ChatLang-8: Can ChatGPT be used to generate Grammatical Error Correction Dataset?

Anonymous ACL submission

Abstract

In this paper, we investigate and enhance the LLM's ability to generate data for grammatical error correction (GEC). When simply generating parallel sentences, the patterns of them are too simple to have value as corpus. To alleviate this problem, we design an automatic framework that consists of Subject Selector, Grammar Selector, Prompt Manager, and Evaluator. We also introduce **ChatLang-8**, new dataset for GEC task. ChatLang-8 covers eight subject noun type and 23 grammar type and is composed of 1M pairs with a distribution of human-like grammatical errors. In experiments, we find that ChatLang-8 has a more uniform pattern composition than the existing GEC datasets. We also observe that the model performed better when trained with ChatLang-8 than with the existing GEC dataset. Experimental results demonstrate that our framework and ChatLang-8 will be a useful resource for improving ChatGPT's data generation capability. The ChatLang-8 dataset and code are available at Our github.

Method	Generated Subjects
Prompt (✖ SS)	{**'I'**: 42, **'She'**: 22, **'He'**: 11, **'Dog'**: 5, **'Cat'**: 4, **'Children'**: 2, **'Boys'**: 1, **'Dogs'**: 1, **'Book'**: 1, **'They'**: 1, **'Person'**: 1, **'Child'**: 1, **'You'**: 1, **'It'**: 1, **'House'**: 1, **'John and Sarah'**: 1, **'Restaurant'**: 1, **'Chair'**: 1, **'Elephant'**: 1, **'Friend'**: 1}
Prompt (✚ SS)	{**'I'**: 6, **'Book'**: 4, **'Patience'**: 4, **'Computer'**: 3, **'Water'**: 3, **'Monkeys'**: 3, **'Dog'**: 3, **'Headphone'**: 2, **'Tree'**: 2, **'Courage'**: 2, **'Flock'**: 2, **'Cat'**: 2, **'Car and Motocycle'**: 2, **'Firefighter'**: 2, **'Chair'**: 2, **'Everest'**: 2, **'Jellyfish'**: 2, **'Dogs'**: 2, **'Eiffel Tower'**: 1, **'Cups'**: 1, **'Wheelchair'**: 1, **'Compassion'**: 1, **'Guitar'**: 1, **'Freedom'**: 1, **'Perserverance'**: 1, **'Grace'**: 1, **'Camera'**: 1, **'She'**: 1, **'Authenticity'**: 1, **'Taj Mahal'**: 1, **'Elephants'**: 1, **'Children'**: 1, **'Windmill'**: 1, **'Road'**: 1, **'Shoe'**: 1, **'Happiness'**: 1, **'Refrigerator and Oven'**: 1, **'Pillow'**: 1, **'Soldiers'**: 1, **'Cupcake'**: 1, **'Airplane'**: 1, **'Chairs'**: 1, **'Bird'**: 1, **'Spoon'**: 1, **'Barack Obama'**: 1, **'Serenity and peace'**: 1, **'Great Barrier'**: 1, **'Thunderstorm'**: 1, **'Door'**: 1, **'Leaves'**: 1, **'Humility'**: 1, **'Mona Lisa'**: 1, **'Desk'**: 1, **'Snowman and Snowwoman'**: 1, **'John'**: 1, **'Pizza'**: 1, **'Place'**: 1, **'Herd'**: 1, **'Cows'**: 1, **'Bookshelf'**: 1, **'Toothbrush'**: 1, **'Respect'**: 1, **'Diversity'**: 1, **'Ink'**: 1, **'Beatles'**: 1, **'Plant'**: 1, **'Pride'**: 1, **'Wolves'**: 1, **'Congress'**: 1, **'Honesty'**: 1}

Table 1: 100 sentences were randomly sampled from each generated result.

1 Introduction

With the advent of large language models (LLM) (OpenAI, 2023a,c,b), fundamental paradigm shift is underway and many subfields of NLP such as machine translation (Jiao et al., 2023), summarization (Yang et al., 2023), question answering (Omar et al., 2023), and grammatical error correction (Wu et al., 2023; Fang et al., 2023) are being validated by LLM (Bang et al., 2023). However, there still lacks investigations into the ability of LLM to correct grammatical errors in text. For instance, ChatGPT gets much lower automatic evaluation scores (Wu et al., 2023) than GECToR (Omelianchuk et al., 2020), based on pretrained RoBERTa model, and GEC product (Grammarly, 2023). ChatGPT also suffers from *Overcorrection* problem (Fang et al., 2023) and shows much lower performance than TagGEC (Stahlberg and Kumar, 2021), T5 large, and T5 xxl (Rothe et al., 2021), despite its cutting-edge prompting techniques.

In order to solve these problems, it is indispensable to improve the performance of existing methods or leverage a hybrid model consisting of a GEC model and LLM, using high-quality training datasets. However, these approaches require a lot of manually annotated sentence pairs and these pairs are very expensive and even hard to obtain. In addition, LLM's data generation capabilities are not sufficiently corroborated and reinforced yet.

Contributions. In this paper, (i) we investigate LLM's ability to generate synthetic data, (ii) propose an automatic framework to bolster up the quality of generated data, and (iii) leverage a new GEC dataset, ChatLang-8, which is composed of 1M pairs with a distribution of human-like grammat-

Method	Generated Conjunction Errors
Prompt (✖ GS)	{**'And'**: 59, **'But'**: 38, **'Nor'**: 2, **'Or'**: 1}
Prompt (✚ GS)	{**'The...the'**: 6, **'Neither...nor'**: 6, **'Both…and'**: 4, **'So'**: 4, **'While'**: 4, **'When'**: 3, **'So...that:'**: 3, **'Yet'**: 3, **'Rather...than'**: 3, **'Now that'**: 3, **'And'**: 3, **'But'**: 3, **'As'**: 3, **'Nor'**: 2, **'Before'**: 2, **'As soon as'**: 2, **'As long as'**: 2, **'As well as'**: 2, **'Either...or'**: 2, **'Even if'**: 2, **'If'**: 2, **'Or'**: 2, **'Although'**: 2, **'Hardly...when'**: 1, **'Because'**: 1, **'In case'**: 1, **'Once'**: 1, **'Not...but'**: 1, **'Besides'**: 1, **'Since'**: 1, **'Whether...or'**: 1, **'If...then'**: 1, **'As many...as'**: 1, **'So that'**: 1, **'Not so...as'**: 1, **'Supposing'**: 1, **'Not only...but also'**: 1, **'Provided that'**: 1, **'For'**: 1, **'Till'**: 1, **'Even though'**: 1, **'Too...to'**: 1, **'Nonetheless'**: 1, **'As...as'**: 1, **'By the time'**: 1, **'Still'**: 1, **'Even if'**: 1, **'No sooner...than'**: 1, **'That'**: 1, **'Scarcely...when'**: 1, **'In order to'**: 1, **'If only'**: 1, **'Indeed'**: 1, **'Otherwise'**: 1, **'Though'**: 1}

Table 2: 100 sentences were randomly sampled from each generated result.

ical errors. We find that ChatLang-8 has a more uniform pattern composition than the existing GEC datasets. We also observe that the model performed better when trained with ChatLang-8 than with the existing GEC dataset with the same corpus size. To the best of our knowledge, we are the first to introduce a GEC dataset using LLM and to demonstrate it outperforms previous human-generated corpora.

2 ChatLang-8

The primary objective is to generate data that reflects human-like grammatical errors. To achieve this, we focus on creating parallel sentence pairs with distributions that resemble realistic human errors, rather than generating unnatural pairs (Grundkiewicz and Junczys-Dowmunt, 2014). A straightforward approach to generating such pairs would be to control all elements of the sentence. However, due to the productivity of language (O'Donnell et al., 2011), sentences can grow infinitely long through recursive structures. It is impractical to control all elements of the generated pairs, and such an approach may even degrade the quality of the data. Therefore, we hypothesize that the key factors influencing the quality of sentence pairs are the subject types and grammatical error types. Based on this hypothesis, we design a framework to diversify these factors systematically.

2.1 Subject Selector

A naive approach would involve designing prompts to generate sentence pairs without considering the subject element. However, as shown in Table 1, this results in a limited range of subject types in the generated sentences. To address this limitation, we propose Subject Selector (SS), a novel algorithm designed to diversify the subjects S in parallel sentences. First, we predefine subject types S_k[1]. The Subject Selector then randomly selects a subject type S_k from these predefined categories and generates subject candidates up to a defined window size. For instance, if the selected type is a proper noun and the window size is set to 30, the Subject Selector generates 30 proper noun candidates. A subject is randomly chosen from these candidates to be used in the sentence.

2.2 Grammar Selector

Similar to the issue discussed in Section 2.1, a basic prompt designed to generate pairs of incorrect and correct sentences results in an overly constrained set of outputs. As shown in Table 2, specifying only the grammar type fails to sufficiently address this problem. For instance, when generating data for *Conjunctions*, the majority of outputs consist solely of *And* and *But*. To resolve this, we propose the Grammar Selector (GS), which considers a diverse range of grammatical errors G to enhance the quality of the generated data. We adopt the same grammatical error categories G_l as in (Bryant et al., 2017, 2019). Unlike the Subject Selector described in Section 2.1, the Grammar Selector first categorizes grammatical errors into two types: those requiring diversification of grammar types and those that do not. For example, errors such as *Word Order (WO)* and *Spelling (SPELL)* do not need extensive pattern variation; accurate results can be achieved by instructing the LLM to simulate typical human errors.

2.3 Prompt Manager

Given a subject and grammatical error generated from Section 2.1 and Section 2.2, respectively, the Prompt Manager combines them into a single prompt. To enhance production stability, we integrate the Chain-of-Thought (CoT) technique (Wei et al., 2022) into the Prompt Manager. To avoid unnecessary errors, the Prompt Manager ensures that the subject in the incorrect sentence is aligned with the subject in the correct sentence. All other elements are kept consistent, except for the grammatical error introduced to maintain parallelism.

[1]Predefined types include: Common Noun, Proper Noun, Collective Noun, Compound Noun, Concrete Noun, Abstract Noun, Countable Noun, and Uncountable Noun.

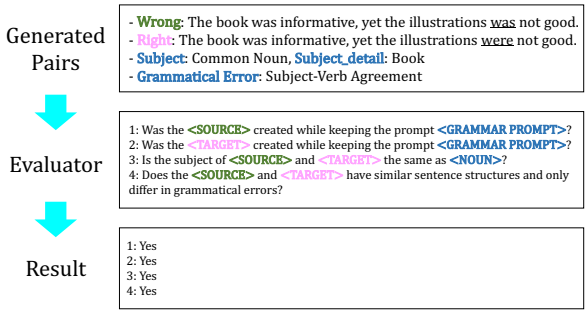

Figure 1: The pipeline of Evaluator.

2.4 Evaluator

While the generated pairs can exhibit diversity by leveraging the Subject Selector, Grammar Selector, and Prompt Manager, the uniform randomness introduced to encourage variation occasionally results in outputs that violate prompt instructions. For example, a prompt specifying *subject: Concrete Noun (car)* and *grammar type: Verb (swim)* may yield pairs such as *"Wrong: The car swims to the shore.", "Right: The car drives to the shore."*. This is a result of incorrectly focusing on semantic errors instead of adhering to the intended verb tense or active/passive voice errors. To mitigate this issue, we propose the Evaluator, a lightweight LLM-based mechanism that assesses generated data against four predefined criteria C (the detailed pipeline of the Evaluator is presented in Figure 1.

The Evaluator $\mathcal{E}$ ensures compliance with the data generation process through a unanimous voting mechanism: if any criterion C_i is unmet, the generated output is discarded:

$$\mathcal{E}(D) = \begin{cases} 1, & if \quad C_1 \wedge C_2 \wedge C_3 \wedge C_4 \\ 0, & otherwise \end{cases} \tag{1}$$

2.5 Dataset Details

Our implementation utilizes gpt-3.5-turbo-1106. We gathered approximately 1M pairs that cover eight subject type and 23 grammar type. Table 3 shows statistics of *ChatLang-8* (ours), FCE (Yannakoudakis et al., 2011), Lang-8 (Mizumoto et al., 2012; Tajiri et al., 2012), NUCLE (Dahlmeier et al., 2013) and W&I+LOCNESS (Bryant et al., 2019). Note that ChatLang-8 has a corpus size similar to Lang-8 and much larger than the other corpora.

3 Experiments

3.1 Statistical Analysis

Note that numerous studies show that the imbalance of a dataset causes performance degradation

S: In India we have various types of Public transport , like Cycle , Bike , Car , Train & Flight .

[Transformer + Lang-8]
A: -1 -1|||noop|||-NONE-|||REQUIRED|||-NONE-|||0

[Transformer + ChatLang-8]
A: 2 2|||M:PUNCT|||,|||REQUIRED|||-NONE-|||0
A: 7 8|||R:ORTH|||public|||REQUIRED|||-NONE-|||0
A: 18 18|||M:PUNCT|||,|||REQUIRED|||-NONE-|||0
A: 18 19|||R:CONJ|||and|||REQUIRED|||-NONE-|||0

S: If I had it to choose between both transportation I think I probably chose car because it is better for me to go by car then to go by bus .

[Transformer + Lang-8]
A 3 4|||U:PRON||||||REQUIRED|||-NONE-|||0
A 13 14|||R:VERB:TENSE|||would have chosen|||REQUIRED|||-NONE-|||0
A 26 28|||U:VERB||||||REQUIRED|||-NONE-|||0

[Transformer + ChatLang-8]
A 3 4|||U:PRON||||||REQUIRED|||-NONE-|||0
A 9 9|||M:PUNCT|||,|||REQUIRED|||-NONE-|||0
A 10 12|||R:OTHER|||would|||REQUIRED|||-NONE-|||0
A 13 14|||R:VERB:TENSE|||choose|||REQUIRED|||-NONE-|||0
A 25 27|||R:OTHER|||than|||REQUIRED|||-NONE-|||0

Figure 2: Comparison of M^2 outputs of GEC models, trained on Lang-8 and ChatLang-8 respectively.

(Huang et al., 2021; Lin et al., 2023; Li and Shi, 2021). In order to demonstrate whether the dataset is created evenly, we conduct statistical analysis by comparing ChatLang-8 with existing datasets. Figure 3 shows a subject distribution of ChatLang-8. We observe that the distribution of subject types in ChatLang-8 is evenly distributed. It suggests that the Subject Selector and evaluator affect data quality improvement even though the evaluator abandons all unsuitable data. In Figure 4 and Table 5, we find that the grammar types of ChatLang-8 are most evenly distributed while the other corpora have highly unbalanced grammar labels. For example, W&I+LOCNESS dataset has 10.7% (averaged) of Determiner (DET) errors, and 17.75% of Punctuation (PUNCT) errors, whereas 0.6133% of Conjunction (CONJ) and 0.21% of Adjective Form (ADJ:FORM). Existing datasets have a very small proportion of important grammatical patterns, which can cause performance degradation during model training. In addition, since ChatLang-8 excluded OTHER and UNK labels when generating data, while the other corpora have a great propor-

	ChatLang-8	FCE	Lang-8	NUCLE	W&I+LOCNESS
Sentences	1,016,588	33,236	1,037,561	57,151	43,169

Table 3: ChatLang-8, FCE, Lang-8, NUCLE, and W&I+LOCNESS dataset statistics.

CoNLL-2014	**ChatLang-8**			**Lang-8**		
	P	**R**	**$F_{0.5}$**	**P**	**R**	**$F_{0.5}$**
Transformer	47.79	26.39	41.12	54.37	13.28	33.58
BART	52.50	34.31	47.47	27.89	16.01	24.29

BEA-2019	**ChatLang-8**			**Lang-8**		
	P	**R**	**$F_{0.5}$**	**P**	**R**	**$F_{0.5}$**
Transformer	34.44	19.69	29.95	44.81	8.5	24.16
BART	34.86	24.38	32.10	23.69	11.00	19.25

GMEG-wiki	**ChatLang-8**			**Lang-8**		
	P	**R**	**$F_{0.5}$**	**P**	**R**	**$F_{0.5}$**
Transformer	46.56	34.33	43.46	41.44	11.38	27.11
BART	49.09	41.77	47.42	23.54	15.68	21.39

Table 4: Evaluation results on English CoNLL-2014, BEA-2019, and GMEG-wiki.

tion of UNK label (e.g., NUCLE: 25.65%) and OTHER label (e.g., NUCLE: 2.57%), it is easier to handle dataset and to predict the model outputs.

3.2 Quantitative Results

To investigate whether ChatLang-8 has value as a dataset, we train a vanilla Transformer (Vaswani et al., 2017) and BART (Lewis et al., 2020), much larger model, on both ChatLang-8 and Lang-8 (Mizumoto et al., 2012; Tajiri et al., 2012). Note that for the fairness of the experiment, we choose Lang-8, which has a similar dataset structure and corpus size, and which effectively improve performance (Lichtarge et al., 2019, 2020; Flachs et al., 2021). In addition, though there is previous attempt to improve the Lang-8 (Flachs et al., 2021), it reports only five error types on BEA test while we considers 25 error types and three benchmarks. Thus, we employ Lang-8 to the experiments, instead of cLang-8. We evaluate performance on three popular GEC benchmarks: CoNLL-2014 (Ng et al., 2014), BEA-2019 (Bryant et al., 2019), and GMEG-wiki (Napoles et al., 2019). As evaluation metrics, we use M^2 scorer (Dahlmeier and Ng, 2012) for CoNLL-2014, and ERRANT (Bryant et al., 2017) for BEA-2019 and GMEG-wiki. Table 4 shows that GEC models trained on ChatLang-8 can surpass its counterparts which are trained on Lang-8 corpus, by a large margin in both recall (**R**) and $\mathbf{F}_{0.5}$. It represents that the GEC system trained with ChatLang-8 can simultaneously detect multiple types grammatical errors in a sentence and correct them, while system trained with Lang-8 can't. Simply a high **R** and a low $\mathbf{F}_{0.5}$ indicate overcorrection. We observe that transformer learned with Lang-8 shows higher precision (**P**) than its counterpart on CoNLL-2014 and BEA-2019 benchmark evaluations, entailing much lower $\mathbf{F}_{0.5}$. This means that it corrects only very obvious grammatical errors and leaves the rest uncorrected. We also find that leveraging a high-quality GEC corpus is overarching problem, as merely enlarging the model does not always ensure improved performance.

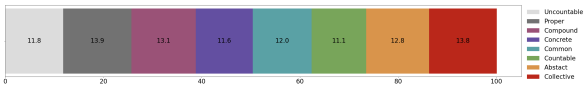

Figure 3: Subject distribution of ChatLang-8.

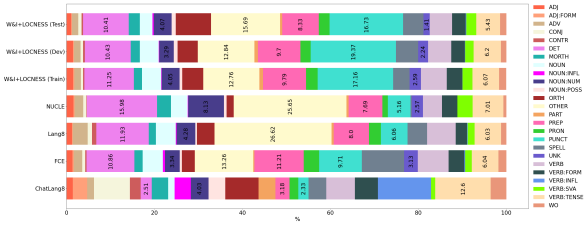

Figure 4: Distributions of the ChatLang-8 and the other datasets with respect to 25 grammatical errors.

3.3 Qualitative Results

Figure 2 shows an example and its corresponding M^2 outputs, which qualitatively demonstrate the effect of dataset used for training, as shown in Section 3.2. We observe that the learned model using ChatLang-8 can detect and correct multiple errors including punctuation (PUNCT), orthography (ORTH), and conjunction (CONJ) while its opponent can not find any errors. Indeed, the output of model learned with ChatLang-8 is *In India, we have various types of public transport, like Cycle, Bike, Car, Train, and Flight*. Results in Figure 2 demonstrates that ChatLang-8 has the error distribution shown in Table 5 very well, and thereby efficiently helps the learning of the GEC system.

4 Conclusion

We hypothesize that determining the quality of dataset is the types of subject and grammatical error. We introduce a high-quality dataset for GEC task that is composed of 1M pairs and covers eight subject type and 23 grammar type. We are the first to build a GEC dataset using ChatGPT and verify it outperforms existing corpora. We hope that our method and dataset can help researchers make a better-informed decision about their research.

Limitations

Factuality and Morality. Though ChatLang-8 effectively can reflect a distribution of human-like grammatical errors in Figure 5 and guide a GEC model to perform well, due to uniform randomness of our method to encourage diversity, sometimes generated pairs have information that is literally not true. For instance, in Section 2.1, Subject Selector, of course, can generate celebrity (e.g., *Michael Jackson* or *Taylor Swift*), or even company name (e.g., *Google* or *Amazon*). By combining the selected subject and grammatical information, sentences that are not true can be generated as follows:

> *Michael Jackson was known for his music career, in order that he could be more successful, he decided to try his hand at acting.*

> *Taylor Swift is really good at using a computer.*

In addition, the generated sentences may contain biased information such as subjective opinions:

> *Barack Obama's speeches were always inspiring the crowds.*

> *Google is the best search engine in the world.*

Note that, although most of pairs are normal and fabricated information is found in some generated examples, this factuality problem must be resolved.

On the other hand, we also observed whether ChatLang-8 has a problem with moral deviations. Fortunately, no morality issue was found in most of the pairs constituting ChatLang-8:

> *The equality between races, genders, and sexual orientations is important.*

> *The importance of maintaining equality among all genders is evident in society.*

However, we found one sentence and removed from ChatLang-8:

> *The equality of the sexes is not a big issue.*

These observations suggest that semantic information as well as grammatical information of the generated sentence should be added to the evaluation criteria of our Evaluator.

Time and Cost. As mentioned in Section 2.5, Our implementation utilizes gpt3.5-turbo version of ChatGPT. It took about two weeks to gather ChatLang-8. To collect 1M pairs dataset, we spent about 1.1K *dollars*. However, It should be noted that the cost was higher than expected because some pairs are discarded by Evaluator. Evaluator disposed of about 0.7M sentences out of a total of 1.7M generated sentences. We consider that though it is much more efficient than having annotators create the dataset, this a great waste of time and resources. Also, there is a trade-off in our framework: The larger window size, the more diverse and high-quality dataset can be obtained, but the time and cost to create dataset increases accordingly.

Discussion and Future Works. In this paper, we strived to validate the ability of LLM to generate datasets. Of course, efficient dataset acquisition and model training are important challenges in the GEC task, but in fact, we also wanted to know what information is important and should be considered in Natural language generation in order to improve quality of outputs generated by LLM. We hope that this study and its results will be helpful to researchers who have the same concerns as us.

Our direction of future work is threefold:

1. We will establish new criteria for factuality and morality, and improve the evaluator through methods such as giving penalties for the generated results that violate the criteria.

2. We also will devise methods such as soft alignment to minimize the amount of data discarded by Evaluator to prevent resource waste.

3. Though we conducted an in-depth analysis through statistical, quantitative, and qualitative methods, we relied on LLM for data collection, leaving human evaluation behind. We will experiment with more off-the-shelf baselines by increasing the size of the corpus, and human evaluation will also be conducted on the generated results.

Ethical Considerations

There are no ethical issues in this study. The data was collected through legitimate methods. Moreover, the experimental results were obtained through an objective comparison. The authors confirm that the research was conducted in accordance with the relevant ethical guidelines and principles.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

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, Vancouver, Canada. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2304.01746*.

Simon Flachs, Felix Stahlberg, and Shankar Kumar. 2021. Data strategies for low-resource grammatical error correction. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 117–122, Online. Association for Computational Linguistics.

Grammarly. 2023. Grammarly website about us page.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2014. The wiked error corpus: A corpus of corrective wikipedia edits and its application to grammatical error correction. In *International Conference on Natural Language Processing*, pages 478–490.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Piji Li and Shuming Shi. 2021. Tail-to-tail non-autoregressive sequence prediction for Chinese 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 1: Long Papers)*, pages 4973–4984, Online. Association for Computational Linguistics.

Jared Lichtarge, Chris Alberti, and Shankar Kumar. 2020. Data weighted training strategies for grammatical error correction. *Transactions of the Association for Computational Linguistics*, 8:634–646.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora generation for grammatical error correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3291–3301, Minneapolis, Minnesota. Association for Computational Linguistics.

Jionghao Lin, Wei Tan, Ngoc Dang Nguyen, David Lang, Lan Du, Wray Buntine, Richard Beare, Guanliang Chen, and Dragan Gasevic. 2023. Robust educational dialogue act classifiers with low-resource and imbalanced datasets.

Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2012. The effect of learner corpus size in grammatical error correction of ESL writings. In *Proceedings of COLING 2012: Posters*, pages 863–872, Mumbai, India. The COLING 2012 Organizing Committee.

Courtney Napoles, Maria Nădejde, and Joel Tetreault. 2019. Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses. *Transactions of the Association for Computational Linguistics*, 7:551–566.

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.

Timothy O'Donnell, Jesse Snedeker, Joshua Tenenbaum, and Noah Goodman. 2011. Productivity and reuse in language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 2023. Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint arXiv:2302.06466*.

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.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. no date provided. introducing assistants api.

OpenAI. 2023c. no date provided. introducing chatgpt.

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.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202, Jeju Island, Korea. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

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.

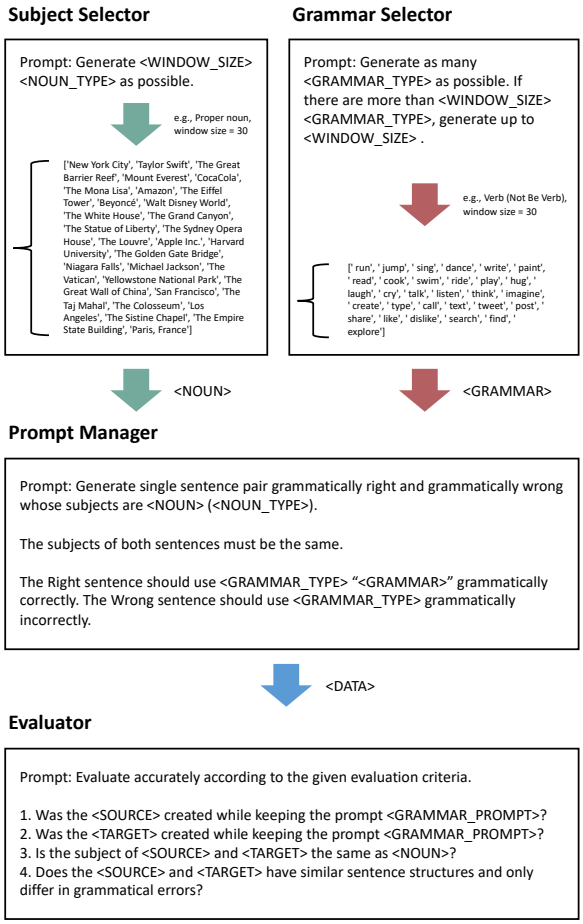

Figure 5: Schematic depiction of overall architecture

A Statistical Analysis: Error Type Distributions

Table 5 shows the ERRANT error type distributions of ChatLang-8 and The Other Corpora. We find that ChatLang-8 are most evenly distributed while the other corpora have highly unbalanced grammar labels.

B Architecture

As shown in Figure 5, our framework consists of Subject Selector, Grammar Selector, Prompt Manager, and Evaluator. Subject Selector and Grammar Selector ensure the diversity of the generated data, Prompt Manager maintains the format of the generated data, and Evaluator manages the quality of the generated data.

	ChatLang-8	FCE	Lang-8	NUCLE	W&I+LOCNESS (Train)	W&I+LOCNESS (Dev)	W&I+LOCNESS (Test)
ADJ	1.44	1.36	1.25	1.58	1.52	1.48	1.05
ADJ:FORM	3.26	0.28	0.19	0.27	0.24	0.21	0.18
ADV	1.55	1.94	3.37	1.95	1.51	1.51	1.45
CONJ	8.18	0.67	0.98	0.71	0.51	0.58	0.75
CONTR	2.44	0.32	0.99	0.11	0.3	0.39	0.32
DET	2.51	10.86	11.93	15.98	11.25	10.43	10.41
MORTH	3.73	1.9	1.62	3.14	1.85	2.07	2.5
NOUN	1.46	4.57	4.51	3.8	4.36	4.3	2.89
NOUN:INFL	3.69	0.5	0.18	0.12	0.12	0.13	0.28
NOUN:NUM	4.03	3.34	4.28	8.13	4.05	3.29	4.07
NOUN:POSS	3.83	0.51	0.35	0.61	0.6	0.87	0.93
ORTH	7.56	2.94	3.99	1.62	4.77	4.61	8.03
OTHER	0	13.26	26.62	25.65	12.76	12.84	15.69
PART	3.83	0.29	0.5	0.46	0.84	0.79	0.49
PREP	3.18	11.21	8	7.69	9.79	9.7	8.33
PRON	2	3.51	2.72	1.26	2.64	2.33	2.45
PUNCT	2.33	9.71	6.06	5.16	17.16	19.37	16.73
SPELL	4.03	9.59	4.45	0.26	3.74	5.07	4.63
UNK	0	3.13	0	2.57	2.59	2.24	1.41
VERB	6.54	7.01	6.52	4.31	5.86	5.27	5.09
VERB:FORM	5.23	3.55	2.56	3.49	3.56	3.09	3.1
VERB:INFL	12.07	0.19	0.15	0.01	0.04	0.07	0.12
VERB:SVA	0.95	1.52	1.58	3.47	2.23	1.94	2.28
VERB:TENSE	12.6	6.04	6.03	7.01	6.07	6.2	5.43
WO	3.58	1.82	1.18	0.66	1.64	1.25	1.4
Total (%)	100	100	100	100	100	100	100

Table 5: Distributions of the ChatLang-8 and the other datasets with respect to 25 grammatical errors.