Comparing and Combining Claude, GPT-3.5 and GPT-4 Large Language Models in the Correction of Finnish Learner Texts

Anonymous ACL submission

Abstract

This paper studies grammatical error correction on challenging authentic Finnish learner texts at CEFR A1 level. 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, various ensemble models combining outputs of multiple single models are evaluated. The best results are obtained by explicitly modeling agreement between single models as a chain of rules in an asymmetric decision tree. The best performing ensemble model obtains an accuracy of 85.7 %, whereas the best single model, which is a GPT-4 model, reaches an accuracy of 82.4 % fully correct sentences. In other words, the ensemble model reduces the sentence error rate by 18.8 % in comparison to the best single model.

1 Introduction

Grammatical Error Correction (GEC) is the task of automatically detecting and correcting errors in text. The term *grammatical* is understood broadly. The errors may be grammatical, such as missing prepositions and mismatched subject-verb agreement, but also orthographic and semantic, such as misspellings and word choice errors, respectively (Bryant et al., 2023). However, GEC is typically seen as a *local* substitution task (Ye et al., 2023), where a few occasional mistakes are corrected in generally intelligible text.

Our aim is to help second-language (L2) learners express themselves fluently and idiomatically in a non-native language that they do not master very well. We work with challenging learner texts that contain numerous mistakes when it comes to inflection, spelling, word choice, word order and even low intelligibility overall. We have previously employed neural machine translation with different data augmentation techniques to solve this task (*self-citations omitted*). Recent developments and the advent of powerful large language models (LLMs) have provided us with new approaches to tackling the problem.

The 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. Additionally, we investigate to what extent an ensemble of multiple models can outperform single models in this task.

2 Data

As data for our experiments we use a subset of ICLFI, the International Corpus of Learner Finnish (Jantunen, 2011; Jantunen et al., 2013). The corpus consists of texts written by students of Finnish as a foreign language from various language backgrounds. It has been compiled with the help of Finnish language teachers around the world. ICLFI is available online through the Language Bank of Finland.[1]

For our study we randomly selected 25 texts that the corpus creators have labeled with the lowest language proficiency level: A1 in the Common European Framework of Reference for Languages (CEFR).[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 et al., 2013) contain reference corrections that can be utilized for evaluation, but that is unfortunately not the case with the ICLFI corpus.[3]

[1] https://www.kielipankki.fi/corpora/iclfi/

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

[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

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.

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] The LLMs were accessed through their APIs (application programming interfaces), 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.

The LLMs are non-deterministic by default. There is a so-called temperature parameter ranging between 0 and 1 that regulates the randomness of the output. A low temperature is expected to produce the most probable and predictable result, whereas higher temperatures increase creativity.[6]

We have tested each of the LLMs on six different temperature values: 0.0, 0.1, ..., 0.5. Every configuration was run twice, because of the non-deterministic nature of the task. Even with the lowest temperature of 0.0, the systems were not fully deterministic, and some variability remained in the output. This left us with 36 correction hypotheses for each of the 25 texts (3 LLMs times 6 temperature values times 2 runs each).

In the following, we will refer to these 36 setups as our *models* or *single models*. Naturally, models may agree amongst each other and produce the same hypotheses, so the total number of unique hypotheses is typically lower than 36.

4 Single Model Results

The 36 correction hypotheses produced by the LLMs for each of the 25 learner texts were tagged as correct or incorrect by the authors of the paper. The tagging was performed on sentence level: either a sentence was fully correct or it was incorrect, considering the context of surrounding sentences. Table 2 shows one proposed correction of a text accompanied by an English translation and illustrates some challenges related to the annotation.

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. Virtually, all GPT-4 models are better than all GPT-3.5 models, which are in turn better than all Claude models. Secondly, the temperature parameter works as expected. Conservative, predictive 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 $T = 0.0$.

5 Ensemble Models

The best single model produces 173 correct sentences out of 210 (82.4 %). However, if look at all 36 models combined, there are only 6 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 $204/210$ (97.1 %). In the following, we will study supervised learning of ensemble models that combine outputs from the single models.

Our approach makes the simplifying assumption that sentences from different hypotheses can always be combined. For instance, in a fictive scenario, where Model 1 proposes the partly correct text *"Hi there! How's you?"* and Model 2 proposes the partly correct text *"Helo! How are you?"*, it would be possible to concatenate the correct parts

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

Minusta tuntuu, että Anna on nyt erilainen kuin yleensä, koska hän on stressaantunut. Annalla ei ole aikaa jutella Jutan kanssa, sillä hänen kotiaan ollaan korjaamassa. Tähän on hyvä syy: pesukone on rikki. Pesukoneen rikkoutuminen on siis hyvä syy. Minusta tuntuu, että Anna on kateellinen, koska hänestä Anssi on hauska mies.	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. There is a good reason for this: the laundry machine is broken. The broken laundry machine is indeed a good reason. I think Anna is jealous, because she thinks Anssi is a fun guy.

Table 2: The correction of the text in Table 1 as proposed by one of the models (GPT-3.5, $T = 0.5$, 1st run). Even though there may be other more likely interpretations, all but the last sentence were annotated as correct. The last sentence was considered incorrect, because the original text explicitly states that Anna is jealous that not herself, but *Jutta* likes Anssi. 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.

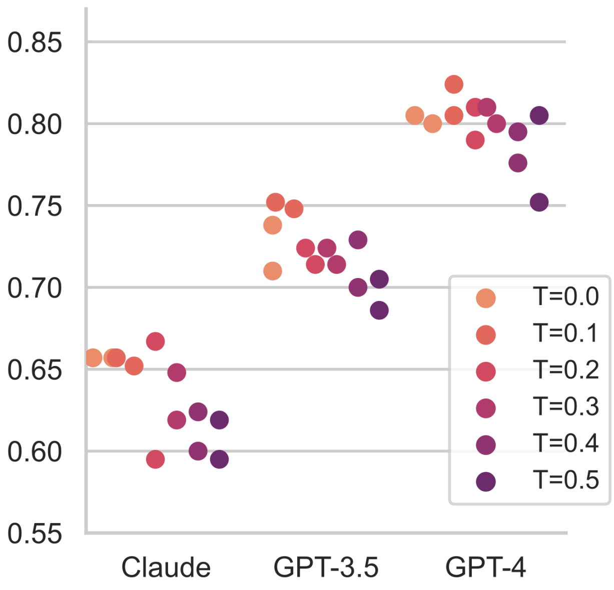

Figure 1: Accuracies of each of the 36 single models. Every model is represented by a dot, and the dots are grouped in "swarms" by LLM type. In every swarm, we progress from left to right as the temperature (T) rises, with higher temperatures rendered in darker color. The best model (GPT-4, $T = 0.1$, 1st run) reaches an accuracy of 0.824, which corresponds to 173 fully correct sentences out of 210 in the data.

from each hypothesis to produce a coherent, correct new text: *"Hi there! How are you?"*[7]

We formulate the problem as a classification task. For every original sentence in the input, each of the 36 models has produced a correction hypothesis, which has been labeled as either correct or incorrect by the annotator. Typically the number of unique hypotheses is lower than 36, because several models produce the same hypotheses. This information can be exploited to train a classifier that predicts when a hypothesis is correct based on the subset of models that have proposed it, as illustrated in Figure 2. During training, the classifier will hopefully learn which models are more reliable than others and discover useful patterns of agreement between models. When the classifier is used for prediction, we proceed sentence by sentence and choose the proposed correction that the classifier assigns the highest likelihood of being correct.

As there is a very limited amount of data available, we do not set aside a separate test set. Instead, we use cross-validation such that every learner text in turn serves as the test set and the remaining 24 texts are used for training. In this way, we obtain test results for all 25 texts and can study how our ensemble models perform in comparison to the single models.

The classifiers that we study are described in the following sections. Due to the limited amount of data, we need to restrict ourselves to fairly simple

[7] Alternatively, we could work on full texts without dividing them into sentences. However, this would be a very crude measure, as we only have 25 texts in total. Additionally, the lengths of the texts vary considerably (between 1 and 15 sentences), and the shorter texts are more likely to be successfully corrected. As we are interested in more fine-grained analysis, where units of similar size are compared, we decided to work on sentence level instead.

Hypotheses	Proposed by models											Label
	1	2	3	4	5	6	7	8	9	...	36	
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 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.

classifiers with a small numbers of parameters to tune, in order to avoid overfitting to the training set.

5.1 Naive Bayes

The first classifier we test is Naive Bayes, using the implementation of the NLTK library (Bird et al., 2019). The 36 individual models being "on" or "off" serve as features as shown in Figure 2. The training of the classifier amounts to solving a closed-form expression, which means that the classifier is not too sensitive to the size of the data set. However, the underlying independence assumption is a simplification that may lead to the exaggeration of the effect of correlated features.

5.2 Maximum Entropy

We also test logistic regression using the Maximum Entropy classifier of NLTK. This classifier does not assume conditional independence, but since it does not have a closed-form solution, it may end up learning a suboptimal set of weights.

5.3 Weighted Sum

As we are not sure whether the Maximum Entropy classifier converges to an optimal solution on our limited data set, we decided to try a simplified, deterministic approach as well. We estimate a weight vector w of the same dimensionality as our binary correction hypothesis vectors x. During prediction, when correction hypotheses are compared, the one 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. The value of w_i is estimated from the training set. Each time Model i proposes a hypothesis that is correct, w_i is increased by $1/n$, where n is the number of models that propose the same hypothesis. That is, if a model is the only one proposing a correct hypothesis, then it will get the full "point", but if the same hypothesis was also proposed by nine other models, then all these models will get $1/10$ of a "point" each. This mitigates the effect of correlated features.

5.4 N Agreeing Models

To explicitly model correlated features we studied another type of classifier, an asymmetric decision tree that branches onto one side only ("`if` *condition 1* `then done else if` *condition 2* `then done else if` *condition 3* `then done ... else done`").

The last fallback condition (last `else` clause) corresponds to using the best single model on the data. However, which single model is the best is determined from the training set, so it is not guaranteed to also be the most accurate single model on the test set.

The preceding conditions in the `if - else` chain correspond to all combinations of $2 \mathinner{..} N$ 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.

For instance, imagine that the best single model is Model 30, and it suggests a hypothesis that is correct for 80 % of the sentences in the training set. It also turns out that in all cases where Models 17 and 31 agree on a hypothesis, that hypothesis is correct in 89 % of the cases. And if Models 17, 25 and 26 agree on a hypothesis, then that hypothesis is correct for 95 % of all such occurrences in the training set. These conditions are sorted, highest accuracy first, such that the classifier first checks if the three models 17, 25 and 26 agree, in which case their proposed hypothesis is chosen. If not, the pair of models 17 and 31 is examined next and if they agree, their hypothesis is selected. Only if none of these conditions are fulfilled, the hypothesis proposed by Model 30 is used.

We have tested N values ranging from 2 to 5, that is, pairs, triples, quadruples and quintuples of models. For higher values of N, all lower-order combinations of models are also included. We will not report results for quintuples ($N = 5$), as their results are identical to those of the quadruples ($N = 4$).

For the pairs of models ($N = 2$), we have additionally tested a minor variant ($N = 2^*$), in which the conditions in the `if - else` chain are ordered differently. In the $N = 2$ classifier, the accuracies of the model pairs are calculated on the full training set "statically" and ordered accordingly. In the training of the $N = 2^*$ classifier, the most accurate model pair is put as the first condition, but after this the accuracies of all other pairs are recalculated "dynamically" on the remainder of the training set, from which the data points that triggered the first condition have been removed. This is repeated at every step until we reach the final fallback single model.[8]

A further tested variant consists in replacing the fallback single model with the Naive Bayes classifier (Section 5.1). The results that we report are in fact based on this variant, since it produced slightly higher accuracies.

6 Ensemble Model Results

The accuracies obtained by the ensemble models are shown in Figure 3 together with the results from the individual single models. Much to our delight, we observe that some of the ensemble models ($N = 2, 3, 4$) do outperform the best single model, whereas the others ($N = 2^*$, Naive Bayes, Maximum Entropy and Weighted Sum) do not. However, the best ensemble model is not radically better than the best single model. Can we do better?

We have observed that the Claude models perform worst in the task and that low temperatures are to be preferred. In our next experiment we reduce the set of single models that are included in the ensemble model. That is, we leave out the Claude models and temperatures above 0.3. The results are shown in Figure 4. Now, the advantage between the best ensemble model ($N = 2^*$) and best single model grows (0.857 vs. 0.824). In other words, the sentence error rate is reduced by 18.8 %, which does make a difference. This difference is statistically significant at the 90 % confidence level, which is decent considering the limited size of the data set.

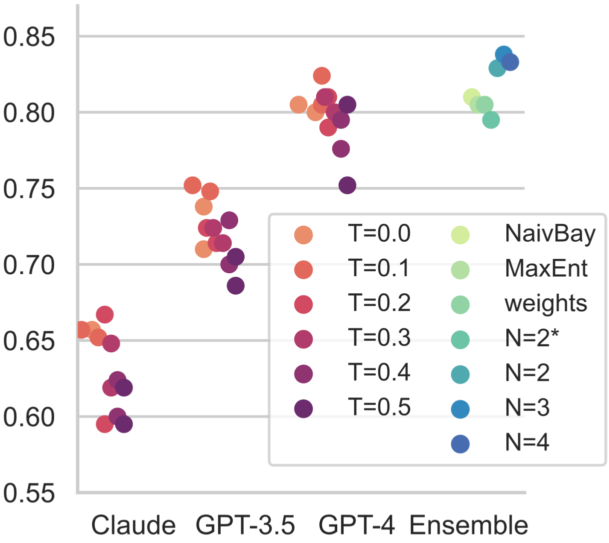

Figure 3: The single models (from Figure 1; in red) plotted together with the ensemble models (in blue-green). The best performing ensemble model is $N = 3$ with an accuracy of 0.838. Also the models $N = 4$ and $N = 2$ outperform the best single model by a slight margin (accuracies 0.833 and 0.829 respectively, compared to the best single model at 0.824).

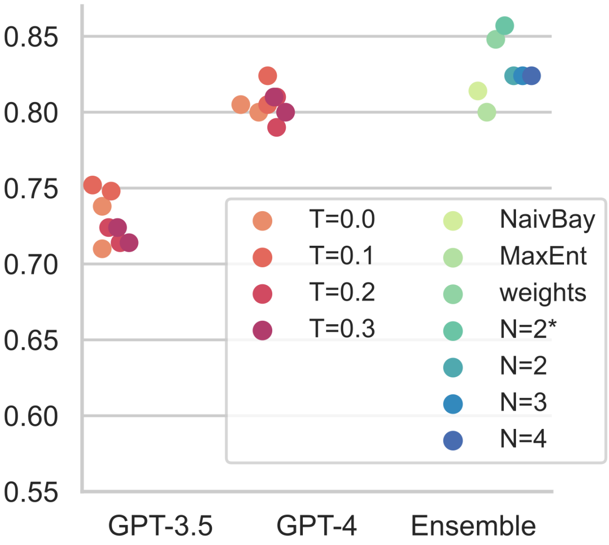

Figure 4: Ensemble models (in blue-green) created from a smaller set of single models (in red), based on GPT-3.5 and GPT-4 only ($T < 0.4$). The best ensemble model $N = 2^*$ obtains an accuracy of 0.857. Second best is the weighted sum at 0.848.

[8] We did not test a similar approach on higher values of N than 2. Therefore, we do not have classifiers called $N = 3^*, 4^*, 5^*$.

6.1 What Rules Are Learned?

The $N = 2^*$ model with the accuracy of 0.857 is the best result we have obtained in our trials involving different combinations of LLMs and temperature values. Therefore it is interesting to see what kind of rules are learned. By *rules* we understand the conditions put in the `if - else` chain, which here correspond to pairs of agreeing single models.

We will, however, need to study 25 separate rule sequences, since we run 25 tests, each time with a different test set and a slightly different training set. Luckily, the differences between the runs are very small. Consistently, the top half of the rules combine a GPT-3.5 and a GPT-4 model, whereas the rest of the rules combine single models of the same type (GPT-3.5 with GPT-3.5 and GPT-4 with GPT-4). This suggests that the most reliable, accurate signal is obtained when different LLM types are combined. In other words, two different LLMs (GPT-3.5 and GPT-4) complement each other better than repeated runs with the same LLM, at the same or different temperatures.

6.2 GPT-3.5 + Claude?

Encouraged by the results from combining GPT-3.5 with GPT-4, can we benefit from combining GPT-3.5 with Claude as well? If, for some reason, we do not have access to the best available LLM (GPT-4), can we compensate by using an ensemble of weaker LLMs (GPT-3.5 and Claude)?

Unfortunately, this does not seem possible. The highest accuracy we have observed for an ensemble of GPT-3.5 and Claude models is 0.762. It is no better than an ensemble of GPT-3.5 models alone, which reaches the same accuracy. This accuracy is indeed better than that of any single Claude or GPT-3.5 model (best Claude: 0.667, best GPT-3.5: 0.752). Compared to the single GPT-4 models, however, it outperforms only one out of twelve and is far below the best single GPT-4 model at 0.824.

6.3 Other Lessons Learned

The Naive Bayes and Maximum Entropy classifiers did not outperform the single models in our 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 our N-agreeing-models decision-tree approach.

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

We also tested pruning of overlapping patterns in our N-agreeing-models implementation, but this had no effect on the results. Overlapping patterns emerge, for instance, if a quadruple of single models contains a triple of single models as its subset and this triple is as accurate as the quadruple. Then the quadruple is superfluous. Alternatively, if the triple never occurs in other contexts than the quadruple, then the triple can be considered superfluous.

The size of the training set appears to affect which models perform the best. With the full data set available (Figure 3), higher values of N are at the top, and the Weighted Sum model is the worst (of the reported ones). When part of the data points are dropped (Figure 4), the accuracy does not increase with higher values of N. $N = 2^*$ is now at the top and the Weighted Sum is second best.

7 Related Work

Bryant et al. (2023) have compiled an overview of the state of art in grammatical error correction. This survey covers data sets (predominantly in English) as well as approaches commonly used to solve the task, most importantly: classifiers, statistical machine translation, neural machine translation, edit-based approaches and language models. Unfortunately, the article was written before the breakthrough of GPT-3.5 and GPT-4, and observations regarding LLMs are therefore limited. The survey mentions small-scale experiments (Wu et al., 2023; Coyne et al., 2023), which generally conclude that LLMs have a tendency to overcorrect for fluency, which causes them to underperform on datasets that were developed for minimal corrections (Fang et al., 2023). This raises the question whether the standard test sets for (English) GEC are good benchmarks or whether more challenging sets should be devised for the evaluation of more advanced error correction.

7.1 GPT Model Performance

Coyne et al. (2023) study English GEC using GPT-3.5 and GPT-4 on the BEA-2019 shared task data set (Bryant et al., 2019) and JFLEG (Napoles et al., 2017). The authors work on sentences in isolation without context. Their study focuses on prompt engineering and includes both automatic and human evaluation. In line with our results they conclude that the tested models demonstrate strong performance and that a low temperature is consistently associated with better performance in this task. GPT-4 performs slightly better than GPT-3.5. On the JFLEG set, GPT-4 produces the highest score yet reported.

Fang et al. (2023) perform correction of not only sentences in isolation, but also of documents, as we do. They also extend their study to German and Chinese data sets. They use ChatGPT as their LLM, which corresponds most closely to the GPT-3.5 (turbo) version that we have used. They find that the sentences corrected by ChatGPT exhibit a high level of fluency and naturalness, but the system "performs poorly on most error types, such as agreement, coreference, tense errors across sentences, and cross-sentence boundary errors." We believe that GPT-4 would have done a better job at fixing this type of errors.

Penteado and Perez (2023) compare GPT-3.5 and GPT-4 against the spelling and grammar error correction features in Google Docs and Microsoft Word for Brazilian Portuguese. In line with the other studies, they observe that LLMs prioritize fluency and coherence over grammatical accuracy, leading to unnecessary changes to the text, increasing false positives. Therefore, higher precision is obtained by rule-based methods that have a narrower focus on grammatical accuracy and make changes only when necessary. However, GPT-3.5 and GPT-4 clearly outperform Microsoft Word and Google Docs on the more challenging texts that had been typed fast or contain slang, abbreviations, and neologisms.

7.2 Claude

We have not found work on GEC, where the Claude LLM would have been assessed. Lin and Chen (2023) evaluate open-domain conversations with large language models. They assess performance on four so-called "dimensions": appropriateness, content, grammar, and relevance. They test Claude (v1.3) and ChatGPT, which are optimized for chat applications, as well as GPT-3.5, which is not. When comparing the Claude and ChatGPT models, both models demonstrate competitive performance across different evaluation dimensions, with Claude slightly outperforming ChatGPT in certain configurations.

Several blog posts compare LLMs. The applications of interest vary and the rigorousness of the analyses can be questioned. Garst (2023) compares the latest version of Claude (v2) to GPT-4 and thinks that Claude 2 shines in key areas, but GPT-4 still leads in general performance: "For natural language processing broadly, GPT-4 remains state-of-the-art. Its sheer model scale and training on a massive internet corpus make it hard to match for conversing, writing, and answering open-ended questions." This is in line with our own observations, although we have used an earlier version of Claude that is allegedly not as strong as Claude 2.

7.3 Ensemble Models

Ensemble models have proven effective in GEC tasks. Unlike our approach, where we select one sentence from a number of proposed sentences, many systems compare individual proposed corrections (edits) to an annotated reference, such as changing the inflection of a word (for instance, *played* to *playing*). Given the reference, it is possible to estimate the precision for specific error types of different single models, and the final hypothesis can combine edits from the different single models. This naturally requires the existence of annotated training data, which we do not have. Li et al. (2019) investigate classification models with bi-directional recurrent neural networks (Bi-RNN) and neural machine translation (NMT) models. Some rules are also involved. Their GEC systems ranked the first in the Unrestricted Track of the BEA-2019 GEC Shared Task (Bryant et al., 2019), and the third in both the Restricted Track and the Low Resource Track. Grundkiewicz and Junczys-Dowmunt (2018) test a variety of ensemble techniques, which combine statistical and neural machine translation as well as a spell-checking component.

In more recent work, Tang et al. (2023) study ensembles of pre-trained language models for Chinese (BART, BERT, GPT-2 etc). Sentence- and edit-level ensembles as well as voting techniques are tested, but the ensemble models do not outperform the best single models.

We have found only little work on ensemble models built on GPT-3.5 or GPT-4 and none of it addresses the GEC task. Jiang et al. (2023) propose an ensemble learning model for ranking and fusing the outputs from multiple LLMs in instruction-following tasks. However, they use ChatGPT as a reference for ranking other models, not as one of the models in the ensemble. Yuan et al. (2023)

utilize GPT-3 (Brown et al., 2020) to generate questions to given answers in given contexts. Different approaches to choosing the best generated question are tested. Fu et al. (2023) work on visual question answering, where pretrained language models first generate a set of possible answers, and a lightweight answer selection model is then trained to identify the correct answer from the set. Further works of interest that involve combinations of different LLMs include hierarchical ensembles of T5 models for summarization (Manakul et al., 2023), ensemble learning of mainly BERT-based LLMs for sentiment analysis of low-resource African languages (García-Díaz et al., 2023), and multiple-prompt agreement confidence scores in question-answering tasks (Portillo Wightman et al., 2023).

7.4 Finnish GEC

Finnish spell checking based on finite-state technology (Pirinen, 2014) as well as grammar checking based on constraint grammar (Karlsson, 1990) have a long history, but systematic research on Finnish grammatical error correction is very scarce because of the lack of annotated data sets. The ICLFI (Jantunen et al., 2013) and TopLing (University of Jyväskylä, 2016) corpora consist of authentic, anonymized learner texts, but there are no correction hypotheses available for model training or testing. There used to be an additional resource, the so-called YKI corpus based on Finnish national certificates of language proficiency exams, but it is unfortunately no longer available because of copyright issues.

An annotated sample of the (since then withdrawn) YKI corpus was used as test data in previous Finnish GEC studies (*self-citation omitted*). The full-sentence accuracy obtained for the best setup was 27.2 %, which falls far behind the figures reported in the current work. Even if conditions are relaxed slightly by allowing a few “minor errors”, the accuracy reaches only 44.5 %. Direct comparisons cannot be made because of the different corpora used in the studies. However, the types of texts and levels of the learners are very similar in both setups, and we can therefore say with a high level of certitude that the current GPT models are far better than our earlier models at producing corrections that are grammatical, fluent and natural in context.

8 Discussion and Conclusion

In general, we are happy with the accuracies obtained in our experiments, especially by the best GPT-4 and ensemble models. The theoretical upper bound on accuracy by an oracle model would be 97.1 %, and our best ensemble reached 85.7 %.

In line with related work, we were able to confirm that lower temperatures are to be preferred in text correction tasks. However, API access to the LLMs tested in this work has only been publicly available for a few months, and it is difficult to find relevant related publications. Some of the cited work needs to undergo peer review and is so far only available on `arxiv.org`.

The conduction of the experiments was made more difficult by the lack of gold-standard reference corrections of the texts. Our own annotation scheme was simple, as it sufficed to determine whether a sentence was fully correct or not. However, despite the rather small data set (210 sentences), each time a model was run on the data, the resulting correction hypotheses had to be annotated manually, and there were 36 such runs in total.

A set of verified gold-standard corrections would allow for automatic evaluation and speed up the testing of further models and configurations. As there are typically multiple correct answers, a multi-reference gold-standard would be ideal. A possible continuation of our work could be to produce such gold-standard sets nearly automatically. Highly accurate correction hypotheses would be generated by the best single and ensemble models. Humans would verify the correctness and post-edit, when necessary. There are other possible approaches as well, involving multiple rounds of prompting, where the LLMs are requested to refine their earlier outputs.

9 Limitations

The size of the data set in this study 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. We have also chosen to visualize all individual test results in plots, which helps to give a better sense of proportions.

A larger data set would have been preferred, but this would also have required a heavier annotation effort (see Section 8). In addition, we would have liked to run more tests with identical configurations (same LLM at same temperature). Since the LLMs are non-deterministic, results change from one run to another. We ran each unique configuration twice. To our understanding, the runs should be independent of each other, but we are unable to exclude the possibility of an ordering effect as far as the Claude models are concerned (see Appendix B). Additional runs could bring clarity into this potential issue.

Some prompt engineering was performed qualitatively, but no systematic quantitative evaluation of the effect of changing the prompts was performed (see Appendix A).

A new version of Claude, Claude 2.0, was published after we had generated correction hypotheses using Claude v1. We did not redo the experiments using Claude 2.0.

In this work, sentence accuracy is used as the evaluation metric. Analyzing the precision and recall of the corrections of individual errors or error types is beyond the scope of this study. In other words, our aim is not to assess how well different types of grammar errors or misspelled words are corrected. The aim is to look at the end result 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 and we have not had access to the original names.

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. We have not, however, observed any such tendencies in the texts generated by the LLMs in our experiments.

References

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

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.

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.

Kim Garst. 2023. Claude 2 vs GPT-4 in 2023: Comparing the top AI models.

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.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *Proceedings of 13th International Conference on Computational Linguistics*, volume 3, pages 168–173, Helsinki, Finland.

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.

Tommi Pirinen. 2014. *Weighted Finite-State Methods for Spell-Checking and Correction*. Ph.D. thesis, University of Helsinki.

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.

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 encourages 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 we did not evaluate the results quantitatively on a test set. Specifically, we 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

We obtained 36 versions of corrected texts for every learner text. 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, we expect results 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 5, organized by runs (first or second).

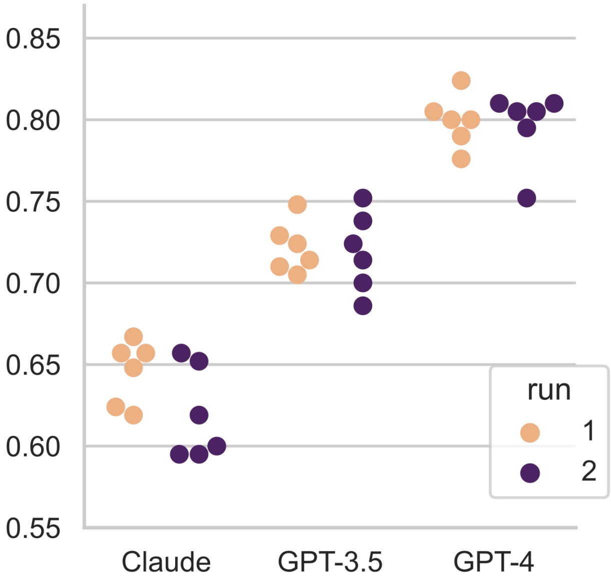

Figure 5: Accuracies obtained for all the single models. The data points are exactly the same as in Figure 1, but they have been grouped into "swarms" differently. Rather than using temperature as the categorizing feature, we now study whether the result was produced by running the configuration for 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 GPT models are effected by the order of the runs. Interestingly enough, this may not be true for Claude, where higher accuracies were obtained in the first run than in the second (p value of sign test: 0.0625). Further runs would be required to understand if this outcome is systematic or occurred by chance after all. Fortunately, this type of unexpected behavior was only observed for the model that was consistently the weakest one and had the least potential for solving the task.