# Towards Fine-Grained Pedagogical Control over English Grammar Complexity in Educational Text Generation

**Dominik Glandorf**
University of Tübingen
Yale University
dominik.glandorf@student.uni-tuebingen.de

**Detmar Meurers**
Leibniz-Institut für Wissensmedien (IWM)
detmar.meurers@uni-tuebingen.de

## Abstract

Teaching foreign languages and fostering language awareness in subject matter teaching requires a profound knowledge of grammar structures. Yet, while Large Language Models can act as tutors, it is unclear how effectively they can control grammar in generated text and adapt to learner needs. In this study, we investigate the ability of these models to exemplify pedagogically relevant grammar patterns, detect instances of grammar in a given text, and constrain text generation to grammar characteristic of a proficiency level. Concretely, we (1) evaluate the ability of GPT3.5 and GPT4 to generate example sentences for the standard English Grammar Profile CEFR taxonomy using few-shot in-context learning, (2) train BERT-based detectors with these generated examples of grammatical patterns, and (3) control the grammatical complexity of text generated by the open Mistral model by ranking sentence candidates with these detectors. We show that the grammar pattern instantiation quality is accurate but too homogeneous, and our classifiers successfully detect these patterns. A GPT-generated dataset of almost 1 million positive and negative examples for the English Grammar Profile is released with this work. With our method, Mistral's output significantly increases the number of characteristic grammar constructions on the desired level, outperforming GPT4. This showcases how language domain knowledge can enhance Large Language Models for specific education needs, facilitating their effective use for intelligent tutor development and AI-generated materials. Code, models, and data are available at https://github.com/dominikglandorf/LLM-grammar.

## 1 Introduction

The arrival and accessibility of well-performing Large Language Models (LLMs) created a flood of applications in personalized education for tutoring and material creation (Kasneci et al., 2023). Despite their ability to follow instructions, it is underexplored to what extent prompting can systematically affect the linguistic properties of the generated output to satisfy educational needs. If LLM-generated text was finely adjustable regarding the grammatical constructs used, personalized and engaging learning materials could systematically support learners' language development by exposing them to the optimal linguistic complexity (Mart, 2013). This control would enable a stronger connection to input-oriented theories of language acquisition.

Due to their data-driven nature, LLMs' grammatical knowledge has to be empirically examined. On the one hand, they have been successfully used for text simplification and grammar construction detection (Jeblick et al., 2023; Weissweiler et al., 2022). On the other hand, transformer models still benefit from explicit syntactic information during training (Hu et al., 2020). Because of missing labeled training data and systematic evaluations, it is uncertain to what extent neural text generation can be controlled for the presence of a comprehensive set of pedagogically relevant and teachable grammatical constructions.

This work pursues the questions of how well LLMs can create valid examples for grammar constructs (RQ1), how well BERT sentence embeddings represent these grammar constructs (RQ2), and how well text generation can be controlled for these constructs (RQ3). We build on an empirically established and validated taxonomy of English grammar, the English Grammar Profile (EGP) (O'Keeffe and Mark, 2017), precisely characterizing the development of English across the proficiency spectrum with 1,222 grammar patterns. We first evaluate how well GPT3.5 and GPT4 can generate positive and negative instances on a subset of the EGP (RQ1). We then alleviate the lack of examples by automatically creating 946K labeled example sentences for all entries of the EGP, which we



*Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 299–308

make available to the public. This unique dataset serves to fine-tune and evaluate BERT-based classification models on detecting examples of the EGP's grammar patterns in sentences (RQ2). Using these models, a grammar-controlled text generation approach to strategically decoding an open pre-trained LLM, Mistral-7B, provides a proof of concept with 600 generated texts (RQ3). To generate them, we sample multiple candidate sentences at inference time and rank them by the grammar patterns detected by the classifiers.

We show that the accuracy of generated instances of grammar patterns is 87.1% with GPT3.5 (92.9% with GPT4), and the classifiers distinguish the positive from negative examples in our generated dataset with an average accuracy of 95.1%. The grammar-controlled text generation approach at least doubles the grammatical constructions on each level of the standard Common European Framework of Reference for Languages (CEFR)({Council of Europe}, 2020).

Going beyond the specific task, our work highlights how explicit domain knowledge relevant to language learning and broader language-sensitive educational contexts can be fused with the versatility of LLMs. It is a step towards better control over a powerful tool compared to pure prompting. The approach can readily be extended to other pedagogically desirable attributes of LLM-based tutors and educational material.

## 2 Related Work

### 2.1 Grammatical complexity in education

Krashen's Input Hypothesis about language learning features the idea that input is an essential driver of language development if understandable to a learner but one step beyond their language level (Krashen, 1992). Although criticized for the vagueness of the theory's predictions, the role of input is broadly accepted in the literature (Lichtman and VanPatten, 2021; Loewen, 2021; Ellis, 2002). Learners benefit from language input adapted to their proficiency level. This assumption manifests itself in *graded readers*, such as simplified literature for learners. Not only do they adapt lexical features but also grammatical complexity (Zakaria et al., 2023). Berendes et al. (2018) systematically analyzed textbooks and highlighted the need to pay more attention to language complexity in subject-matter teaching regarding learner appropriateness. Indeed, research on language-sensitive education in science and other subjects stresses that learning difficulties often arise due to factors such as the syntactic complexity of the language used (Wellington and Osborne, 2001). The success of graded readers and these shortcomings underline the importance of controlling grammar in learning materials for effective language development and the potential impact of automated control.

O'Keeffe and Mark (2017) compiled and published the English Grammar Profile based on the systematic analysis of learner data from language proficiency exams. The EGP includes 1,222 grammar constructs that learners use on different levels, categorized by the standard CEFR level, from A1 (beginner) to C2 (native). They are organized into 19 categories (e.g., adverbs) and can be of type *FORM*, *FORM/USE*, or *USE*. FORM means constructs that can be described lexically and syntactically, whereas USE refers to a semantic function of a linguistic form. The EGP includes a brief description in the form of a can-do statement and one to five authentic learner examples for each structure, as illustrated in Figure 1.

Research on fostering adaptive language learning has started to use developmentally proximal input, though it typically does so by selecting from existing materials (Chen et al., 2022). The EGP's instance-based characteristics of grammatical development allow for fine-grained adaptivity in language teaching because each construct is teachable (and indeed, many are explicitly specified as part of school curricula), which contrasts with the typical aggregate measures and ratios used in linguistic complexity research as part of the Complexity, Accuracy, and Fluency triad (Housen et al., 2012). Thus, the EGP can be a milestone in measuring the grammar complexity of learner input, which is especially valuable when generating material for learners in earlier stages of development, for which little authentic language material exists. However, no large-scale corpus annotated with the EGP constructions is publicly available, yielding the need for our novel dataset.

### 2.2 Grammar-related tasks in natural language processing

Recent LLMs are performant on high-level grammar-related tasks such as essay complexity scoring (Yancey et al., 2023) and text simplification (Jeblick et al., 2023), suggesting a general grasp of grammatical structures. Low-level tasks include grammar annotation, for example with a pre-trained

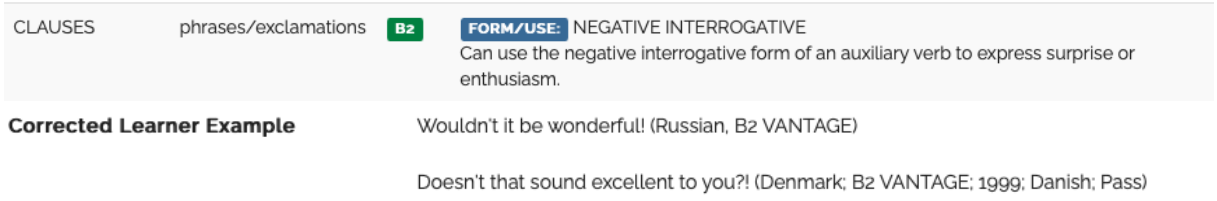


Figure 1: An English Grammar Profile construct at level B2 with two examples

BERT model (Devlin et al., 2019). Weissweiler et al. (2022) successfully detected the presence of the comparative correlative in English with logistic regressions on BERT sentence representations. Yu et al. (2023) also argue for the potential of LLMs for linguistic annotation compared to traditional natural language processing techniques, especially for semantic features without a mapping to lexical forms. Their results for annotating acts of apologizing hint that LLMs can distinguish complex grammatical functions of words and can potentially solve tasks demanding grammatical knowledge. The only work that classified an EGP-alike set of constructions from SCoRE (Chujo et al., 2015) used BERT models to detect three constructions and was successful in increasing their likelihood in generated dialog responses via reinforcement learning (Okano et al., 2023). Unfortunately, the construction-wise reinforced models cannot be combined, making the approach challenging to scale.

Controlled text generation has developed from decoding strategies and supervised fine-tuning (Xiao et al., 2023) to prompt engineering (Koraishi, 2023) and preference optimization approaches (Rafailov et al., 2023). Apart from Okano et al. (2023), past work on syntactic constraints usually worked on parse trees or part-of-speech sequences, which are not directly mappable to curricular grammar patterns (Sun et al., 2023). Especially EGP patterns of the type *USE* are semantic and impossible to represent in this form. Advanced controlled text generation approaches are out of the scope of this work, but the resulting classifiers of this work can be incorporated into all of these approaches.

## 3 Method

Our approach comprises validating the EGP instantiation capabilities of state-of-the-art LLMs, training neural rule detectors on a generated large-scale grammar dataset, and using these rule detectors to rank candidates when sampling from an open text generation model. The analysis was conducted with standard Python libraries for natural language processing and deep learning on up to 16 Nvidia GeForce RTX 2080 Ti GPUs provided by the computing cluster of the University of Tübingen. The code and data are available on GitHub[1]. Seeds are provided for reproducibility.

### 3.1 Instantiating the English Grammar Profile

This step evaluates the possibility of automatically sourcing a high-quality labeled dataset of single grammar constructions. The English Grammar Profile is obtained from its official website[2]. Its structure is characterized in Section 2.1. The information about the learner and the uncorrected examples are removed. We prompt the OpenAI Chat Completion API[3] to generate more examples, namely positive instances of the rule and negatives that ought to have the same meaning without using the construct (i.e., a minimal pair). We evaluate two model checkpoints for comparison, `gpt-3.5-turbo-1106` and `gpt-4-0125-preview`, using in-context learning with a prompt template to describe the grammar rule and append the one to five available examples. If present, the numerical value for the lexical range is translated into *low*, *medium*, and *high*. After the list of positive examples is returned, a second prompt asks to rewrite every example as a minimal pair without using the construction. These are the exact prompts:

```
1. Learn the grammar rule "{Can-do statement}" ({Super Category}, {Sub Category}, {Guideword}). It is CEFR level {Level}. {Lexical Range}
Examples:
{Examples}
```

[1] https://github.com/dominikglandorf/LLM-grammar

[2] https://www.englishprofile.org/english-grammar-profile/egp-online

[3] https://platform.openai.com/docs/models/overview

```
Create {Batch Size} more examples
    using that rule.

2. {Previous Prompt and Response}
Rewrite each created example as a
    minimal pair that does not
    show the usage of the given
    rule.
```

Using regular expressions, the model responses are parsed based on the enumeration, cleaned from prefixes and explanations in parenthesis, and cleared from repetitions of the positive examples in the case of negative instances. The `presence_penalty` parameter that penalizes repetitions of tokens during sampling from the model was increased to 0.5 for the initial prompt to diversify the vocabulary within one response. The model temperature that makes the output more random for higher values was decreased to 0.5 for the negative prompt to favor correctness over diversity. This assumes that there is only a small number of possible modifications to make a positive example negative and therefore the sampling should favor the most likely tokens. The EGP may or may not be part of the training set of OpenAI's models. Even if this is the case, it remains unclear how well they can transfer the patterns to a wider range of topics and sentence meanings than the few included examples.

For a small-scale quality assessment (before generating the large dataset in the next step), 36 EGP patterns are randomly drawn, stratified by CEFR level and type, and the two models generate each 20 (in two batches of 10) positive and 20 negative examples each, resulting in 2,880 examples. The set of sentences is hand-coded on whether they include the intended grammar pattern or not in a blinded manner, i.e., without knowing the model or intended label. These labels serve to calculate the models' accuracy. An automatic evaluation based on the ROUGE and BLEU scores assesses how close the negative examples are to the most similar positive example. The ROUGE-1 score (ranging from 0 to 1) reflects the number of common unigrams between a text and the set of reference texts, measuring lexical similarity. The BLEU score is in the same range but focusses more on precision instead of recall and also takes longer subsequences into account. Furthermore, the average cosine similarity of embeddings with the recent `ember-v1` model[4] between all positive example sentences and between all negative sentences is calculated per EGP pattern and compared to the baseline of the renowned Brown corpus (Kučera et al., 1967). To improve the diversity of negative examples, positive examples from other EGP entries are mixed in, assuming that these do not contain the pattern.

### 3.2 Detecting instances of grammar patterns

This step poses the challenge of learning a binary classifier that detects the presence of a single EGP construct in a given sentence. The bidirectional transformer architecture led to a breakthrough in natural language understanding and was also used by prior work on grammar detection (Okano et al., 2023; Weissweiler et al., 2022). Due to the large number of EGP constructs, we use multi-task training. We choose BERT instead of non-neural tools due to the much lower cost of development, only requiring accurate training data. We fine-tune a pretrained instance of `bert-base-uncased` (Devlin et al., 2019) with model dimensionality 768 and 12 attention layers (110M parameters) as a shared embedding model for each of the six CEFR levels. We train for each single construction a two-layer feedforward network with a hidden dimensionality of 16 on the mean-pooled output from the shared model (12,320 extra parameters per construction). This is a compromise between optimal performance by fine-tuning an entire BERT for each construction and saving the vast amount of GPU memory that this would entail. We did not explore other model architectures because preliminary results have been satisfying.

We use `gpt-3.5-turbo-1106` to create 500 unique positive and 250 unique negative examples for *each* EGP construct in batches of 25 because the model often refused to create larger batches. This results in the large-scale dataset of 946,246 sentences we release with this work. During training, we add 250 random positive examples from other constructs labeled as negative to increase the diversity of the dataset, assuming these do not contain the rule. This leads to a total of 109K (CEFR A1) to 338K (CEFR B1) sentences to train and evaluate each of the six models. Gradients were accumulated across batches of all constructs before taking an optimizer step to balance the influence of a single construct. The batch size was 8, the learning rate for the AdamW optimizer was 0.0001, and training was stopped as soon as the validation loss increased or after a maximum of 5 epochs. We release the trained models with data.

[4] https://huggingface.co/llmrails/ember-v1

We use 5-fold cross-validation for evaluation and do not pursue systematic hyperparameter tuning due to satisfying initial experimentation results. Because of the balanced classes, accuracy is the primary evaluation metric besides precision and recall.

### 3.3 Controling text generation for grammar patterns

This step uses the trained classifiers to control language model output for grammar patterns. Caused by the lack of authentic text annotated with single EGP entries, the CEFR level is used as a proxy. Ideally, a text for a certain level exposes the reader to a high amount of grammar constructions on that level. Thus, the goal is to generate texts with the most EGP constructs of a given level, as indicated by the previous step's classifiers. A CEFR-labeled dataset that was compiled from online resources[5] serves as the static baseline. It contains 1,494 texts on all CEFR levels, 37,008 sentences in total.

We generate 600 texts (100 per level) with each method for comparison. As the LLM baselines, `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023) and `gpt-4-0125-preview` are prompted to continue the first words of given writings with as many grammar constructs as possible on a specific CEFR level, explained with its official description ({Council of Europe}, 2020). Mistral-7B is a model with an architecture and training procedure comparable to the GPT models but with efficiency adjustments. We relied on Mistral-7B due to its appealing trade-off between model size and performance and added GPT4 as the best-performing, closed model at the time. We ran Mistral in inference mode on our cluster infrastructure on two of the GPU cores.

In our proposed ranking approach, the model, prompted in the same way as the baselines, generates five sentence candidates, and the candidate with the most grammar constructs on the desired level is chosen in the remaining generation procedure. This approach is supposed to succeed if the generated candidates show a significant variance in grammar constructions. Tyen et al. (2022) chose a similar ranking approach to generate dialog responses of a specified CEFR level but was using a classifier predicting the CEFR level of candidates instead of explicitly the presence of grammatical structures. Although possible, we did not use a smaller set of preferred EGP patterns because of the large number of constructions and the potential inefficiency of sampling. Past work has also emphasized that a single grammar rule may not be sensible in every generated sentence (Okano et al., 2023). The prompt comprised at least the first 50 characters (adding characters up to the following space) from randomly drawn texts in the CEFR dataset to set different topics of the stories:

```
[INST] Continue the writing using
    as many grammar constructs on
    CEFR level {level} as
    possible ({level description})
    . Do not talk about the CEFR
    level.
[/INST] {story beginning}
```

We stop generation when the continuation exceeds 1,024 characters (Mistral) or 256 tokens (GPT4). The evaluation metric is the average percentage of detected constructions in the corresponding text level.

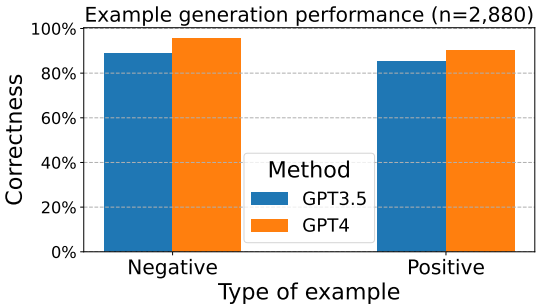


Figure 2: Ratio of correctly generated instances by model and type of example.

## 4 Results

### 4.1 Grammar Pattern Instantiation

#### 4.1.1 Accuracy

Figure 2 summarizes the manually evaluated quality of the two models on generating instances of 36 sample EGP entries. On average, GPT4 got overall 92.9% of the generated instances right, while GPT3.5 scored 87.1%. This difference holds for positive and negative examples, while both models score a few percentage points worse on positive examples. This indicates that they got some rules wrong in the first place. Since all four conditions are accurate far above the random baseline of 50%, the accuracy of the LLM-generated examples is satisfying, and the next steps can build on this technique.

[5] https://www.kaggle.com/datasets/amontgomerie/cefr-levelled-english-texts

Table 1: ROUGE and BLEU scores of negative examples versus positive examples

| Model & Parameters | ROUGE-1↑ | BLEU↑ |
|---|---|---|
| GPT3.5, temp=1 | 0.704 | 0.268 |
| GPT3.5, temp=0.5 | **0.783** | **0.368** |
| GPT4, temp=1 | 0.721 | 0.283 |

**EGP#777:** Can use the past perfect simple to talk about situations which changed.
+ They had expected to win the match, but their opponents played exceptionally well.
- They expected to win the match, but their opponents played exceptionally well.
+ She had thought she had everything under control, but then the unexpected happened.
- She thought she had everything under control, but then the unexpected happened.
+ We had believed we had enough time to finish the project, but unforeseen complications arose.
- We believed we had enough time to finish the project, but unforeseen complications arose.

**EGP#288:** Can use no article before a limited range of singular, plural and uncountable nouns when referring to things in general.
+ Dogs are my favorite animals.
- The dogs in the park are friendly.
+ Milk is good for your bones.
- The milk in the fridge is expired.
+ I don't like carrots.
- The carrots in the salad are fresh.

Figure 3: Generated positive (+) and negative (-) examples for an EGP entry with very high average ROUGE and BLUE scores (top) and one with very low scores (bottom).

#### 4.1.2 Minimality

Table 1 shows the automatic quality assessment of the minimality of the negative examples, measured by their ROUGE and BLEU scores with respect to the positive examples. Interestingly, the temperature is more critical than the general performance of the model. Concretely, GPT3.5 with decreased temperature performs better than GPT4 with the default temperature. This hints at the importance of reducing the randomness when sampling from the language model output. Figure 3 shows generated examples for two EGP entries. These instances show that there may be rules for which minimal negative examples are easier to create. For construct #288, one could just add the article but would make the sentences potentially ungrammatical. This shows that the model also takes care of the correctness of the generated examples.

Table 2: Average cosine similarities between sentences in authentic text (Brown corpus) and the positive and negative examples generated by GPT3.5. *Random others refer to negative examples with random positive examples from other constructs.

| Corpus | Similarity | Std. Dev. |
|---|---|---|
| Brown | **0.334** | 0.002 |
| Positive examples | 0.462 | 0.052 |
| Negative examples | 0.451 | 0.045 |
| Random others* | 0.369 | 0.007 |

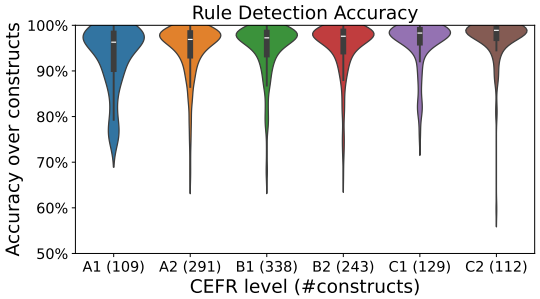


Figure 4: Accuracy distributions of the grammar classifiers across CEFR difficulty levels. Variation between cross-validation folds is negligible. The baseline is 50%. The white dash indicates the median and the pronounced black strip the interquartile range.

#### 4.1.3 Diversity

The diversity of the generated examples, indicated by the average sentence similarity within the generated EGP patterns, is represented by Table 2.

To some extent, the similarity between the examples is expected to be higher due to the presence of the grammar pattern. Still, we observe a significantly increased cosine similarity for both positive and negative pairs compared to the Brown reference corpus. Adding positive examples from other EGP constructs increases the diversity, yielding an average cosine similarity increased by only 10% compared to the reference corpus. Overall, the evaluation confirms the capability of state-of-the-art LLMs to augment a grammar pattern dataset from a class description and a few examples with accurate positive and negative examples, only lacking diversity within the positive examples.

### 4.2 Grammar Pattern Detection

Figure 4 depicts the accuracies of our BERT-based models at detecting whether a given grammar construction is present in a sentence.

The average accuracy of all classifiers is 95.1%, precision is 93.3% and recall is 97.3%. The distri-

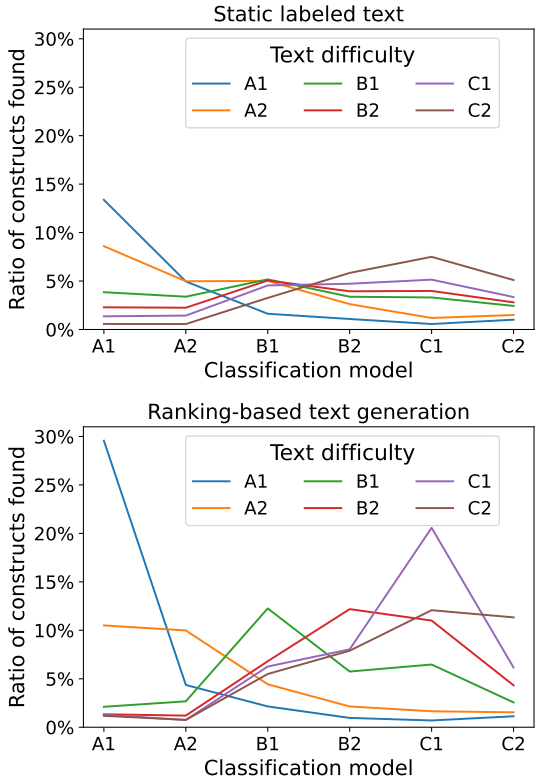


Figure 5: Number of grammatical constructions per CEFR levels for the static baseline (top) and our proposed approach (bottom). The trained classifiers from the previous step detected constructions.

butions of precisions and recalls are mean-shifted but overall very similar and are included in our GitHub repository. The average recall attains at least 91% among all CEFR levels. The lower precision may be explainable by false negatives added for diversification and the quality differences between positive and negative examples. The accuracy distribution within CEFR level A1 reveals slight problems detecting some of these constructions. This may be due to the very basic character of many A1 grammar patterns. This likely also increases the number of false negatives in the random negatives from other constructs. Overall, the classification quality seems near optimal given the quality of the augmented data, which sets an upper performance bound. Due to eliminating duplicates and having 25% random examples from other constructs in the validation set, the accuracy can exceed the 87.1% example accuracy from the manual evaluation.

### 4.3 Grammar-Controlled Text Generation

Table 3 lists the average ratio of detected grammar constructions on the given level across all sentences, as detected by the trained classifiers from the previous step.

The two LLM baselines, which employ pure prompting, already show improvements over the static baseline of CEFR-annotated texts. GPT4 increases the frequency of EGP entries on all requested levels. The Mistral baseline shows less pronounced improvements and fails to increase the number of grammar constructs on levels A2 and C2. Generally, the pre-trained models have difficulties using more constructs of the levels A2, B1, and B2. Our approach to ranking sentence candidates during text generation has a severe positive impact on the distribution of grammar constructions across all six levels. For all levels, the ratio of applied grammar rules has at least doubled, on level C1 it has even quadrupled versus the baseline. This proves that the variance within different generated candidates regarding the grammatical constructions is sufficient, although the prompt included the instruction to control text complexity. Figure 5 provides a bigger picture of the generated text characteristics between the static baseline versus our method.

While the grammatical constructions in the corpus are much more evenly distributed across all text difficulties, our ranking approach can create visible spikes on the desired complexity level while roughly maintaining the frequencies of other levels' patterns. Only on requested level B2, constructs of level C1 are also increased which may even help scaffolding. Overall, the intervention seems to help control the desired pedagogical properties of generated text.

## 5 Discussion and Conclusion

This work showcases how Large Language Models can be controlled based on the qualitative EGP augmented to a large-scale dataset to align with pedagogical use cases, specifically – but not limited to – language teaching. We first verified the sufficient quality of LLM-generated instances of an established grammar repository, the English Grammar Profile. The validation emphasizes the strength of the most recent closed-source model, GPT4. Nevertheless, the quality of instances generated by GPT3.5 could almost keep up with the flagship model. Because of this positive finding, we generated 946K labeled grammar construction examples, which we publicly share for further research. The binary grammar construction classification on this data shows satisfying results within the distribution

Table 3: Ratio of detected constructs by CEFR level of the corresponding texts on the same level.

| Method | A1↑ | A2↑ | B1↑ | B2↑ | C1↑ | C2↑ |
|---|---|---|---|---|---|---|
| Static Baseline | 13.4% | 5.0% | 5.2% | 3.9% | 5.1% | 5.1% |
| GPT4 | 22.2% | 5.7% | 7.0% | 7.3% | 14.2% | 10.9% |
| Mistral Prompting | 16.1% | 4.2% | 6.1% | 6.5% | 9.7% | 4.6% |
| Mistral Candidate Ranking (Ours) | **29.6%** | **10.0%** | **12.3%** | **12.2%** | **20.6%** | **11.3%** |

of generated data. The results are close to similar research that has not used minimal pairs and shared embedding models and solved a potentially easier problem (Okano et al., 2023). Controlling an open LLM such as Mistral on the used grammar constructions with these classifiers significantly affects the frequency of desired grammar patterns. It can even beat the baseline of prompting GPT4. While the prompt-based strategies already improve over the static baseline for most CEFR levels, our proposed approach has improved text on every proficiency level and at least doubled the default frequency of constructs on all levels. It also solves the shortcoming of Tyen et al. (2022) that had difficulties generating text on the simpler levels A1 and A2.

With the advent of performant open LLMs, such as Llama3 and Mixtral of Experts, educational applications can be further tailored to align with pedagogical expectations than with prompting alone. Currently, instructors can only use commercial model interfaces such as ChatGPT or third-party wrappers around the model endpoints. Our method advances the possibilities from prompt engineering approaches to fine adjustment of the model output. We freely release our augmented dataset and the trained grammar classifiers to provide learning engineers with a tool to introduce this level of control to their applications. A possible application is adjusting the grammatical complexity of an AI tutor in science to the language proficiency level of each student. Non-native speakers in the same classroom can interact with the seemingly "same" agent that adapts its language to them under the hood. Language instructors can use models to generate texts of students' interests while ensuring the use of particular grammar that aligns with their curriculum.

### 5.1 Ethical considerations

The EGP was created and annotated by experts to empirically identify the characteristics of the English used by learners at different levels of proficiency. While the data stems from official proficiency tests taken by a wide variety of language learners worldwide, the language used may still be biased by the test tasks, the opinions expressed by the learners who took the tests, and the selection of learner data selected as examples for the EGP. Instructing the LLM to focus on grammatical structure instead of content should mitigate such bias in the generated dataset, though this is not guaranteed. The grammar classification may thus work better for topics typically used by a specific student subgroup in particular language tasks. The authors also acknowledge the potential critique of the CEFR classification as eurocentric (O'Keeffe and Mark, 2017).

Another consideration related to the use of LLMs is the potential generation of toxic or biased language, which is especially sensitive when underage students are working with an LLM-based language learning tool (Meyer et al., 2023). On the pedagogical side, the use of artificially generated text may also limit authenticity and thereby reduce learner motivation. Finally, interacting with a machine to foster language acquisition will not offer the same social benefits and challenges as human interaction.

### 5.2 Future Work

Future work should build real applications for the educational text generation approach. Then, a controlled field experiment should be pursued to assess the impact on students' language acquisition. It should survey the perception of the generated texts by teachers and students and measure learning gains. This may reveal details about potential weaknesses and issues for example with lexical complexity for which our approach does not explicitly control. With more invasive adaptation techniques, the approach can be easily extended to single grammar constructions and adapt grammar not only to the holistic proficiency level of the learner but to the knowledge of single grammar patterns. The grammar constructions should be further located within the sentences to increase the detection quality and enable annotations. This enables

more precise input enhancement applications.

## 6 Limitations

The training data for grammar classification has some drawbacks. Having only many positive and negative examples is likely insufficient for robust control over single grammar patterns in educational text generation. The models usually use the same sentence structure for creating new instances, especially given the scarcity of seed examples in the EGP. Although the classifiers can learn most of the differences within the generated dataset, it remains unclear how well the classifiers generalize to other models' generative distributions or real-world corpora. More diverse examples must be fostered, and a manual validation of grammar construction detection in real corpora would be needed.

In the text generation step, we only maximized the amount of constructs on the desired CEFR level. A suitable text likely also requires reducing the number of overly difficult constructs to not confuse the learner and better target the zone of proximal development. One could add a parameter that balances how large the penalty for the presence of more difficult grammar should be. Furthermore, some grammar patterns may occur too infrequently in sampling from a pre-trained model, and generating many candidates to obtain at least one positive instance would be inefficient. This can only be overcome by adapting the weights of the pre-trained language model or advanced decoding strategies. Therefore, we tested the approach only on the six groups of grammar construction, as given by their CEFR level, which limits the current approach to less fine-grained control over text generation. However, we believe this can still serve as a proof of concept.

We are also aware that the English Grammar Profile is a description of the typical proficiency level when learners start to use a grammar pattern. This can only serve as a proxy for reading comprehension, which is the focus of this work. Fortunately, our grammar classifiers can serve to analyze existing materials that are expert-curated to create a valid mapping to reading comprehension levels instead of written production.

## 7 Acknowledgements

We gladly acknowledge the use of the EGP using their officially requested statement: This publication has made use of the English Grammar Profile. This resource is based on extensive research using the Cambridge Learner Corpus and is part of the English Profile program, which aims to provide evidence about language use that helps to produce better language teaching materials. See http://www.englishprofile.org for more information.

## References

Karin Berendes, Sowmya Vajjala, Detmar Meurers, Doreen Bryant, Wolfgang Wagner, Maria Chinkina, and Ulrich Trautwein. 2018. Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*, 110(4):518–543.

Xiaobin Chen, Detmar Meurers, and Patrick Rebuschat. 2022. ICALL offering individually adaptive input: Effects of complex input on L2 development. *Language Learning and Technology*, 26(1):1–21.

Kiyomi Chujo, Kathryn Oghigian, and Shiro Akasegawa. 2015. A corpus and grammatical browsing system for remedial EFL learners. In Agnieszka Leńko-Szymańska and Alex Boulton, editors, *Studies in Corpus Linguistics*, volume 69, pages 109–128. John Benjamins Publishing Company, Amsterdam.

{Council of Europe}. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing, Strasbourg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Nick C. Ellis. 2002. FREQUENCY EFFECTS IN LANGUAGE PROCESSING: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*, 24(2):143–188.

Alex Housen, Folkert Kuiken, and Ineke Vedder. 2012. *Dimensions of L2 Performance and Proficiency*. John Benjamins Publishing Company, Amsterdam.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P. Levy. 2020. A Systematic Assessment of Syntactic Generalization in Neural Language Models.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, and Michael Ingrisch. 2023. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European Radiology*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Osama Koraishi. 2023. Teaching English in the age of AI: Embracing ChatGPT to optimize EFL materials and assessment. *Language Education and Technology*, 3(1).

Stephen Krashen. 1992. The input hypothesis: An update. *Linguistics and language pedagogy: The state of the art*, pages 409–431.

Henry Kučera, Winthrop Francis, William Freeman Twaddell, Mary Lois Marckworth, Laura M. Bell, and John Bissell Carroll. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence.

Karen Lichtman and Bill VanPatten. 2021. Was Krashen right? Forty years later. *Foreign Language Annals*, 54(2):283–305.

Shawn Loewen. 2021. Was Krashen right? An instructed second language acquisition perspective. *Foreign Language Annals*, 54(2):311–317.

Cagri Tugrul Mart. 2013. Teaching grammar in context: why and how? *Theory & Practice in Language Studies*, 3(1):124–129.

Jesse G. Meyer, Ryan J. Urbanowicz, Patrick C. N. Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J. Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, and Jason H. Moore. 2023. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1):20.

Yuki Okano, Kotaro Funakoshi, Ryo Nagata, and Manabu Okumura. 2023. Generating Dialog Responses with Specified Grammatical Items for Second Language Learning. In *BEA 2023*, pages 184–194, Toronto, Canada. ACL.

Anne O'Keeffe and Geraldine Mark. 2017. The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4):457–489.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating Large Language Models on Controlled Generation Tasks. In *EMNLP 2023*, pages 3155–3168, Singapore. ACL.

Gladys Tyen, Mark Brenchley, Andrew Caines, and Paula Buttery. 2022. Towards an open-domain chatbot for language practice. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 234–249, Seattle, Washington. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your Syntax, the better your Semantics? Probing Pretrained Language Models for the English Comparative Correlative. In *EMNLP 2022*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jerry Wellington and Jonathan Osborne. 2001. *Language and Literacy in Science Education*. McGraw-Hill Education, UK.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications. In *BEA 2023*, pages 610–625.

Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating Short L2 Essays on the CEFR Scale with GPT-4. In *BEA 2023*, pages 576–584, Toronto, Canada. ACL.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2023. Assessing the potential of AI-assisted pragmatic annotation: The case of apologies.

Azrifah Zakaria, Willy A. Renandya, and Vahid Aryadoust. 2023. A Corpus Study of Language Simplification and Grammar in Graded Readers. *LEARN Journal: Language Education and Acquisition Research Network*, 16(2):130–153.