comprehensive evaluation of the newest Large Language Models (LLMs), the reliance on limited set of evaluation metrics for comparative analysis, and the underestimation of system performance by reference-based and reference free metrics. We address these limitations by first fine-tuning state-of-the-art LLMs, which outperformed existing sequence-tagging, sequence-to-sequence, and previous generation LLM-based GEC systems across multiple reference-based metrics. This comparative analysis of multiple metrics helped capture not only correction precision but also fluency preservation and meaning retention, which are essential dimensions for practical deployment of GEC systems. Additionally, using LLM-as-a-Judge with human validation, we demonstrate that 73.76% of LLM corrections that differed from gold reference are either grammatically equally valid or preferred over gold standard, revealing that reference-based metrics significantly underestimate GEC system performance.

1 Introduction

Grammatical Error Correction (GEC) is the task of automatically detecting and correcting errors in texts, where a system takes an erroneous sentence as input and produces a grammatically correct version as output. Effective GEC systems would ideally not only achieve high accuracy but also align closely with human correction patterns, preserving meaning and fluency while providing comprehensive error correction. In order to address this challenge, three categories of GEC systems have been developed: (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, sequence tagging approaches struggle with long-range dependencies (Omelianchuk et al., 2020) and Seq2Seq models suffer from over-correction (Omelianchuk et al., 2024). These challenges have remained difficult to solve because GEC systems must balance between making enough corrections (recall) without making unnecessary ones (precision), while also accounting for grammar's contextual nature. While newer LLMs might potentially address some of these limitations, comprehensive evaluations of newest LLMs (e.g., LLaMA 3.3 70B (Grattafiori et al., 2024), GPT-4o (OpenAI, 2024), and DeepSeek v3 671B (DeepSeek-AI et al., 2025)) are notably absent, as previous studies evaluated older LLMs (e.g., LLaMA-2) (Omelianchuk et al., 2024).

Beyond these challenges, most research relies on a single reference-based evaluation metric, typically ERRANT $F_{0.5}$ (Bryant et al., 2017), which evaluates correction quality on individual edit level (Omelianchuk et al., 2024). This results in a scarcity of research evaluating GEC systems on higher-level dimensions like fluency preservation or meaning retention after correction, which can be captured using other reference-based metrics like GLEU (Napoles et al., 2016a) and PT-ERRANT (Gong et al., 2022). This narrow evaluation focus limits our understanding of how GEC systems perform across dimensions critical for student learning and writing assistance.

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 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) and Scribendi score (Islam and Magnani, 2021), are developed to directly assess semantic and grammatical acceptability without references, but they often exhibit bias towards minimal edits. These evaluation limitations create uncertainty about the true performance of GEC systems especially in high-stakes settings like educational applications.

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. To further increase performance, we developed an ensemble architecture by combining our best models using sentence-level majority voting, with fallback strategy that selects corrections with the highest word-level n-gram overlap across candidate corrections when there is no majority. These helped us build and present the first multi-metric comparative analysis of latest generation LLMs against multiple sequence tagging, Seq2Seq, and previous generation LLM-based GEC approaches, capturing different dimensions of correction quality, including edit accuracy, fluency preservation, and meaning retention. 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:

- Fine-tuned GPT-4o model established a new state-of-the-art for standalone GEC systems, further validated by our top-ranking performance in XXX Shared Task.[2]
- First extensive GEC system comparison of latest LLMs using multiple metrics to capture grammatical dimensions critical for real-world deployment.
- 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, models and data will be available at xxx
[2] Shared task link anonymized for review

2 Related Work

This section examines established GEC systems and their evaluation challenges.

2.1 Evolution of GEC Systems

Sequence tagging GEC approaches like GECToR (Omelianchuk et al., 2020) use pre-trained encoders (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2020)) to predict specific edit operations for each token from a large vocabulary of transformation tags, with larger versions of these encoders giving better performance (Tarnavskyi et al., 2022). However, these approaches struggle with complex, interconnected errors which require broader contextual understanding.

On the other hand, Seq2Seq models approach GEC as a translation task from incorrect to correct sentences. These systems have evolved 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 over-correction tendencies.

Alternatively, LLMs approach GEC through conditional text generation, leveraging knowledge acquired during pre-training. In zero-shot settings, models like GPT-4 and LLaMA-2 excel at error detection but exhibit low precision due to unnecessary corrections (Fang et al., 2023), with fine-tuning improving performance to a level comparable to specialized Seq2Seq and sequence tagging systems (Omelianchuk et al., 2024). Despite extensive research and comparative analysis (Kobayashi et al., 2024a), latest-generation LLMs haven't been adequately compared to established GEC systems, creating a gap in comparative performance assessment.

While these individual models show promising results, ensemble approaches consistently outperformed them historically, such as Tarnavskyi et al. (2022)'s 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 on specific datasets, rather than a single generalizable model. Building on this, Omelianchuk et al. (2024) analyzed various ensembling techniques (GRECO, GPT-based rank-

ing, 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.

2.2 Evaluation of GEC Systems

Current evaluation methods fall into two main categories: reference-based and reference-free metrics. Reference-based approaches, including ERRANT (Bryant et al., 2017), GLEU (Napoles et al., 2016a), and PT-ERRANT (Gong et al., 2022), compare system outputs to gold-standard corrections. While widely used, they systematically underestimate performance by failing to account for the diversity of valid corrections (Napoles et al., 2016b; Katinskaia and Yangarber, 2023). Rozovskaya and Roth (2021) demonstrated that system scores improved by 20–40 points when references were adjusted to accept any valid correction. Additionally, Woo and Choi (2021) showed that evaluation of GEC systems rely mostly on ERRANT $F_{0.5}$ that omits key dimensions such as preserving fluency and the intended meaning of the sentence. Rozovskaya and Roth (2014) further demonstrated that architecturally similar systems can produce different $F_{0.5}$ scores, revealing the limitations of relying solely on edit-level correction precision.

Alternatively, reference-free metrics attempt to address these issues by evaluating semantic and grammatical acceptability without requiring gold standards. However, metrics like IMPARA (Maeda et al., 2022) and Scribendi score (Islam and Magnani, 2021) are biased toward minimal edits and struggle with scoring comprehensive corrections (Bryant et al., 2023). Additionally, Kobayashi et al. (2024b) found that both reference-based and reference-free metrics exhibit lower correlations with human judgments for LLM outputs compared to sequence tagging or Seq2Seq models, partly because LLMs produce more extensive stylistic, yet valid, improvements.

Human evaluation can address these gaps but is prohibitively costly for large-scale assessment. To bridge this gap, LLM-as-a-Judge methods (Kobayashi et al., 2024b; Goto et al., 2025b) have emerged as promising alternatives, offering inter-annotator agreement levels comparable to human evaluators at lower cost. However, relying on a single LLM for evaluation risks introducing biases that could compromise assessment objectivity and reliability.

3 Methodology

To address these challenges, we conduct a multi-metric comparison of recent LLMs, and use hybrid LLM-human framework to validate corrections.

3.1 Model Selection and Ensembles

We initially evaluated three leading LLMs: GPT-4o (OpenAI, 2024), LLaMA-3.3-70B-Instruct (Grattafiori et al., 2024), and DeepSeek-V3-671B (DeepSeek-AI et al., 2025), in zero-shot setting using the prompt in Figure 1 in the Appendix A.2. Following this, two of the three models are fine-tuned (GPT-4o[3] and LLaMA 3.3) on the BEA training data (details in subsection 4.1) for two epochs and using the same prompt template as in zero-shot inference. DeepSeek-V3 was excluded due to the prohibitive computational costs of fine-tuning a 671B-parameter model. Further fine-tuning details and computational cost information is in Appendix A.3.

Following this, we experimented with ensembles by combining the outputs from our two fine-tuned models (GPT-4o and LLaMA-3.3) and the zero-shot DeepSeek-V3 model. These ensemble approaches address whether different latest generation LLMs exhibit complementary strengths across error types, and leverage these complementary capabilities to provide more consistent and comprehensive corrections. All our ensemble variants use a majority voting mechanism as their primary decision rule. However, the ensembles differ in their fallback strategy for cases where no agreement exists among the three models:

Best Model Fallback: relies on the best-performing model's correction as fallback when no majority exist, reflecting how educational institutions often rely on their most trusted correction tool when uncertainty arises.

Qwen Fallback: employs Qwen-2.5-7B-Instruct (Qwen et al., 2025) model as a meta-selector, similar to having a neutral third party help resolve differing correction approaches.[4]

Perplexity Fallback: computes the perplexity score of each candidate correction using base Qwen-2.5-7B[5] and selects the correction with the lowest perplexity. LLMs are pre-trained primarily on grammatically correct text, causing them to assign higher probabilities (lower perplexity) to

[3] GPT-4o fine-tuned using OpenAI API
[4] Prompt used is in Appendix A.4
[5] https://huggingface.co/Qwen/Qwen-7B

Metric	Description
Reference-based Metrics	
ERRANT $F_{0.5}$ (Bryant et al., 2017)	Evaluates edit-level precision ensuring systems make reliable corrections without inappropriately modifying grammatically correct text.
GLEU (Napoles et al., 2016a)	Assesses fluency through n-gram overlap with reference corrections, measuring whether corrections produce naturally fluent text that aligns with human judgment.
PT-ERRANT (Gong et al., 2022)	Evaluates phrase-level and structural modifications with semantic understanding, measuring a system's ability to preserve intended meaning while making corrections.
Reference-free Metrics	
IMPARA (Maeda et al., 2022)	Accommodates valid alternative corrections, with the aim of supporting linguistic diversity.
Scribendi (Islam and Magnani, 2021)	Uses GPT-2 perplexity, token sort ratio, and Levenshtein distance ratio to evaluate correction quality.

Table 1: Overview of evaluation metrics used in this study.

corrections that follow grammatical patterns encountered during pre-training.

N-gram Fallback: selects 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. We use word-level n-gram (n=3) with Jaccard Similarity, $\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$, and select $\arg\max_i \left(\sum_{j \neq i} \text{Jaccard}(\text{n-grams}_i, \text{n-grams}_j) \right)$.

We initially implemented perplexity-based scoring, meta-model selection, and N-gram based best correction selection as primary ensemble strategies. However, evaluation showed that these methods performed worse than simple majority voting. We achieved better performance only when implementing these as fallback strategies for cases where majority voting failed to reach consensus.

3.2 Automated Metrics Evaluation

For comprehensive evaluation across different grammatical dimensions, we use five metrics, as shown in Table 1 (equations in Appendix A.5):

3.3 LLM-as-a-Judge and Human Evaluation

To address the limitations of existing automated GEC evaluation, we introduce a hybrid evaluation framework that combines LLM-as-a-Judge with targeted human evaluation. This evaluation protocol first employs two LLMs (independent from the LLMs used for GEC evaluation in this study), Claude 3.7 Sonnet[6] and DeepSeek-R1 (DeepSeek-AI, 2025), 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 LLM's demonstrated high correlation with human evaluations in GEC tasks (Kobayashi et al., 2024a).

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 (Kobayashi et al., 2024b; Goto et al., 2025b). Reliability is further improved through two-model consensus mechanisms that mitigate individual LLM biases.

In cases where the LLM judges disagree, two qualified human evaluators applied the same three-category framework to resolve discrepancies. Both annotators have professional English proficiency and prior GEC annotation experience, were completely blind to system identity and disagreements between them were resolved through discussion. This evaluation framework substantially reduces time and resources, as human judgment is invoked only for contested cases, making the evaluation approach highly scalable.

4 Experiments and Results

Our experimental results reveal that GPT-4o outperforms established individual GEC systems across multiple dimensions ($p < 0.001$, permutation test). Additionally, our LLM-as-a-Judge with human validation framework reduces human evaluation effort by 64%, while showing that around 73% of model corrections (which were penalized by reference based metrics) are equally valid or preferred.

[6] Anthropic Claude 3.7 Sonnet System Card

Model	ERRANT	GLEU	PT-ERRANT	IMPARA	Scribendi
Our Individual Models					
Fine-Tuned GPT-4o	0.6599	**0.8400**	0.7064	0.7768	0.5189
Fine-Tuned LLaMA 3.3	0.6420	0.8281	0.6842	0.7705	0.4986
Base DeepSeek	0.4926	0.7677	0.5666	0.7754	0.5977
Base GPT-4o	0.4592	0.7420	0.5484	0.7564	**0.6843**
Base LLaMA 3.3	0.4826	0.7345	0.4973	0.7122	0.4710
Our Ensemble Systems					
Majority Voting + GPT-4o Fallback	0.6607	0.8392	0.7039	0.7746	0.5128
Majority Voting + Qwen Fallback	0.6249	0.8312	0.6905	0.7855	0.5294
Majority Voting + Perplexity Fallback	0.6251	0.8304	0.6879	0.7858	0.5381
Majority Voting + N-gram Fallback	**0.6623**	0.8347	**0.7122**	0.7749	0.5103
LLMs (Decoder Only Transformers)					
Fine-Tuned LLaMA 2 7B (Omelianchuk et al., 2024)	0.5530	0.7985	0.6157	0.7529	0.5018
Fine-Tuned LLaMA 2 13B (Omelianchuk et al., 2024)	0.5640	0.8027	0.6399	0.7554	0.4952
Seq2Seq Models (Encoder-Decoder Transformer)					
Fine-Tuned T5 11B (Omelianchuk et al., 2024)	0.5860	0.8231	0.6656	0.7629	0.4929
Fine-Tuned FLAN 20B (Omelianchuk et al., 2024)	0.5770	0.8149	0.6630	0.7582	0.4799
Sequence Tagging Systems					
GeCTOR (XLNet) (Omelianchuk et al., 2020)	0.5630	0.7687	0.6248	0.7058	0.4035
CTC-Copy (Zhang et al., 2023)	0.5270	0.7714	0.6096	0.7302	0.4154
EditScorer (Sorokin, 2022)	0.5740	0.7565	0.6285	0.7072	0.4035
Ensemble and Model Ranking Approaches					
Ensemble Best 7 (Omelianchuk et al., 2024)	0.6290	0.7854	0.7040	0.7153	0.4370
Ensemble Best 3 (Omelianchuk et al., 2024)	0.6250	0.7907	0.7000	0.7216	0.4396
GRECO Rank 7 (Omelianchuk et al., 2024)	0.6200	0.8032	0.7084	0.7366	0.4451
GPT-4 Rank 3 (Omelianchuk et al., 2024)	0.5810	0.8270	0.6654	0.7753	0.5374
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 2: Performance comparison of models on BEA-dev dataset. Existing systems' results are obtained from Omelianchuk et al. (2024), except shared task systems which were provided by organizers. As stated by Omelianchuk et al. (2024), Ensemble Best 7 includes all 7 models; 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).

4.1 Dataset

For our experiments, we trained on the ABC (combining beginner, intermediate, and advanced English texts) train partition of the W&I+LOCNESS dataset from the BEA 2019 Shared Task (Bryant et al., 2019; Granger, 1998). We tested our results on the ABCN development set (including native English texts) of BEA 2019, and on CoNLL-14 test set (Ng et al., 2014), which contains English essays written by undergraduate students.

4.2 Model Performance

Table 2 and 3 reveal that fine-tuned GPT-4o outperforms all individual models tested, with notable improvements in correction accuracy (ERRANT $F_{0.5}$), fluency (GLEU), and meaning retention (PT-ERRANT). Among zero-shot models, DeepSeek shows superior performance across all metrics due to its mixture-of-experts (MoE) architecture. DeepSeek selectively activates specialized sub-networks for each token, possibly enabling it to better handle diverse linguistic patterns. On the other hand, GPT-4o is outperformed by both open-source LLMs in zero-shot setting. After fine-tuning, GPT-4o's modest advantage over LLaMA demonstrates the competitiveness of open-source models, which offer greater accessibility and transparency for deployment of GEC systems, especially in high-stakes educational settings. Statistical confidence test results are provided in Tables 6, 7, 8 9 and 10 in Appendix A.6.

Our systems' performance was further validated in the XXX Shared Task[7], where our best performing models ranked highest across all three reference-based metrics, as shown in Table 2. 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, XLNet, and DeBERTa) with majority voting.

[7]Anonymized for review

Model	ERRANT	GLEU	PT-ERRANT	IMPARA	Scribendi
Our Individual Models					
Fine-Tuned GPT-4o	0.6752	**0.7183**	0.7390	0.8973	0.6517
Fine-Tuned LLaMA 3.3	0.6287	0.6969	0.6942	0.8987	0.5511
Base DeepSeek	0.5923	0.6917	0.6651	0.8795	**0.6799**
Our Ensemble Systems					
Majority Voting + GPT-4o Fallback	0.6752	**0.7183**	0.7390	0.8973	0.6517
Majority Voting + Qwen Fallback	0.6572	0.7145	0.7180	**0.9048**	0.6662
Majority Voting + Perplexity Fallback	0.6381	0.7070	0.7084	0.9033	0.6509
Majority Voting + N-gram Fallback	0.6846	0.7161	**0.7433**	0.8904	0.6509
LLMs (Decoder-only Transformers)					
Fine-Tuned LLaMA 2 7B (Omelianchuk et al., 2024)	0.6375	0.6773	0.6789	0.8330	0.5503
Fine-Tuned LLaMA 2 13B (Omelianchuk et al., 2024)	0.6397	0.6721	0.6935	0.8236	0.5023
Seq2Seq Models (Encoder-Decoder Models)					
Fine-Tuned T5 11B (Omelianchuk et al., 2024)	0.6309	0.7016	0.6991	0.8883	0.5968
Fine-Tuned FLAN 20B (Omelianchuk et al., 2024)	0.6265	0.6960	0.6949	0.8809	0.5899
Sequence Tagging Systems					
GeCTOR (XLNet) (Omelianchuk et al., 2024)	0.6308	0.6716	0.6709	0.8059	0.5053
CTC-Copy (Zhang et al., 2023)	0.6185	0.6773	0.6664	0.8395	0.5229
EditScorer (Sorokin, 2022)	0.6300	0.6530	0.6683	0.7940	0.4718
Ensemble and Model Ranking Approaches					
Ensemble Best 7 (Omelianchuk et al., 2024)	**0.7030**	0.6750	0.7400	0.8086	0.5152
Ensemble Best 3 (Omelianchuk et al., 2024)	0.6849	0.6691	0.7260	0.8092	0.5000
GRECO Rank 7 (Omelianchuk et al., 2024)	0.6981	0.6879	0.7324	0.8311	0.5336
GPT-4 Rank 3 (Omelianchuk et al., 2024)	0.6414	0.7114	0.7056	0.8945	0.6410

Table 3: Performance comparison of models on CoNLL-2014 dataset.

4.3 Comparative Analysis of Existing Systems

Table 2 demonstrates that even latest LLMs, in zero-shot setting, significantly under-perform individual sequence tagging and Seq2Seq GEC systems. This makes them unsuitable for high-stakes educational use where incorrect feedback could mislead students about grammar rules. However, fine-tuning proves essential, with GPT-4o improving by 22.07 points to surpass all existing individual GEC systems.

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 and Scribendi score, metrics known to favor minimal edits, suggests these external evaluation criteria prioritize conservative corrections over comprehensive ones even when comprehensive corrections are necessary.

Additionally, none of our ensemble configurations significantly outperform the single fine-tuned GPT-4o model (permutation test, $p > 0.05$). On BEA-dev, our best ensemble using N-gram fallback achieves ERRANT $F_{0.5}$ of 0.6623 compared to GPT-4o's 0.6599 ($p > 0.05$), with similarly insignificant differences across GLEU (0.8347 vs. 0.8400, $p > 0.05$) and PT-ERRANT (0.7122 vs. 0.7064, $p > 0.05$). This pattern extends to CoNLL-14 ($p > 0.05$), demonstrating that combining multiple newest LLMs provides no measurable benefit over deploying a single fine-tuned model.

Furthermore, on BEA-dev, both our N-gram ensemble and fine-tuned GPT-4o achieve performance statistically equivalent to the previous best ensemble baseline (Omelianchuk et al., 2024), with no significant differences ($p > 0.05$). Additionally, on CoNLL-14, the ensemble from Omelianchuk et al. maintains slightly superior performance compared to our single GPT-4o model ($p < 0.0001$). This makes it more suitable to deploy single fine-tuned models as they achieve close to equivalent performance to ensemble approaches while requiring substantially lower computational resources.

4.4 Error Type Analysis

To investigate why homogeneous LLM ensembles fail to provide complementary coverage, we conducted systematic error-type analysis examining correction rates and false insertion rates across 54 ERRANT error categories for our three models (detailed results in Tables 13 and 14 in Appendix A.8). Table 15 presents Spearman rank correlations revealing exceptionally strong agreement in LLMs' per-error-type performance patterns. So, all LLMs seem to excel at correcting similar error types, while struggling with the same difficult error categories.

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 4: GEC examples for reference free metric IMPARA. Ensemble is the Majority Voting with GPT-4o Fallback.

For correction rates, fine-tuned GPT-4o and LLaMA show near-perfect correlation ($\rho = 0.947$, $p < 0.0001$), with similarly high correlations for other LLM pairs (Table 15 in Appendix A.8). When GPT-4o achieves high correction rates for spelling errors, the other models perform similarly well. Conversely, all models struggle equally with conjunction replacements. False insertion patterns show comparable correlation with ρ values from 0.54 to 0.82 (all $p < 0.0001$). This means LLMs share tendencies toward over-correction on identical error categories. These findings show LLMs, despite differences in model families (GPT, LLaMA, DeepSeek), converge on similar error patterns due to shared architectural principles and training objectives.

Further analysis of correction rates reveal that all models demonstrate strong performance on closed-class grammatical categories with limited lexical variation. These include determiners, prepositions, and rule-governed morphological transformations such as subject-verb agreement, noun number, and spelling corrections. These error types rely primarily on syntactic patterns and morphological rules that are well-represented in pre-training corpora and can be corrected through pattern matching. On the other hand, all models struggle with open-class semantic decisions requiring broader contextual understanding. Correction rates remain below 50% for content word replacements such as nouns, verbs, adjectives, and adverbs. Contextual connectives, like conjunction insertions, have very low correction rates as these errors require semantic judgment and contextual understanding that extends beyond surface-level linguistic patterns. Furthermore, LLMs exhibit over-correction patterns for discourse-level elements. False insertion rates exceed 50% for conjunctions and adverbs, suggesting that pre-training biases towards explicit discourse connectivity lead to unnecessary corrections made by LLMs.

4.5 Analysis of Evaluation Metric Limitations

Reference based evaluation penalize linguistic diversity. This particularly disadvantages multilingual learners whose diverse language choices are marked as incorrect despite being grammatically valid. The most common differences occur in replacement edits, indicating that users can often correct grammatical errors in multiple valid ways. Yet reference-based metrics penalize GEC systems for recognizing this flexibility. Multiple fundamental linguistic elements can be expressed in numerous grammatically valid ways such as punctuation, articles, prepositions, and verbs.

Punctuation varies across different style traditions and linguistic backgrounds, such as users may add commas for clarity, or employ semicolons versus periods based on their native language conventions. Articles present multiple valid options because many languages lack article systems entirely, leading their speakers to develop different but equally correct approaches to English article usage, such as, sometimes not using "the". Prepositions naturally allow variation because different languages structure spatial and temporal relationships differently. For example, "at the beach" versus "on the beach" can both be contextually appropriate, and "different from" versus "different than" represents legitimate regional variation. Similarly, verbs offer multiple valid forms through different tense choices, voice selections, and formality levels. Notably, "people are suffering" versus "people suffer" can both correctly express ongoing conditions.[8]

To address these limitations, reference-free metrics like IMPARA and Scribendi were developed. However, our analysis, also supported by Sugiyama et al. (2025) and Gotō et al. (2025), reveals IMPARA favors minimal edits while severely penalizing essential comprehensive corrections, as demonstrated in Table 4. This limitation stems from IMPARA's 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. Scribendi score also favours minimal edits, as it incorporates Levenshtein distance ratio, which measures the similarity between source and

[8]More examples in Appendix A.7

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 5: LLM-as-a-Judge and Human Evaluation Results (On the BEA-dev dataset, for 1,730 edits for which fine-tuned GPT-4o produced different edit corrections compared to gold reference).

corrected sentences and penalizes larger edits. It can also be seen in Table 2 that base GPT-4o which is less inclined to make comprehensive corrections, is given a higher Scribendi score compared to fine-tuned GPT-4o which has learned to make comprehensive corrections wherever necessary. Given these documented limitations, we focused our comparative analysis on reference-based metrics.

4.6 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. For cases where dual LLMs reached consensus, LLM-as-a-Judge demonstrated strong inter-annotator agreement with human experts (kappa = 0.72, 82% overall agreement), with particularly high reliability for clear-cut judgments: 91% agreement on identifying instances where GPT-4o corrections are preferred, and 85% agreement on recognizing gold standard preference. Disagreements were primarily concentrated in cases where humans judged items as equivalent, suggesting the LLMs are inclined to make more decisive evaluations (Confusion Matrix provided in Appendix A.10). This approach reduces traditionally time-intensive human evaluation effort by 64%, making rigorous correction quality assessment more feasible while revealing the shortcomings of current reference-based metrics.

We identified all 1,730 corrections where GPT-4o output differed from gold standard corrections. As demonstrated in Table 5, two LLM judges reached consensus in 64.34% of cases out of 1,730, preferring our model's corrections (30.87%) over gold standards (19.77%), with 13.70% rated as equally valid. Following human evaluation to resolve all LLM disagreements (by two evaluators with inter annotator agreement, Cohen's κ: 0.975), 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. Figure 4 in Appendix A.11 demonstrates the counts of different edit types according to the three preference categories: gold standard preferred, GPT-4o correction preferred, or no preference (rated equally).

In 73.76% of cases where our model produced corrections different from gold standards, these were actually judged as either superior or equally valid compared to gold references. This further validates that GPT-4o not only outperforms existing individual GEC approaches on automated metrics but also produces corrections that are frequently preferred over or considered equivalent to the gold standards themselves. This complementary evaluation demonstrates 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) for assessing correction quality.

5 Conclusion

Fine-tuned GPT-4o establishes a new state-of-the-art for individual GEC systems, surpassing previous benchmarks across multiple reference-based metrics. Alternatively, reference-based metrics 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 preferred. Through error type analysis, we show 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. These findings shows the necessity of complementing automated metrics with human evaluation, aided by LLMs for scalability, to accurately assess GEC system capabilities. Our dual LLM-as-a-Judge framework provides a solution to this by reducing human evaluation effort by 64% while maintaining reliability, making comprehensive quality assessment more feasible.

Limitations

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 fine-tune this model due to prohibitive computational requirements. In contrast, we successfully fine-tuned GPT-4o via OpenAI's API without hosting the model locally (OpenAI employs parameter-efficient methods, though OpenAI does not disclose the specific techniques), and adapted LLaMA 70B using Low-Rank Adaptation (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.

These experiments were conducted using GPT-4o, LLaMA-3.3-70B-Instruct, and DeepSeek-V3-671B. Given the rapid development cycle of LLMs, newer versions of these models, or entirely new architectures, may have been released subsequent to our experiments, potentially yielding improved performance. As the field continues to evolve rapidly, our findings represent a snapshot of current capabilities. Future work will incorporate updated model versions to assess performance improvements and validate the generalizability of our findings.

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.

Ethical Considerations

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.

This research utilizes established public benchmark datasets with appropriate consent and no additional personal data collection. 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. 2025. 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, and 1 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

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 Goto, Yusuke Sakai, and Taro Watanabe. 2025a. gec-metrics: A unified library for grammatical error correction evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 524–534, Vienna, Austria. Association for Computational Linguistics.

Takumi Goto, Yusuke Sakai, and Taro Watanabe. 2025b. Rethinking evaluation metrics for grammatical error correction: Why use a different evaluation process than human? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1165–1172, Vienna, Austria. 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, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Md Asadul Islam and Enrico Magnani. 2021. Is this the end of the gold standard? a straightforward referenceless grammatical error correction metric. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3009–3015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anisia Katinskaia and Roman Yangarber. 2023. Grammatical error correction for sentence-level assessment in language learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 488–502, Toronto, Canada. Association for Computational Linguistics.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024a. Large language models are state-of-the-art evaluator for grammatical error correction. *Preprint*, arXiv:2403.17540.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024b. Revisiting meta-evaluation for grammatical error correction. *Transactions of the Association for Computational Linguistics*, 12:837–855.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

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. 2016a. Gleu without tuning. *arXiv preprint arXiv:1605.02592*.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016b. There's no comparison: Referenceless evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115, Austin, Texas. 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 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, Junyang Lin, and 24 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. 2014. Building a state-of-the-art grammatical error correction system. *Transactions of the Association for Computational Linguistics*, 2:419–434.

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.

Jin Ha Woo and Heeyoul Choi. 2021. Systematic review for ai-based language learning tools. *CoRR*, abs/2111.04455.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding. *Preprint*, arXiv:1906.08237.

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.

A Appendix

A.1 Licenses and Data Availability

We use the W&I+LOCNESS dataset under the LOCNESS non-commercial research LOCHNESS license and CoNLL-2014 dataset under the NUCLE license agreement. Model fine-tuning was conducted using GPT-4o via OpenAI API (subject to OpenAI Terms of Service), LLaMA 3.3-70B-Instruct under Meta's Llama 3.3 Community License, and DeepSeek V3 under the MIT License. Our code, annotated data and models will be released under MIT license.

Our use of all datasets and models is consistent with their intended purposes as specified by their creators. The W&I+LOCNESS and CoNLL-2014 datasets are used solely for GEC research as intended. Our released code and fine-tuned models are intended for GEC research and educational purposes only. Due to training on research-only licensed datasets, these artifacts should not be used for commercial purposes and are subject to the same research-use restrictions as the original training data.

A.2 Prompt for GEC Inference and Fine-tuning

Figure 1 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 1: Inference and Fine-tune Prompt.

A.3 Fine-tuning Method and Costs

This subsection provides additional explanation on the fine-tuning of GPT-4o and LLaMA 3.3, so the results can be reproducible. Training data consisted of 34,308 sentences from the ABC train partition of W&I+LOCNESS dataset. GPT-4o was fine-tuned using OpenAI's fine-tuning API for 2 epochs, by employing parameter-efficient methods (though OpenAI does not disclose the specific techniques). We acknowledge that the lack of transparency in OpenAI's fine-tuning process might limit reproducibility. LLaMA 3.3 was adapted using LoRA via the PEFT library with the following parameters: rank=8, alpha=16, and target modules=attention layers (q_proj, k_proj, v_proj, o_proj). Training of LLaMA 3.3 was conducted with a learning rate of 2e-5, cosine scheduler (warmup ratio 0.1), AdamW optimizer (betas=(0.9, 0.999), epsilon=1e-8), gradient accumulation steps=8, and a total batch size of 32 across 2 GPUs over 2 epochs. All code, models (except GPT-4o as it is on OpenAI platform) and data will be available at xxx.

In terms of costs, GPT-4o fine-tuning via OpenAI's API cost approximately $88.48 in total (for both fine-tuning and inference) for 34,308 training samples over 2 epochs. For LLaMA-3.3, fine-tuning required approximately 3 hours (using two rented H200 GPU, 141GB VRAM, 188GB RAM, 12 VCPU on Runpod) with $3.47 per hour on per GPU. The total cost inclusive of fine-tuning and inference of the LLaMA model is $27.856.

A.4 Prompt for Qwen Meta-Model

Figure 2 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.5 Information of Automated Metrics

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

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

The equation for GLEU is given below:

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

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.

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 2: Prompt used for Qwen to judge which model correction is best.

We computed evaluation metrics using the gec-metrics library (Goto et al., 2025a) with default configurations for ERRANT, GLEU, PT-ERRANT, IMPARA, and Scribendi. The link to the Github repository is gec-metrics.

A.6 Permutation Test Scores

Tables 6, 7 and 8 present pairwise permutation tests, on the BEA dev dataset, comparing the performance of different grammatical error correction (GEC) systems across three key evaluation metrics, ERRANT $F_{0.5}$, GLEU and PT-ERRANT respectively. Each table reports p-values, performance differences, and 95% bootstrap confidence intervals for system performance difference comparisons. Tables 9 and 10 presents permutation tests on the CoNLL test dataset.

A.7 Examples of Edit Differences

Tables 11 and 12 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.

A.8 Correction rates of different error types by different models

Tables 13 and 14 present detailed error-type analysis of model performance. Table 13 shows the percentage of gold errors that each model successfully corrects, while Table 14 reports the rate of false insertions (corrections made by models which are not present in the actual sentence, or in other words, the model is over-correcting). Table 15 shows the Spearman rank correlations between models' per-error-type correction rates and false insertion rates across all 54 ERRANT error categories. Table 16 in the Appendix A.9 provides detailed descriptions of all edit type codes.

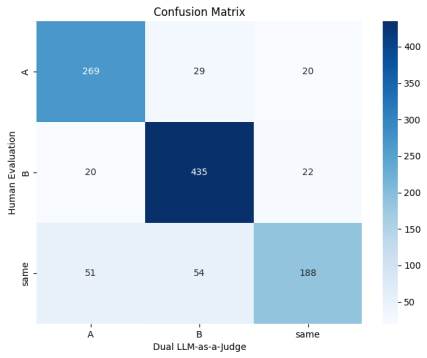

Figure 3: Confusion Matrix for Dual LLM-as-a-Judge agreement with human evaluation.

A.9 ERRANT Error Type Descriptions.

Table 16 provides descriptions of the ERRANT edit types referenced in this study.

A.10 Confusion Matrix for LLM-as-a-Judge

Figure 3 shows the confusion matrix for the agreement between dual LLM-as-a-Judge and human evaluation, with rows representing human annotations and columns representing dual LLM-as-a-Judge's annotations, when the two LLM judges themselves are in agreement. The confusion matrix reveals that disagreements are not random but follow systematic patterns. Dual LLM-as-a-Judge (when in agreement) demonstrates strong reliability in distinguishing between clearly preferred options (A vs B), with cross-category errors comprising only 4-9% of judgments. The primary source of disagreement occurs when human judge classifies corrections as equivalent. In these cases, dual LLM-as-a-Judge tends to identify a preferred option rather than declaring a tie, suggesting a more decisive evaluation approach.

A.11 Edit Preference

Figure 4 shows the edit counts for different error types according to three preference categories: gold standard preferred, GPT-4o correction preferred, or no preference (rated equally). Table 16 in the Appendix A.9 provides detailed descriptions of all edit type codes.

System 1	System 2	P-Value	Difference	Bootstrap CI
Fine-tuned GPT-4o	Zero-shot GPT-4o	0.0000	0.2012	(0.1926, 0.2106)
Fine-tuned LLaMA	Zero-shot LLaMA	0.0000	0.1591	(0.1486, 0.1713)
Fine-tuned GPT-4o	Ensemble Best 7	0.3800	-0.0047	(-0.0109, 0.0031)
Fine-tuned GPT-4o	EditScorer	0.0000	0.0698	(0.0671, 0.0773)
Fine-tuned GPT-4o	Fine-tuned LLaMA 13B	0.0000	0.0525	(0.0463, 0.0554)
Fine-tuned GPT-4o	Fine-tuned T5 11B	0.0000	0.0276	(0.0224, 0.0315)
N-gram	Best Model	0.9100	0.0004	(-0.0041, 0.0018)
N-gram	Qwen	0.0000	0.0374	(0.0330, 0.0417)
N-gram	Perplexity	0.0000	0.0374	(0.0340, 0.0443)
N-gram	Ensemble Best 7	0.5400	-0.0028	(-0.0067, 0.0023)
N-gram	Fine-tuned GPT-4o	0.6600	0.0019	(-0.0048, 0.0080)

Table 6: Statistical significance tests comparing system performance on ERRANT $F_{0.5}$ scores on BEA dev. P-values, differences, and 95% bootstrap confidence intervals are shown for pairwise comparisons.

System 1	System 2	P-Value	Difference	Bootstrap CI
Fine-tuned GPT-4o	Zero-shot GPT-4o	0.0001	0.2012	(0.1934, 0.2062)
Fine-tuned LLaMA	Zero-shot LLaMA	0.0000	0.1591	(0.1492, 0.1794)
Fine-tuned GPT-4o	Ensemble Best 7	-0.0047	0.3800	(-0.0168, 0.0052)
Fine-tuned GPT-4o	EditScorer	0.0000	0.0698	(0.0651, 0.0753)
Fine-tuned GPT-4o	Fine-tuned LLaMA 13B	0.0000	0.0525	(0.0445, 0.0589)
Fine-tuned GPT-4o	Fine-tuned T5 11B	0.0000	0.0276	(0.0234, 0.0349)
N-gram	Best Model	0.9100	0.0004	(-0.0023, 0.0056)
N-gram	Qwen	0.0000	0.0374	(0.0343, 0.0407)
N-gram	Perplexity	0.0000	0.0374	(0.0348, 0.0478)
N-gram	Ensemble Best 7	0.5400	-0.0028	(-0.0105, 0.0014)
N-gram	Fine-tuned GPT-4o	0.6600	0.0019	(-0.0035, 0.0093)

Table 7: Statistical significance tests comparing system performance on GLEU scores on BEA dev. P-values, differences, and 95% bootstrap confidence intervals are shown for pairwise comparisons.

System 1	System 2	P-Value	Difference	Bootstrap CI
Fine-tuned GPT-4o	Zero-shot GPT-4o	0.0000	0.1570	(0.1568, 0.1672)
Fine-tuned LLaMA	Zero-shot LLaMA	0.0000	0.1651	(0.1646, 0.1929)
Fine-tuned GPT-4o	Ensemble Best 7	0.7000	0.0030	(-0.0060, 0.0090)
Fine-tuned GPT-4o	EditScorer	0.0000	0.0785	(0.0683, 0.1028)
Fine-tuned GPT-4o	Fine-tuned LLaMA 13B	0.0000	0.0672	(0.0614, 0.0729)
Fine-tuned GPT-4o	Fine-tuned T5 11B	0.0000	0.0415	(0.0324, 0.0476)
N-gram	Best Model	0.7000	0.0014	(-0.0032, 0.0013)
N-gram	Qwen	0.0000	0.0217	(0.0161, 0.0241)
N-gram	Perplexity	0.0000	0.0246	(0.0224, 0.0269)
N-gram	Ensemble Best 7	0.5000	0.0078	(-0.0057, 0.0129)
N-gram	Fine-tuned GPT-4o	0.2000	0.0048	(0.0001, 0.0106)

Table 8: Statistical significance tests comparing system performance on PT-ERRANT scores on BEA dev. P-values, differences, and 95% bootstrap confidence intervals are shown for pairwise comparisons.

System 1	System 2	P-Value	Difference	Bootstrap CI
Fine-tuned GPT-4o	Ensemble Best 7	0.0000	-0.0278	(-0.0501, -0.0263)
Fine-tuned GPT-4o	EditScorer	0.0000	0.0452	(0.0385, 0.0643)
Fine-tuned GPT-4o	Fine-tuned LLaMA 13B	0.0000	0.0355	(0.0240, 0.0535)
Fine-tuned GPT-4o	Fine-tuned T5 11B	0.0000	0.0443	(0.0331, 0.0501)
N-gram	Ensemble Best 7	0.0300	-0.0183	(-0.0265, -0.0043)
N-gram	Fine-tuned GPT-4o	0.0800	0.0095	(0.0050, 0.0154)

Table 9: Statistical significance tests comparing system performance on ERRANT $F_{0.5}$ scores on CoNLL-2014. P-values, differences, and 95% bootstrap confidence intervals are shown for pairwise comparisons.

System 1	System 2	P-Value	Difference	Bootstrap CI
Fine-tuned GPT-4o	Ensemble Best 7	0.0000	0.0435	(0.0384, 0.0475)
Fine-tuned GPT-4o	EditScorer	0.0000	0.0660	(0.0587, 0.0699)
Fine-tuned GPT-4o	Fine-tuned LLaMA 13B	0.0000	0.0468	(0.0392, 0.0548)
Fine-tuned GPT-4o	Fine-tuned T5 11B	0.0000	0.0167	(0.0144, 0.0214)
N-gram	Ensemble Best 7	0.0000	0.0413	(0.0360, 0.0454)
N-gram	Fine-tuned GPT-4o	0.1600	-0.0022	(-0.0045, -0.0000)

Table 10: Statistical significance tests comparing system performance on GLEU scores on CoNLL-2014. P-values, differences, and 95% bootstrap confidence intervals are shown for pairwise comparisons.

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 11: 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.

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

Error Type	Count	GPT-4o Correction %	Llama Correction %	DeepSeek Correction %
M:ADJ	20	40.0	40.0	25.0
M:ADV	29	37.9	20.7	37.9
M:CONJ	24	8.3	4.2	37.5
M:CONTR	2	0.0	50.0	0.0
M:DET	361	82.0	76.7	74.8
M:NOUN	44	22.7	25.0	27.3
M:NOUN:POSS	26	73.1	73.1	65.4
M:OTHER	160	43.1	46.9	35.6
M:PART	10	50.0	40.0	30.0
M:PREP	158	61.4	60.8	55.7
M:PRON	58	63.8	60.3	60.3
M:PUNCT	1104	85.7	81.2	78.2
M:VERB	59	55.9	49.2	47.5
M:VERB:FORM	23	91.3	87.0	87.0
M:VERB:TENSE	59	72.9	67.8	50.8
R:ADJ	93	36.6	39.8	25.8
R:ADJ:FORM	11	81.8	81.8	81.8
R:ADV	63	36.5	34.9	20.6
R:CONJ	7	28.6	14.3	14.3
R:CONTR	21	52.4	52.4	9.5
R:DET	178	52.8	51.7	46.1
R:MORPH	199	73.9	65.3	69.3
R:NOUN	290	41.4	37.9	22.8
R:NOUN:INFL	5	100.0	100.0	100.0
R:NOUN:NUM	247	80.6	73.3	73.3
R:NOUN:POSS	21	76.2	71.4	71.4
R:ORTH	325	84.3	75.4	77.8
R:OTHER	603	53.1	50.7	42.8
R:PART	45	68.9	64.4	42.2
R:PREP	496	72.2	69.2	52.0
R:PRON	85	54.1	56.5	51.8
R:PUNCT	293	75.1	64.8	34.1
R:SPELL	470	96.0	94.7	96.2
R:VERB	357	48.2	47.9	26.9
R:VERB:FORM	210	77.1	78.1	76.2
R:VERB:INFL	7	85.7	85.7	100.0
R:VERB:SVA	142	84.5	84.5	90.8
R:VERB:TENSE	373	65.7	63.8	45.8
R:WO	88	62.5	59.1	58.0
U:ADJ	11	27.3	18.2	18.2
U:ADV	50	40.0	32.0	32.0
U:CONJ	9	33.3	44.4	55.6
U:CONTR	8	87.5	75.0	37.5
U:DET	257	73.9	65.8	54.9
U:NOUN	46	45.7	39.1	30.4
U:NOUN:POSS	10	70.0	50.0	60.0
U:OTHER	81	33.3	33.3	28.4
U:PART	6	66.7	66.7	50.0
U:PREP	103	70.9	61.2	59.2
U:PRON	44	68.2	65.9	56.8
U:PUNCT	93	44.1	40.9	40.9
U:VERB	41	39.0	48.8	34.1
U:VERB:FORM	7	42.9	42.9	71.4
U:VERB:TENSE	47	57.4	61.7	48.9

Table 13: Correction rates: Percentage of errors fixed by each LLM on BEA dev. Count represents total errors for each error type. The exact LLMs are fine-tuned GPT-4o, fine-tuned LLaMA 3.3 and Zero-shot Deepseek V3.

Error Type	Count	GPT-4o False Insert %	Llama False Insert %	DeepSeek False Insert %
M:ADJ	9	52.9	52.9	61.5
M:ADV	19	52.2	73.9	61.3
M:CONJ	50	71.4	90.9	84.7
M:CONTR	1	100.0	50.0	0.0
M:DET	98	19.5	19.0	26.5
M:NOUN	12	41.2	52.2	45.5
M:NOUN:POSS	7	13.6	26.9	5.6
M:OTHER	57	34.6	34.2	50.4
M:PART	4	37.5	50.0	25.0
M:PREP	77	26.2	25.0	46.1
M:PRON	24	32.1	32.7	40.7
M:PUNCT	589	21.1	18.1	38.4
M:VERB	37	38.2	39.6	56.9
M:VERB:FORM	10	28.6	24.0	33.3
M:VERB:TENSE	31	38.6	39.4	50.8
R:ADJ	37	33.3	29.4	60.7
R:ADJ:FORM	4	30.8	0.0	30.8
R:ADV	19	45.2	37.1	53.6
R:CONJ	2	0.0	50.0	66.7
R:CONTR	16	20.0	20.0	88.9
R:DET	59	27.8	28.0	42.1
R:MORPH	99	23.7	27.2	42.1
R:NOUN	71	30.2	32.2	53.4
R:NOUN:INFL	1	16.7	0.0	16.7
R:NOUN:NUM	106	24.0	27.1	35.3
R:NOUN:POSS	8	27.3	34.8	25.0
R:ORTH	341	17.6	17.5	54.6
R:OTHER	227	27.7	33.4	49.8
R:PART	11	26.2	23.7	32.1
R:PREP	88	20.1	18.9	26.8
R:PRON	44	35.2	27.3	48.9
R:PUNCT	60	22.0	23.2	25.8
R:SPELL	59	6.7	6.0	11.1
R:VERB	40	19.3	18.9	30.8
R:VERB:FORM	100	24.2	24.8	38.6
R:VERB:INFL	5	25.0	14.3	41.7
R:VERB:SVA	83	27.0	27.3	38.6
R:VERB:TENSE	72	22.8	22.9	21.2
R:WO	37	34.5	39.8	42.0
U:ADJ	6	40.0	75.0	66.7
U:ADV	16	33.3	46.9	51.6
U:CONJ	9	40.0	42.9	64.3
U:CONTR	0	0.0	0.0	0.0
U:DET	86	24.0	28.2	37.6
U:NOUN	24	53.5	60.0	53.8
U:NOUN:POSS	1	0.0	16.7	0.0
U:OTHER	28	38.1	43.5	56.0
U:PART	3	42.9	42.9	50.0
U:PREP	34	27.7	34.3	34.4
U:PRON	20	16.7	17.6	43.5
U:PUNCT	92	41.2	42.2	70.8
U:VERB	17	40.7	35.5	54.8
U:VERB:FORM	18	40.0	57.1	78.3
U:VERB:TENSE	14	30.8	27.5	37.8

Table 14: False insertion rates: Percentage of false LLM corrections on BEA dev. Count represents maximum false insertions made by all LLMs for each error type. The exact LLMs are fine-tuned GPT-4o, fine-tuned LLaMA 3.3 and Zero-shot Deepseek V3.

Model Pair	Correction Rates		False Insertion Rates	
	Spearman	p-value	Spearman	p-value
GPT-4o vs. LLaMA 3.3	0.947	<0.0001	0.816	<0.0001
GPT-4o vs. DeepSeek V3	0.811	<0.0001	0.541	<0.0001
LLaMA 3.3 vs. DeepSeek V3	0.808	<0.0001	0.603	<0.0001
Mean	0.855	—	0.653	—

Table 15: Spearman rank correlations between models' per-error-type performance across 54 ERRANT error categories. All correlations significant at $p < 0.0001$.

Edit Code	Description
M:ADJ	Missing adjective - adjective present in the correction but absent in gold standard
M:ADV	Missing adverb - adverb present in the correction but absent in gold standard
M:CONJ	Missing conjunction - conjunction present in the correction but absent in gold standard
M:CONTR	Missing contraction - contraction present in the correction but absent in gold standard
M:DET	Missing determiner - determiner present in the correction but absent in gold standard
M:NOUN	Missing noun - noun present in the correction but absent in gold standard
M:NOUN:POSS	Missing possessive noun - possessive noun present in the correction but absent in gold standard
M:OTHER	Other missing words not falling into above categories
M:PART	Missing particle - particle present in the correction but absent in gold standard
M:PREP	Missing preposition - preposition present in the correction but absent in gold standard
M:PRON	Missing pronoun - pronoun present in the correction but absent in gold standard
M:PUNCT	Missing punctuation - punctuation present in the correction but absent in gold standard
M:VERB	Missing verb - verb present in the correction but absent in gold standard
M:VERB:FORM	Missing verb form - verb form present in the correction but absent in gold standard
M:VERB:TENSE	Missing verb tense - verb tense present in the correction but absent in gold standard
R:ADJ	Replacement of adjective - different adjective used in correction compared to gold
R:ADJ:FORM	Replacement of adjective form - different form of the same adjective
R:ADV	Replacement of adverb - different adverb used in correction compared to gold
R:CONJ	Replacement of conjunction - different conjunction used in correction compared to gold
R:CONTR	Replacement of contraction - different contraction used in correction compared to gold
R:DET	Replacement of determiner - different determiner used in correction compared to gold
R:MORPH	Replacement of morphology - different word form used in correction compared to gold
R:NOUN	Replacement of noun - different noun used in correction compared to gold
R:NOUN:INFL	Replacement of noun inflection - different inflected form of the same noun
R:NOUN:NUM	Replacement of noun number - singular/plural variation of the same noun
R:NOUN:POSS	Replacement of possessive noun - different possessive form used in correction
R:ORTH	Replacement of orthography - spelling or formatting differences
R:OTHER	Other replacements not falling into above categories
R:PART	Replacement of particle - different particle used in correction compared to gold
R:PREP	Replacement of preposition - different preposition used in correction compared to gold
R:PRON	Replacement of pronoun - different pronoun used in correction compared to gold
R:PUNCT	Replacement of punctuation - different punctuation used in correction compared to gold
R:SPELL	Replacement for spelling correction - corrected spelling of a word
R:VERB	Replacement of verb - different verb used in correction compared to gold
R:VERB:FORM	Replacement of verb form - different form of the same verb used in correction
R:VERB:INFL	Replacement of verb inflection - different inflected form of the same verb
R:VERB:SVA	Replacement for subject-verb agreement - verb changed to agree with subject
R:VERB:TENSE	Replacement of verb tense - different tense of the same verb used in correction
R:WO	Replacement for word order - reordering of words in correction compared to gold
U:ADJ	Unnecessary adjective - adjective present in gold standard but omitted in correction
U:ADV	Unnecessary adverb - adverb present in gold standard but omitted in correction
U:CONJ	Unnecessary conjunction - conjunction present in gold standard but omitted in correction
U:CONTR	Unnecessary contraction - contraction present in gold standard but omitted in correction
U:DET	Unnecessary determiner - determiner present in gold standard but omitted in correction
U:NOUN	Unnecessary noun - noun present in gold standard but omitted in correction
U:NOUN:POSS	Unnecessary possessive noun - possessive noun present in gold standard but omitted
U:OTHER	Other unnecessary words not falling into above categories
U:PART	Unnecessary particle - particle present in gold standard but omitted in correction
U:PREP	Unnecessary preposition - preposition present in gold standard but omitted in correction
U:PRON	Unnecessary pronoun - pronoun present in gold standard but omitted in correction
U:PUNCT	Unnecessary punctuation - punctuation present in gold standard but omitted in correction
U:VERB	Unnecessary verb - verb present in gold standard but omitted in correction
U:VERB:FORM	Unnecessary verb form - verb form present in gold standard but omitted in correction
U:VERB:TENSE	Unnecessary verb tense - verb tense present in gold standard but omitted in correction

Table 16: ERRANT Edit Type Descriptions.

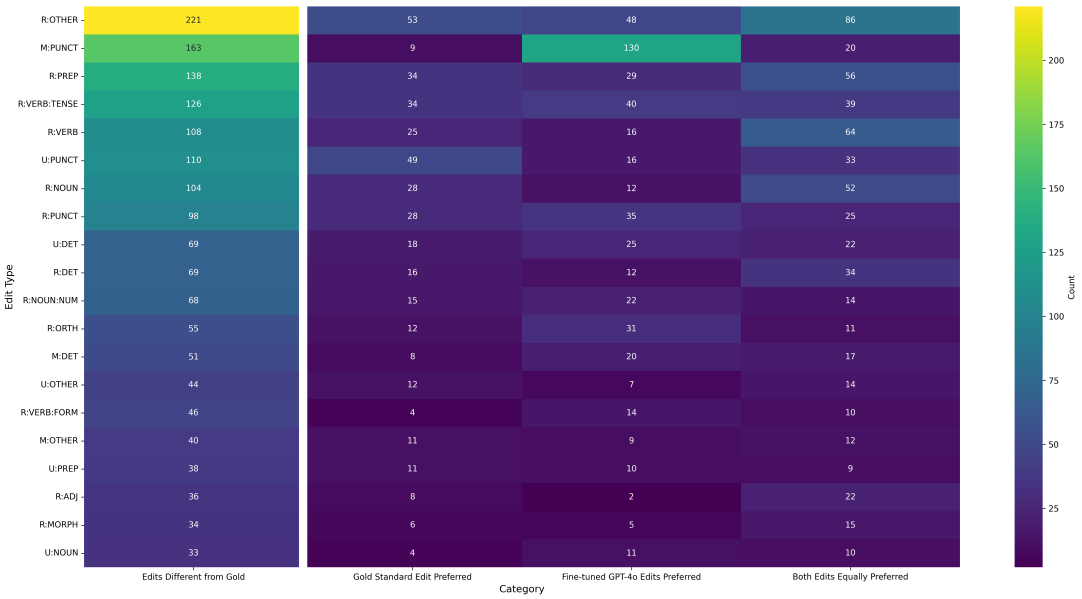

Figure 4: Edit Type Counts where fine-tuned GPT-4o produced different corrections compared to gold standard.