Correcting Challenging Finnish Learner Texts With Claude, GPT-3.5 and GPT-4 Large Language Models

Mathias Creutz
Department of Digital Humanities, University of Helsinki, Finland
mathias.creutz@helsinki.fi

Abstract

This paper studies the correction of challenging authentic Finnish learner texts at beginner level (CEFR A1). Three state-of-the-art large language models are compared, and it is shown that GPT-4 outperforms GPT-3.5, which in turn outperforms Claude v1 on this task. Additionally, ensemble models based on classifiers combining outputs of multiple single models are evaluated. The highest accuracy for an ensemble model is 84.3 %, whereas the best single model, which is a GPT-4 model, produces sentences that are fully correct 83.3 % of the time. In general, the different models perform on a continuum, where grammatical correctness, fluency and coherence go hand in hand.

1 Introduction

The motivation behind the present work is to help second-language (L2) learners express themselves fluently and idiomatically in a non-native language that they do not master very well. The problem can be studied through the automatic correction of challenging learner texts that contain numerous mistakes when it comes to inflection, spelling, word choice, word order and even low intelligibility overall. Previously, neural machine translation with different data augmentation techniques have been employed to solve this task (Sjöblom et al., 2021), but the advent of powerful large language models (LLMs) opens up new possibilities to tackle the problem.

Bryant et al. (2023) present an overview of the state of art in Grammatical Error Correction (GEC). The term *grammatical* is understood broadly and does not only refer to grammatical errors. However, GEC is typically seen as a *local* substitution task (Ye et al., 2023), where occasional mistakes are corrected in generally intelligible text. The survey covers methods and data sets (predominantly in English). The article was written before the breakthrough of GPT-3.5 and GPT-4, and observations regarding LLMs are therefore limited. Some small-scale experiments are mentioned (Wu et al., 2023; Coyne et al., 2023), concluding that LLMs tend to overcorrect for fluency, which causes them to underperform on datasets that were developed for minimal corrections (Fang et al., 2023). By contrast, Penteado and Perez (2023) find that LLMs outperform earlier methods on more challenging texts, typed in a hurry or containing slang, abbreviations, and neologisms.

The main goal of this paper is to study how well state-of-the-art large language models are capable of rephrasing beginner-level learner texts into idiomatic, correctly formulated texts. As advocated by Sakaguchi et al. (2016), the focus is not on the detection and correction of specific errors in isolation, but on the fluency and naturalness of entire correction hypotheses. As ensemble models have proven effective in earlier GEC tasks (Grundkiewicz and Junczys-Dowmunt, 2018; Li et al., 2019; Bryant et al., 2019), additional experiments are carried out, where multiple model outputs are combined.

2 Data

A subset of ICLFI, the International Corpus of Learner Finnish (Jantunen, 2011; Jantunen et al., 2013) is used as data for the experiments. [1] A random selection of 25 texts were selected for the study, all of them labeled with the lowest language proficiency level: CEFR A1.[2] The A1 level was chosen in order to obtain as challenging data as possible. Table 1 shows one text extracted from this data, with an approximate English translation. The total number of sentences in all 25 texts is 210.

Some English learner corpora, such as FCE (Yannakoudakis et al., 2011) and NUCLE (Dahlmeier

[1] Available online through the Language Bank of Finland: https://www.kielipankki.fi/corpora/iclfi/

[2] https://www.coe.int/en/web/common-european-framework-reference-languages

Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024), pages 1–10

Minä lulee etttä, Anna on nyt niin erilainen kuin tavallisesti, koska hänellä on stressi. Anna ei ole aikaa puhumaan Jutan kanssa, koska korjata tule hänen kotiinsa. Annalla ei ole siihen jokin hyvä syy, koska pesukone on rikki, pesukone on siihen jokin hyvä syy. Minusta Anna on kateellinen, koska Juttasta Anssi on hauska mies.	I belives thatt, Anna is now so different than usually, because she is stressed. Anna is no time talking with Jutta, because repair come to her house. Anna has not some good reason for this, because the laundry machine is broken, the laundry machine is a good reason for that. I think Anna is jealous, because according Jutta Anssi is a fun guy.

Table 1: An example text from the ICLFI corpus (CEFR level A1). The Finnish text is on the left with an approximate English translation on the right. The intended meaning is not entirely clear, because one sentence contradicts itself.

et al., 2013) contain reference corrections that can be utilized for evaluation, but that is unfortunately not the case with the ICLFI corpus.[3] TopLing (University of Jyväskylä, 2016) is another Finnish learner corpus that lacks correction hypotheses. There used to exist an additional resource, the so-called YKI corpus based on Finnish national certificates of language proficiency exams (Yleiset kielitutkinnot), but it is no longer available because of copyright issues.

3 Models

Three different commercial LLM systems were tested in this study: Claude v1 by Anthropic[4], as well as GPT-3.5 (turbo) and GPT-4 by Open AI (OpenAI, 2023).[5] Claude may be an interesting complement to the GPT models, as it has been seen to outperform ChatGPT (GPT-3.5) in certain open-domain conversation tasks (Lin and Chen, 2023).

The LLMs were accessed through their APIs, Claude at the end of June and GPT-3.5 and GPT-4 at the end of July and beginning of August 2023. The models were prompted to reformulate the learner texts into fluent, impeccable Finnish language that contains no factual or grammatical errors. The exact prompts used can be found in Appendix A. Each prompt contained an entire text in order for the model to be able to exploit context across sentence boundaries.

LLMs are non-deterministic. The temperature parameter ranging between 0 and 1 regulates the randomness of the output. Low temperatures result in the most predictable result, whereas higher temperatures increase creativity.[6]

Each of the LLMs was tested on six different temperature values: 0.0, 0.1, ..., 0.5. Even with the lowest temperature of 0.0, the systems were not fully deterministic, and some variability remained in the output. Every configuration was run twice, because of the non-deterministic nature of the task. These runs were confirmed not to depend on the outcome of the previous run (see Appendix B). This resulted in 36 correction hypotheses for each of the 25 texts (3 LLMs times 6 temperature values times 2 runs each). In the following, these 36 setups will be referred to as *models* or *single models*.

4 Annotation

The 36 correction hypotheses produced by the LLMs for each of the 25 learner texts were manually tagged as correct or incorrect. The tagging was performed on the sentence level: either a sentence was fully correct or it was incorrect, considering the context of surrounding sentences.

The annotation was performed independently by two persons, the author of the paper and one of his colleagues. The annotators could see the full original text and the suggested corrections, sentence by sentence. When multiple models had produced the same sentence in the same context, it was sufficient to annotate that sentence only once. Theroretically, there would have been $36 * 210 = 7560$ sentences to annotate, but because of duplicates, the actual number was reduced to one fifth of that.

Initially, the annotators agreed in 83.9 % of the cases (type count, after sentence deduplication). This corresponds to 87.5 % of all generated sentences (token count). In a second round, the annotators discussed the results and decided which category to choose for the remaining cases. The main reasons for initial disagreement were minor errors that had gone unnoticed by either annotator, different levels of tolerance for the incorrect use of punctuation,[7] and confusion about the intended meaning of the original sentence.

[3] In fact, ICLFI has been automatically lemmatized and parsed, and some of the misspelled words have been corrected in the process, but this representation is not accurate enough to be used as a proper reference.

[4] https://claudeai.pro/what-is-claude-v1/

[5] https://platform.openai.com/

[6] https://platform.openai.com/docs/guides/gpt/how-should-i-set-the-temperature-parameter

[7] In the end, we decided not to be very strict about comma rules.

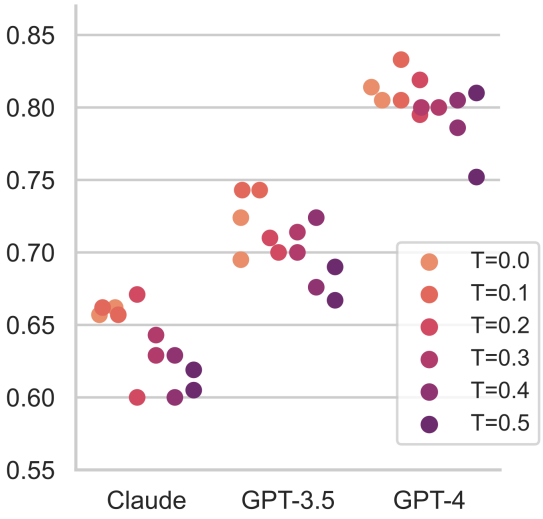

Figure 1: Accuracies of each of the 36 single models. Every model is represented by a dot, and the dots are grouped in "swarms" by LLM type. In every swarm, we progress from left to right as the temperature (T) rises, with higher temperatures rendered in darker color. The best model (GPT-4, $T = 0.1$, 1st run) reaches an accuracy of 0.833, which corresponds to 175 fully correct sentences out of 210 in the data.

5 Single Model Results

The accuracies of the 36 single models have been plotted in Figure 1. The results reveal two things: Firstly, there are clear differences in the performance levels of the LLMs. All GPT-4 models are better than all GPT-3.5 models, which are in turn better than all Claude models (with the exception of the one weakest GPT-3.5 model). Secondly, the temperature parameter works as expected. Conservative, predictable results are to be preferred in this correction task, and thus lower temperatures work better than higher temperatures. However, the best results are in general obtained for $T = 0.1$, not the lowest possible value $T = 0.0$.

In line with these findings, Coyne et al. (2023) observe that GPT-4 outperforms GPT-3.5 on English GEC data (Napoles et al., 2017; Bryant et al., 2019). They also confirm that a low temperature yields better performance in this task.

In previous work on Finnish GEC (Creutz and Sjöblom, 2019), an annotated sample of the (since then withdrawn) YKI corpus was used as test data. The full-sentence accuracy obtained for the best setup was 27.2 %, which falls far behind the accuracies in Figure 1. Direct comparisons cannot be made because of the different corpora used in the studies. However, the types and levels of the texts are very similar.

Hypotheses	Proposed by models									...		La-
	1	2	3	4	5	6	7	8	9	...	36	bel
How are you?	●		●		●	●						✓
How you are?											●	✗
How are things?				●				●				✓
What are you like?		●							●			✗
How old are you?							●					✗

Figure 2: Possible correction hypotheses for a fictive sentence *"How yuo are?"* (in English for illustration purposes). Among other things, we see that models 1, 3, 5 and 6 propose the first correction hypothesis *"How are you?"*, which is correct, whereas model 36 proposes *"How you are?"*, which is incorrect. From this example we get five data entries to train a supervised classification model. The inputs consist of 36-dimensional binary vectors, where every dimension corresponds to one of the single models and is zero or one depending on whether that model produced this particular hypothesis. The outputs are binary as well, indicating whether the hypothesis is correct or not.

6 Ensemble Models

The best single model produces 175 correct sentences out of 210 (83.3 %). However, if we look at all 36 models combined, there are only 7 sentences that all models get wrong. This suggests that by being very smart at combining sentences from different models, we could ideally reach an accuracy of 203/210 (96.7 %).

In the following, we will study supervised learning of ensemble models that combine outputs from the single models. The simplifying assumption is made that sentences from different hypotheses can always be combined. For instance, the two partly correct texts *"Hi there! How's you?"* and *"Helo! How are you?"* can be combined coherently into *"Hi there! How are you?"*.

The problem is formulated as a classification task. For every input sentence, each of the 36 models has produced a correction hypothesis, but typically the number of unique hypotheses is lower than 36, because several models produce the same hypotheses. This is exploited by a classifier, which is trained to predict when a hypothesis is correct based on the subset of models that have proposed it, as illustrated in Figure 2.

As there is limited amount of data available, rather than setting aside a separate test set, cross-validation is used, such that every learner text in

turn serves as the test set and the remaining 24 texts are used for training. In this way, test results are obtained for all 25 texts and direct comparisons can be made to the single model results (Figure 1). The feature extraction (Figure 2) produces 1532 vectors in total. As one text is left out in turn, on average 1470 vectors (24/25) are available for training.

6.1 Classifiers Used

The limited amount of data available calls for fairly simple classifiers with a small numbers of parameters to tune, in order to avoid overfitting.

Naive Bayes. (NLTK implementation, Bird et al., 2019) This classifier is not very sensitive to the size of the data set, because the training amounts to solving a closed-form expression. However, the underlying independence assumption may lead to the exaggeration of correlated features.

Maximum Entropy. This is logistic regression using the Maximum Entropy classifier of NLTK. Conditional independence is not assumed, but the lack of a closed-form solution may lead to suboptimal weights in the model.

Weighted Sum. This is a simplified, deterministic alternative to Maximum Entropy. A weight vector w of the same dimensionality as the binary correction hypothesis vectors x is estimated. During prediction, the hypothesis with the highest score s is selected: $s = w \cdot x$. The elements w_i of w correspond to the prominence of the ith model in the weighted sum and is proportional to the number of times that model has predicted a correct hypothesis, divided by the total number of models that predicted the same hypothesis. This mitigates the effect of correlated features.

N Agreeing Models An asymmetric decision tree is trained in order to explicitly model correlated features. The tree branches onto one side only ("`if` *condition 1* `then done else if` *condition 2* `then done ... else done`").

The conditions correspond to all combinations of $2 .. N$ models that are more accurate than the best single model when they are in agreement on what hypothesis to propose. These model combinations are sorted, most accurate first. The last fallback condition was originally the best single model, but was later replaced by the Naive Bayes classifier for better performance.

N values ranging from 2 to 5 have been tested. For higher values of N, all lower-order combinations of models are also included. The results for $N = 5$ turn out to be identical to those of $N = 4$.

For the pairs of models ($N = 2$), a minor variant ($N = 2^*$) was tested as well. In the basic case, the sorting order of the conditions is statically determined from the entire training set, whereas the extended version ($N = 2^*$) incrementally recalculates accuracies on the remainder of the training set, from which data points that triggered previous conditions in the chain have been removed.

6.2 Ensemble Model Results

If all single models are combined into ensemble models, only one of the resulting ensembles (N Agreeing Models with $N = 2^*$) outperforms the best single model (see Appendix C). The best ensemble obtains an accuracy of 0.838, compared to the best single model: 0.833. This is a rather insignificant improvement.

We have observed that the Claude models perform worst in the task and that low temperatures are to be preferred. By excluding the Claude models and temperatures above 0.3, the results in Figure 3 are obtained. Now, the advantage between the best ensemble model (Weighted Sum) and best single model is slightly larger (0.843 vs. 0.833). In other words, the sentence error rate is reduced by 6.0 %. This is the best result of all trials involving different combinations of single models. The theoretical upper bound on accuracy by an oracle model would be 0.967. None of the ensembles reach accuracies even close to that. Further analysis can be found in Appendix C.

Finding related work on ensemble models built on GPT-3.5 or GPT-4 is hard, and none of it addresses the GEC task. Work by Jiang et al. (2023), Yuan et al. (2023), Fu et al. (2023), Manakul et al. (2023), García-Díaz et al. (2023), and Portillo Wightman et al. (2023) relate to other NLP tasks, such as summarization, sentiment analysis and question answering. Tang et al. (2023) create ensembles of less advanced pre-trained language models (BART, BERT, GPT-2 etc.) for Chinese GEC, but fail to outperform the best single models.

7 Qualitative Evaluation

When the generated hypotheses were tagged as correct or incorrect, it was not known to the annotators which model had produced them. Therefore, no systematic qualitative evaluation of the differences between Claude, GPT-3.5 and GPT-4 is available.

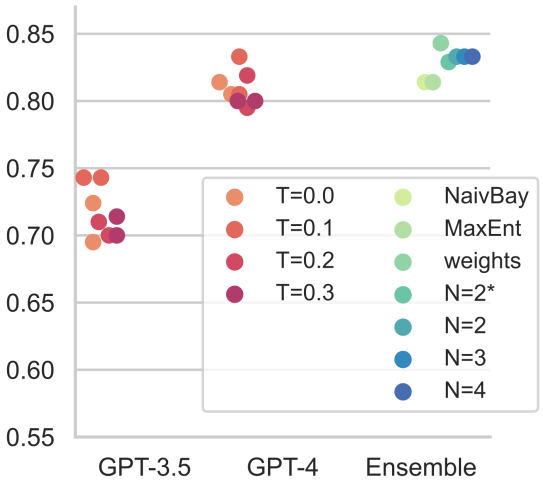

Figure 3: Ensemble models (in blue-green) created from a selection of single models (in red), based on GPT-3.5 and GPT-4 only ($T < 0.4$). The best ensemble model (Weighted Sum) obtains an accuracy of 0.843. Second best are the asymmetric decision trees $N = 2, 3, 4$ at 0.833, which is the same accuracy as for the best single model.

Nonetheless, it appears that the models perform on a continuum, where grammatical correctness, fluency and coherence go hand in hand.

In general, the Claude models most faithfully reproduce the original texts. However, this comes at the expense of not correcting all grammatical errors or resolving contradictions. The GPT models produce higher-quality output, but these models also reformulate the texts to a higher extent. Very few typos or grammar errors remain in their output. The GPT models may have a tendency to "overcorrect" for fluency, but whether that is considered good or bad is subjective.

The best ensemble model fluently combines sentences from GPT-3.5 and GPT-4 output, but often fails to replace the trickiest parts that go wrong in the best single model with sentences that some less reliable single model actually got right.

A full example of a text that is corrected by each model type (Claude, GPT-3, GPT-4 and Ensemble) is shown in Appendix D.

8 Discussion and Conclusion

Finnish is a morphologically rich language that is considered hard to learn. This study has shown the capacity of state-of-the-art large language models to produce accurate correction hypotheses for challenging learner texts. Experiments could have been conducted on simpler, established data sets in other languages, but that would not have served the purpose. However, the lack of appropriate annotated data sets meant that a low-resource scenario was adopted, with a data set consisting of 210 sentences. As the output of every run had to be tagged manually and there were 36 runs, the number of sentences to tag was still rather high.

The annotation was performed using a binary scheme: Either a sentence was considered fully correct or incorrect. This obscures any differences between "almost correct" and "totally wrong". Whereas this may seem too coarse an analysis on the level of individual sentences, it is unlikely to make a large difference for the data set as a whole and the performance ranking of the models.

A verified gold-standard would allow for automatic, faster testing. There are typically multiple correct answers, however, and it is hardly possible to know all possible alternatives in advance.

The benefit of the ensemble models turned out to be limited. Alternative directions for improvement might involve few-shot chain-of-thought prompting and finetuning (Kwon et al., 2023; Fan et al., 2023).

9 Limitations

The present study is exploratory and the size of the data set is small (25 learner texts consisting of 210 sentences in total). This means that very fine-grained conclusions cannot be made, since some observed differences are not statistically significant. Nevertheless, the higher-level distinctions are statistical significant, such as the difference in performance between the different types of LLMs. Additionally, all individual test results are plotted as "swarms" in order to clearly visualize the magnitude of the variance between different setups.

A larger data set would have been preferred, but this would also have required a heavier annotation effort. The annotation could also have been performed differently. Initially, the two independent annotators were in agreement on the category of approximately 5/6 of the sentences. A joint decision then needed to be made for the remaining 1/6. This was a pragmatic decision suitable for an exploratory feasibility study. If the goal had been to create a solid gold-standard reference for wide public dissemination, more rigorous and time-consuming approaches could have been considered.

Some prompt engineering was performed qualitatively, but no systematic quantitative evaluation of the effect of changing the prompts was per-

formed (see Appendix A).

A new version of Claude, Claude 2.0, has been published after the experiments were run. New experiments were not performed using Claude 2.0.

In this work, sentence accuracy is used as the evaluation metric. Analyzing the precision and recall of the corrections of specific error types is beyond the scope of this study. The aim is to look at the end result as a whole and investigate to what extent challenging learner texts can be reformulated into natural, correct, idiomatic language.

10 Ethical Considerations

The data set used in this study is a subset of the International Corpus of Learning Finnish (ICLFI). The corpus has been curated from authentic texts written by students of the Finnish language at international universities. The identities of the authors have nonetheless been protected. Names of people and places have been anonymized in the texts.

Large language models are trained on very large amounts of text data and may therefore learn harmful biases and prejudices that are reflected in some portions of the training data. Such tendencies have not been observed in the texts generated by the LLMs in this work.

Acknowledgments

I would like to express my sincere thanks and great appreciation to my colleague Mikko Aulamo for volunteering as an annotator of the data. I would also like to thank my colleagues Teemu Vahtola, Anssi Moisio and Jörg Tiedemann for valuable discussions and comments during the work on this paper. Likewise, I am very grateful to the reviewers of the manuscript for their insightful comments. This work has been supported by the *Behind the Words* project, funded by the Research Council of Finland 2021–2023.

References

Steven Bird, Ewan Klein, and Edward Loper. 2019. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*.

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, 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*, pages 1–59.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of GPT-3.5 and GPT-4 in Grammatical Error Correction.

Mathias Creutz and Eetu Sjöblom. 2019. Toward automatic improvement of language produced by non-native language learners. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 20–30, Turku, Finland. LiU Electronic Press.

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.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2023. GrammarGPT: Exploring open-source llms for native Chinese grammatical error correction with supervised fine-tuning.

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.

Xingyu Fu, Sheng Zhang, Gukyeong Kwon, Pramuditha Perera, Henghui Zhu, Yuhao Zhang, Alexander Hanbo Li, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, Dan Roth, and Bing Xiang. 2023. Generate then select: Open-ended visual question answering guided by world knowledge. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2333–2346, Toronto, Canada. Association for Computational Linguistics.

José Antonio García-Díaz, Camilo Caparros-laiz, Ángela Almela, Gema Alcaráz-Mármol, María José Marín-Pérez, and Rafael Valencia-García. 2023. UMUTeam at SemEval-2023 task 12: Ensemble learning of LLMs applied to sentiment analysis for low-resource African languages. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 285–292, Toronto, Canada. Association for Computational Linguistics.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2018. Near human-level performance in grammatical error correction with hybrid machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 284–290, New Orleans, Louisiana. Association for Computational Linguistics.

Jarmo Jantunen. 2011. Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. *Lähivõrdlusi. Lähivertailuja*, 21:86–105.

Jarmo Jantunen, Sisko Brunni, and University of Oulu, Department of Finnish Language. 2013. International Corpus of Learner Finnish.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Sang Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Beyond English: Evaluating LLMs for Arabic grammatical error correction. In *Proceedings of ArabicNLP 2023*, pages 101–119, Singapore (Hybrid). Association for Computational Linguistics.

Ruobing Li, Chuan Wang, Yefei Zha, Yonghong Yu, Shiman Guo, Qiang Wang, Yang Liu, and Hui Lin. 2019. The LAIX systems in the BEA-2019 GEC shared task. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–167, Florence, Italy. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Yassir Fathullah, Adian Liusie, Vyas Raina, Vatsal Raina, and Mark Gales. 2023. CUED at ProbSum 2023: Hierarchical ensemble of summarization models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 516–523, Toronto, Canada. Association for Computational Linguistics.

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

OpenAI. 2023. GPT-4 technical report.

Maria Carolina Penteado and Fábio Perez. 2023. Evaluating GPT-3.5 and GPT-4 on grammatical error correction for Brazilian Portuguese.

Gwenyth Portillo Wightman, Alexandra Delucia, and Mark Dredze. 2023. Strength in numbers: Estimating confidence of large language models by prompt agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 326–362, Toronto, Canada. Association for Computational Linguistics.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4:169–182.

Eetu Sjöblom, Mathias Creutz, and Teemu Vahtola. 2021. Grammatical error generation based on translated fragments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 398–403, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Chenming Tang, Xiuyu Wu, and Yunfang Wu. 2023. Are pre-trained language models useful for model ensemble in Chinese grammatical error correction? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 893–901, Toronto, Canada. Association for Computational Linguistics.

University of Jyväskylä. 2016. The Finnish Subcorpus of Topling - Paths in Second Language Acquisition.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark.

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.

Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. 2023. CLEME: Debiasing multi-reference evaluation for grammatical error correction.

Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2023. Selecting better samples from pre-trained LLMs: A case study on question generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12952–12965, Toronto, Canada. Association for Computational Linguistics.

Appendices

A Prompts

The following zero-shot prompt, written in Finnish, was utilized to ask GPT-3.5 and GPT-4 to produce corrected texts:

`Hei! Korjaisitko seuraavan tekstin siten, että siitä tulee sujuvaa, erinomaista suomen kieltä eikä sisällä asiavirheitä eikä kielioppivirheitä. Älä kirjoita ylimääräistä tekstiä. Pelkkä korjattu teksti riittää. Tekstin alku:\n <LEARNER TEXT GOES HERE>\n Teksti päättyy.`

In English the prompt reads: *Hi, could you please correct the following text in such a way that it becomes fluent, impeccable Finnish language and does not contain factual errors or grammar errors. Do not write superfluous text. Just the corrected text is enough. Start of the text:*\n <LEARNER TEXT GOES HERE> \n *Text ends.*

The same prompt was basically used for the Claude LLM as well, with the exception that Claude requires the use of the keywords "Human:" and "Assistant:" to mark the roles in the dialog:

`\n\nHuman: Hei! Korjaisitko seuraavan tekstin siten, että siitä tulee sujuvaa, erinomaista suomen kieltä eikä sisällä asiavirheitä eikä kielioppivirheitä. Älä kirjoita ylimääräistä tekstiä. Pelkkä korjattu teksti riittää.\n <LEARNER TEXT GOES HERE>\n\nAssistant:`

Some exploratory prompt engineering went into the design of the final prompt, but no quantitative evaluation was made. Specifically, it was observed that the LLMs tended to embed their answers in polite phrases to create the impression of a natural dialog. Therefore the prompt was modified to explicitly state that only the actual correction hypothesis was desired in the output.

B Random Fluctuation

For every learner text, 36 versions of corrected texts were obtained. Three LLMs were used with six temperature values each, and every such configuration was run twice. That is, every prompt was submitted twice to the same LLM with the same temperature.

As the LLMs are non-deterministic by nature, results are expected to be slightly different on every run. However, there should not be a systematic difference, such that better (or worse) results are consistently obtained the first (or second) time the same configuration is used. The accuracies produced by all single models are plotted in Figure 4, organized by runs (first or second).

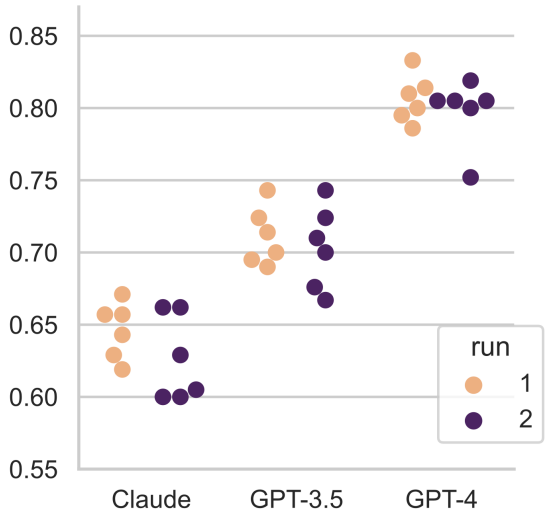

Figure 4: Accuracies obtained for all the single models. The data points are exactly the same as in Figure 1, but they have been grouped into "swarms" differently. Rather than using temperature as the categorizing feature, we now study whether the result was produced by running the configuration for the first or the second time. Thus, for every LLM, there are six dots in light color from running the prompts with six different temperatures for the first time, and six dots in dark color, from running the same setup again. If there is no systematic ordering effect, the averages from both runs should be approximately the same.

Statistical significance tests reject the hypothesis that the models are effected by the order of the runs. That is, the Claude, GPT-3.5 and GPT-4 models behave as expected in this respect.

C Further Analysis of Ensemble Models

Ensemble models based on all 36 single models were created. The accuracies obtained by the ensemble models are shown in Figure 5 together with the results from the individual single models. As discussed in Section 6.2, this is not the best possible result. A slightly better ensemble is obtained by using the Weighted Sum model and excluding all the Claude models and any models with temperatures above 0.3.

Claude + GPT-3.5? Inspired by the results from combining GPT-3.5 with GPT-4, can we benefit from combining GPT-3.5 with Claude as well? If,

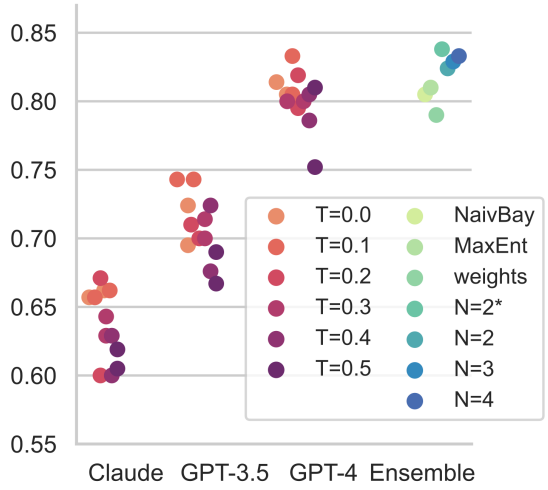

Figure 5: The single models (from Figure 1; in red) plotted together with the ensemble models (in blue-green). The best performing ensemble model is the asymmetric decision tree variant called $N = 2^*$, which attains an accuracy of 0.838. The model $N = 4$ performs on par with the best single model (accuracy 0.833), but the remaining ensemble models perform worse than the best single model.

for some reason, the best available LLM is not available, can this be compensated by using an ensemble of weaker LLMs? Unfortunately, this does not seem possible. The highest accuracy observed for an ensemble of GPT-3.5 and Claude models is 0.748. It is no better than an ensemble of GPT-3.5 models alone (accuracy: 0.752), and this setup outperforms none of the twelve single GPT-4 models.

The Naive Bayes and Maximum Entropy classifiers did not outperform the single models in the experiments. Possibly, the training sets were insufficient, or these classifiers simply failed to capture the correlations between features accurately. The Naive Bayes classifier did, however, prove useful as the fallback model in the decision-tree approach.

Further tests involved "standard", symmetric decision trees, using information gain as a splitting criterion for features. Their learning ability was poor on this task.

D Example Corrections

The differences between the different LLMs are illustrated in Table 2 using an example text. Models at temperature 0.1 have been selected as they are generally the strongest performing single models. Also the best ensemble model is included.

The text is challenging. In addition to spelling and grammar errors, it contains a contradiction. The Claude model most faithfully reproduces the original text, leaving some grammatical errors and a contradiction in the text.

The GPT models reformulate the text to a higher extent. No typos or grammar errors remain. However, these models are not able to resolve all factual errors. GPT-4 is more successful than GPT-3.5 at this, by simply dropping a part of a sentence that it cannot make sense of.

The ensemble model combines sentences from GPT-3.5 and GPT-4 without arriving at a fully correct text in this case.

These examples also highlight some challenges related to the annotation. Regarding the house being repaired when the laundry machine is broken, the original text is not clear. Apparently something needs to be fixed in the house because of the broken laundry machine, and therefore that sentence was annotated as correct. If Jutta thinks that Anssi is a fun guy, does that mean that they have fun together? The annotators found that to be the case. However, Anna is not jealous because she *herself* finds Anssi a fun guy, but because Jutta does. (Anna apparently also likes Anssi, but not necessarily because of him being fun.)

Original

Minä lulee etttä, Anna on nyt niin erilainen kuin tavallisesti, koska hänellä on stressi. Anna ei ole aikaa puhumaan Jutan kanssa, koska korjata tule hänen kotiinsa. Annalla *ei ole siihen jokin hyvä syy*, koska pesukone on rikki, pesukone on siihen jokin hyvä syy. Minusta Anna on kateellinen, koska Juttasta Anssi on hauska mies.	I belives thatt, Anna is now so different than usually, because she is stressed. Anna is no time talking with Jutta, because repair come to her house. Anna *has not some good reason for this*, because the laundry machine is broken, the laundry machine is a good reason for that. I think Anna is jealous, because according Jutta Anssi is a fun guy.

Claude ($T = 0.1$)

Minä luulen, että Anna on nyt niin erilainen kuin tavallisesti, koska hänellä on stressiä. Anna ei ole aikaa puhua Jutan kanssa, koska korjaaja tulee hänen kotiinsa. Annalla *ei ole siihen jotain hyvää syytä*, koska pesukone on rikki. Pesukone on siihen jokin hyvä syy. Minusta Anna on kateellinen, koska Juttasta Anssi on hauska mies.	I think that Anna is now so different than usual, because she is stressed. Anna is no time to talk to Jutta, because a repairman is coming to her house. Anna *does not have some good reason for this,* because the laundry machine is broken. The laundry machine is a good reason for that. I think Anna is jealous, because according Jutta Anssi is a fun guy.

GPT-3.5 ($T = 0.1$)

Minä luulen, että Anna on nyt niin erilainen kuin tavallisesti, koska hänellä on stressiä. Annalla ei ole aikaa puhua Jutan kanssa, koska hänellä on korjaaja tulossa kotiinsa. Annalla *ei ole siihen mitään hyvää syytä*, koska pesukone on rikki. *Pesukoneella on siihen jokin hyvä syy.* Minusta Anna on kateellinen, koska Jutan ja Anssin välillä on hauskaa.	I think that Anna is now so different than usual, because she is stressed. Anna doesn't have time to talk to Jutta, because she will have a repairman come to her house. Anna *has no good reason for this*, because the laundry machine is broken. *The laundry machine has a good reason for that.* I think Anna is jealous, because Jutta and Anssi have fun together.

GPT-4 ($T = 0.1$)

Minusta tuntuu, että Anna on nyt erilainen kuin tavallisesti, koska hän on stressaantunut. Anna ei ehdi jutella Jutan kanssa, koska hänen kotiaan ollaan korjaamassa. Annalla on hyvä syy tähän, sillä hänen pesukoneensa on rikki. Minusta Anna vaikuttaa kateelliselta, koska *hänestä* Anssi on hauska mies.	I think that Anna is now different than usual, because she is feeling stressed. Anna doesn't have time to talk to Jutta, because her house is being repaired. There is a good reason for this, because her laundry machine is broken. To me Anna seems jealous, because *she* thinks Anssi is a fun guy.

Ensemble (Weighted Sum policy based on GPT-3.5 and GPT-4 models only with $T < 0.4$)

Minä luulen, että Anna on nyt niin erilainen kuin tavallisesti, koska hänellä on stressiä. Anna ei ehdi puhua Jutan kanssa, koska hänen kotiaan korjataan. Annalla *ei ole siihen mitään hyvää syytä*, koska pesukone on rikki. *Pesukoneella on siihen jokin hyvä syy.* Minusta Anna on kateellinen, koska Jutan ja Anssin välillä on hauskaa.	I think that Anna is now different than usual, because she is stressed. Anna doesn't have time to talk to Jutta, because her house is being repaired. Anna *has no good reason for this*, because the laundry machine is broken. *The laundry machine has a good reason for that.* I think Anna is jealous, because Jutta and Anssi have fun together.

Table 2: A learner text (from Table 1) with corrections suggested by a Claude, GPT-3.5, and GPT-4 model as well as an ensemble model. The Finnish text on the left is accompanied by an approximate English translation on the right. Spelling mistakes and grammatical errors have been underlined. Factual errors, such as contradictions and incorrect coreference are rendered in italics.