Generation and Evaluation of English Grammar Multiple-Choice Cloze Exercises

Nicolò Donati[1,2,*,†], Matteo Periani[1,†], Paolo Di Natale[3,†], Giuseppe Savino[2] and Paolo Torroni[1]

[1]University of Bologna, Viale del Risorgimento, 2, 40136 Bologna BO, Italy
[2]Zanichelli editore S.p.A., Via Irnerio 34, 40126 Bologna, Italy
[3]University of Bologna, Corso della Repubblica, 136, 47121 Forlì FC, Italy

Abstract

English grammar Multiple-Choice Cloze (MCC) exercises are crucial for improving learners' grammatical proficiency and comprehension skills. However, creating these exercises is labour-intensive and requires expert knowledge. Effective MCC exercises must be contextually relevant and engaging, incorporating distractors—plausible but incorrect alternatives—to balance difficulty and maintain learner motivation. Despite the increasing interest in utilizing large language models (LLMs) in education, their application in generating English grammar MCC exercises is still limited. Previous methods typically impose constraints on LLMs, producing grammatically correct yet uncreative results. This paper explores the potential of LLMs to independently generate diverse and contextually relevant MCC exercises without predefined limitations. We hypothesize that LLMs can craft self-contained sentences that foster learner's communicative competence. Our analysis of existing MCC exercise datasets revealed issues of diversity, completeness, and correctness. Furthermore, we address the lack of a standardized automatic metric for evaluating the quality of generated exercises. Our contributions include developing an LLM-based solution for generating MCC exercises, curating a comprehensive dataset spanning 19 grammar topics, and proposing an automatic metric validated against human expert evaluations. This work aims to advance the automatic generation of English grammar MCC exercises, enhancing both their quality and creativity.

Keywords

Large Language Models, Distractor Generation, Multiple-Choice Cloze, Evaluation Metric

1. Introduction

English grammar Multiple-Choice Cloze (MCC) exercises are widely used tools for enhancing a learner's grammatical proficiency and comprehension skills. They consist of fill-the-gap questions where the gap must be filled by choosing one correct solution (*key*) among several options. The incorrect alternatives are called *distractors*. Devising these exercises is a labour-intensive process requiring expert knowledge in language teaching and content creation. The exercises must be contextually relevant to help learners understand how rules apply in real-life situations. This requires crafting sentences and scenarios that are both engaging and educational. Learners have different levels of proficiency, from beginners to advanced. Striking the right balance ensures that learners are neither bored nor frustrated, which is crucial for maintaining their motivation and progress. In MCC exercises this is done by choosing distractors that are incorrect but plausible, thus keeping the exercise challenging for the learner. Studies in Communicative Language Teaching demonstrate that the learner must possess the knowledge of grammatical structures and the ability to compose syntactically well-formed propositions, and they must also acquire the ability to employ grammatical forms in discourse [1][2].

Recently, there has been a growing interest in applying LLMs in education [3]. However, the adoption of LLMs for English grammar MCC exercise generation is still limited. Some proposals focus on testing vocabulary [4] or use LLMs by constraining their generation capability, for example using fixed part-of-speech sequences [5]. Although the outputs of these models are grammatically correct typically they lack creativity [6].

In this work, we investigate the potential of LLMs in automatic exercise generation without hampering their creativity. Our working hypothesis is that LLMs can generate self-contained sentences, recreating situational contexts that elicit the *communicative competence* of the learner [7]. Our main objective is to understand to what extent can LLMs generate accurate grammar exercises without providing predefined constraints or POS sequences. To pursue this objective, we analyzed the available English grammar MCC exercises dataset [8]. We observed that it has limited diversity, some topics are underrepresented, and there are often mistakes. Existing literature does not offer a single agreed-upon automatic metric for evaluating the quality of the generated gram-

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
†These authors contributed equally.
n.donati@unibo.it (N. Donati); matteo.periani2@studio.unibo.it (M. Periani); paolo.dinatale3@studio.unibo.it (P. D. Natale); gsavino@zanichelli.it (G. Savino); p.torroni@unibo.it (P. Torroni)
0009-0000-5673-5274 (N. Donati); 0000-0002-9253-8638 (P. Torroni)

mar exercises. Therefore, we set out to identify such a metric and validate its alignment with human judgment. In this paper, we present a novel solution utilizing an LLM to generate English grammar MCC exercises. Our contribution also focuses on curating an MCC dataset that spans 19 topics. Lastly, we propose an automatic metric to evaluate the exercise's correctness and verify the validity of our contribution thanks to human expert evaluation.

2. Task description

Grammar exercises should define the range of abilities to be assessed and avoid the influence of irrelevant factors like past knowledge or cultural background [9]. We followed the Best-practice guidelines for creating grammar MCC items defined in [10] [11]. According to them, each item consists of three components.

- **Body**: the sentence with a gap in place of the key.
- **Key**: the correct answer.
- **Distractor**: the incorrect answer.

The body plays a central role in designing effective exercises. Learners should be able to infer the key based on the helpful elements present in the body. However, the effectiveness of an exercise depends mainly on the quality of its distractors. Ideally, challenging distractors should be homogeneous, plausible, and unambiguous. Homogeneous distractors share the same syntactic category as the key [12]. Plausible distractors provide a credible alternative to the key. Lastly, unambiguous distractors ensure that none of them could be considered correct if used in place of the key [10].

3. Related Works

The generation of MCC exercises has been explored from various perspectives. In this section, we will briefly discuss the main related approaches.

3.1. MCC Dataset

Prior works in creating MCC datasets are very limited. To the best of our knowledge, the only one in English was presented by Liu et al. in their work SC-Ques [8]. It comprises real English test items for students developed by teaching professionals. The dataset contains roughly 300k MCC sentence completion exercises, composed of the question body, a varying number of alternative answers, and the key (i.e. the correct alternative). It comprises both exercises with only single or multiple blanks. It has various limitations, discussed in Section 5.

3.2. Grammar MCC Exercise Generation

A large share of prior works uses rules to create Grammar MCC Exercises (Sumita et al. [13], Brown et al. [14], Smith et al. [15], Majumder and Saha [16], Lin et al. [17]). They all follow a three-fold process: (1) select sentences from arbitrary sources, (2) insert the blank into the sentence, and (3) generate distractors for the blank. Sentences usually come from corpora or user-submitted passages. Many solutions restrict gap detection into fixed schemes: Sumita et al. [13] picked out the leftmost single verb, Lin et al. [17] only selected adjectives as a blank. One of the few exceptions is Goto et al. [18], who proposed a method based on Conditional Random Fields (CRFs) [19]. Methods that extract sentences from arbitrary text suffer from several limitations. First of all, they lack customization options, such as adjusting for the subject or difficulty level of the exercise. Additionally, they are limited by the length and quality of the extracted texts, which can negatively impact the system's results.

Recently, parts of MCC generation have been executed by Neural Networks instead of rule-based algorithms. Bitew et al. [20] use a variation of the RoBERTa [21] model to predict the gap positions within the sentence. To decrease the ambiguity Matsumori et al. [22] trained a Masked Language Model for gap score prediction of each candidate sentence. Chomphooyod et al. [23] proposed a system that uses Transformers [24] to generate candidate sentences given a POS sequence, a keyword and a desired grammar topic.

3.3. Metrics

In the literature, the evaluation of MCC exercises is mainly based on judgments expressed by human annotators. Slavuj et al.[25] asked annotators to perform the language tasks, assuming that the presence of incorrect answers would be a sign of ill-formed exercises. Teachers were then asked to provide feedback on any pitfalls they encountered. Malafeev [26] simply attended to suitability for classroom use. Chomphooyod et al. [23] evaluates for each exercise different aspects such as the grammatical and semantic correctness, the relevance with respect to the topic, and its acceptability.

Very few automatic metrics have been proposed to evaluate exercise generation. Bitew et al. [20] rely on span overlap with respect to ground truth to assess the consistency of gap detection. March et al. [27] test the effectiveness of distractors by their selection rate.

Since an important criterion for exercise collection is diversity, often similarity measures have been applied to MCC exercise. Metrics like BLUE [28], ROUGE [29], and METEOR [30] have been used even though originally designed for different applications.

4. Approach

To overcome the limitations of existing solutions, we utilized an LLM to generate exercises in a single, constraint-free step. We chose Llama3 [31] due to its acceptable balance between computational cost and performance. To evaluate its effectiveness, we engineered a well-structured prompt (Appendix B.2). However, the results were unsatisfactory. The model exhibited significant difficulties with certain grammar topics and consistently failed to generate effective distractors. Therefore we decided to fine-tune the model using a well-formatted dataset containing exercises with distractors that meet our criteria. Each dataset example includes four features: the grammar topic, the exercise text, the key, and the distractors. The model is trained to produce the exercise text, key, and distractors when given a specific grammar topic as input. The prompt used during the fine-tuning and an example of input-output text can be found in the appendix section B.1.

To assess the correctness of the generated items, we devised metrics that evaluate the minimal structural requirements of an exercise thanks to rule-based analysis. These are defined in section 7. To monitor the results we used SELF-BLEU [6], a metric that inspects repetitions checking continuous lexical overlap.

5. Dataset Curation

We developed the fine-tuning dataset based on the data released by [8]. The data underwent three pre-processing steps: cleaning, grammar topic identification, and removal of similar examples.

Data cleaning First, we got rid of improperly formatted examples and cleaned the text to comply with the tokenizer specifications and limit potential noise. Items with multiple blank spaces or fewer than two distractors were discarded. Next, we filtered out exercise texts containing instructions, non-Latin symbols or letters, emails, phone numbers, and links.

Extraction of the grammar topic The second step involves the assignment of the grammar topic to each exercise thanks to the Pattern Matcher. First, grammar topics are defined in a tailor-made grammar taxonomy with the aid of spaCy Dependency Matcher. Given a set of sentences, this tool allows one to identify whether each sentence features the described grammar topics, and if so, at what position. The relevant topic is chosen by comparing the overlap between the position of the topic detected by Pattern Matcher and the key span[1]. To ensure the exclusively grammatical nature of the exercises, distractors are checked using the metrics proposed in Section 3.3. All exercises lacking valid distractors are then discarded.

Deduplication We deduplicated and removed all the similar exercises, to increase the quality of our dataset [32]. Exercises are clustered by topic and compared in terms of embeddings through cosine similarity. Using a threshold T_p, where p denotes the topic, all elements exceeding the limit are discarded. Lastly, we noticed that SC-Ques [8] had an unbalanced representation of grammar topics. For example, in half of the WH-questions have "How" as the key. For each topic, a maximum ratio of key presence is established, and superfluous data are discarded.

After pre-processing, the least represented class contained a quarter of the examples present in the most represented one. The only exception was the "WH-questions" class, which was underrepresented. Therefore, we upsampled the class with synthetic exercises using GPT-4 [33]. The dataset is composed by several fields: the filled_text (complete exercise sentence), the gapped_text (sentence with a blank gap), the key (the text removed to create the gap), and the list of distractors.

6. Fine-Tuning

We designed the fine-tuning process to generate exercises on specific grammar topics with a fixed number of distractors. The model's expected response is a JSON-encoded exercise coherent to the dataset structure described in Section 5. We observed that including the filled_text in the output improves overall accuracy and reduces similarity among exercises. An example from the fine-tuning dataset can be found in the appendix section B.1. To reduce the computational resources required for fine-tuning, we employed the Quantized Low-Rank Adapters (QLoRA) [34] approach. Our tests on small models revealed that this strategy prevents significant shrinkage of the model's dictionary during fine-tuning. Consequently, the generated exercises exhibit greater variability, enhancing the model's creativity.

7. Evaluation Metrics

Two metrics are used to track the model's performance on diverse aspects. First, we introduce a metric that evaluates the minimal structural requirements of an exercise. Secondly, we control for language diversity to have more interpretability on the results.

[1]The key span is the range of positions the key belongs to.

7.1. Structural Compliance

This metric evaluates the structure and well-formedness of the exercise. Decomposing the validation stage into two steps, we design two rule-based components, namely *pertinence* and *homogeneity*.

The former oversees that the gap placeholder is located in the intended position and that the key includes the correct grammar form. The second component checks that the distractor fulfils the criterion of homogeneity as described in the section 2. To achieve this, grammar topics have been grouped into two classes.

Inflectional They must have the same lemma as the key so as to rule out the influence of lexis and semantics. We also make adjustments to account for circumstances when the key and the distractor are identical, as well as for handling variation of the auxiliary verb.

Free morphemes Exercises of this group limit acceptable keys and distractors to a narrow range of options. So, we manually compile a list of admitted words for each grammar topic. If the distractor belongs to that list and is not identical to the key, it is deemed homogeneous.

Some grammar topics may be built with distractors of any of the two classes. If either of the checks is successful, the distractor passes the test of fitness.

7.2. Language Diversity

LLMs often experience the so-called repetition problem, where their output includes excessively repeated segments of text, creating an undesirable effect [35]. In the context of the generation of thousands of exercises, duplicates or overly similar sentences are highly likely to occur. In order to assess this phenomenon we decided to rely on continuous lexical overlap by using Self-BLEU [6] onto 2-to-5-grams to capture multi-word repetitions.

8. Experiments

We fine-tuned the Huggingface implementation of Meta-Llama-3-8B-Instruct[2]. The model was first quantized to 4-bit precision and then fine-tuned using LoRA adapters, with the following configuration: rank equal to 64, alpha 16, and a dropout percentage of 0.1. The adapters have been added on top of all the attention linear layers to not significantly degrade performance. The training hyperparameters are: a constant learning rate of $2e{-}4$, max gradient norm of 0.3, and a weight decay equal to $1e{-}2$. The number of epochs was set to 3, using a batch size of 1 and gradient accumulation equal to 16. The train lasted two hours on a NVIDIA RTX A6000.

[2]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

9. Results

To evaluate performances, for each grammar topic we generated 50 exercises, setting the number of distractors to 1. We use the sampling decoding strategy with a temperature equal to 0.7 to balance the creativity and the coherence of the output.

The exercises are categorized according to their grammar topic. For each exercise, we assessed its structural compliance and its similarity to the exercises within the same grammar topic that has been labelled structurally correct, using the metrics described in section 3.3. The results are then averaged to obtain the accuracy for each grammar topic. In the end, the model performances are computed by averaging the topic scores. The results are reported in Table 1.

Overall, the outcomes are satisfactory. The model on average scores a Structural Compliance (SC_H) equal to 85%, indicating its ability to generate well formed exercises. It achieves a self-BLEU similarity of 7%, demonstrating that text repetitions are limited. Looking at the individual SC scores, we observe that the model tends to perform better on free morphemes grammar topics. We suppose this is due to the limited number of possible key/distractor options. Furthermore, we observed that due to spaCy limitations in properly labelling certain verbs, grammar topics related to verbal tenses are more prone to be misidentified. This limitation causes occasional misjudgment of the exercise's structural compliance, leading to a negative effect on the topic performance.

9.1. Human Evaluation

To assess classroom suitability a human evaluation was performed on all 950 exercises by a computational linguist with a background in pedagogy in language teaching. Each generated exercise (**EC**) was evaluated on four criteria:*Plausibility*, *Ambiguity* (defined in section 2), *Common Sense*, *Acceptability*. Common Sense means that the exercise sentence should be coherent with common sense. Acceptability indicates that a sentence does not perpetuate stereotypes or display inappropriate content, such as violence. If any of these criteria is not met, the item is flagged as incorrect.

The results presented in table 1 have established that 79% of the items satisfy all the requirements to be administered to learners. We conducted an error analysis. The results are summarized in Table 2. *Common sense* was the most frequently observed inaccuracy, although the magnitude of the issue is modest. As expected, ambiguous

grammar topic	$\mathbf{SC}_A$	self-BLEU	$\mathbf{SC}_H$	EC
articles	0.94	0.03	0.94	0.74
comparison adjectives	0.90	0.09	0.92	0.72
conditional statements	0.76	0.07	0.90	0.66
future simple	0.82	0.06	0.90	0.90
modal verbs	0.62	0	0.78	0.70
infinitive and gerund verbs	0.76	0	0.96	0.86
passive tenses	0.84	0	0.86	0.74
past continuous	0.98	0.16	0.98	0.88
past perfect	0.94	0.12	0.96	0.82
past simple	0.88	0	0.86	0.82
personal pronouns	0.85	0.07	0.92	0.74
possessive adjectives	0.82	0.12	0.90	0.72
prepositions	0.84	0	0.92	0.72
present continuous	0.96	0.11	0.98	0.88
present perfect	0.66	0.08	0.98	0.84
present simple	0.88	0.05	0.88	0.86
quantifiers	0.88	0.07	0.88	0.84
relative clauses	0.94	0.03	0.94	0.74
WH- question	0.98	0.18	1.00	0.90
average	**0.85**	**0.07**	**0.92**	**0.79**

Table 1
Results of the evaluation on the generated exercises. $\mathbf{SC}_A$ is the Structural Compliance evaluate by our metric, $\mathbf{SC}_H$ evaluated by the human annotator and **EC** is the exercise correctness. The double lines divide the results from the automatic metric (left) to those obtained by the human-eval (right). More results on error analysis can be found in table 2.

distractors remain an open matter in the field, especially for tense-based topics. Instead, we can notice that the generation of sentences with bias or trivial exercises is almost absent.

Furthermore, we asked the annotator to evaluate the structural compliance of the exercises ($\mathbf{SC}_H$). Then we computed the Precision, Recall and F1 scores using annotator judgements as golden labels. The results show that our automatic structural compliance metric ($\mathbf{SC}_A$) has an F1 score of 95% w.r.t the human evaluation, with a Precision of 98% and a Recall of 91%. This highlights its effectiveness in predicting the overall structural quality of the exercises.

10. Conclusion

We investigated the use of an LLM to generate English MCC grammar exercises. To that end, we curated a new English grammar MCC exercises dataset. We devised metrics for the automatic evaluation of such exercises. We evaluated our work using said metrics, and a human study involving domain experts. Our findings demonstrate the model's ability to generate exercises suitable for educational use. The generated exercises exhibit a low similarity score, indicating that our method can effectively produce original exercises: a significant advantage from prior art, mostly relying on rule-based methods. We observe that human evaluation correlates positively with the proposed structural compliance metric, corroborating our metric as an indicator of exercise structure correctness and alignment with human expert preferences. We found that a key factor of our method was the availability of high-quality fine-tuning data.

One limitation was the presence of many similar exercises in the SC-Dataset [8] we used to build our resource from. After removing similar exercises, only 30% of the original data was left. Another limitation is the sensitivity of the evaluation metric to the Pattern Matcher, concerning the evaluation of the key and the distractors, which caused some false negatives.

The curated dataset and model will be available to the community.[3].

Acknowledgments

We wish to thank *Zanichelli editore* for their support which enabled data up-sampling, human evaluation, and experimentation with their infrastructure. We also thank Eleonora Cupin for her valuable contribution to the human evaluation of the dataset.

[3] https://github.com/ZanichelliEditore/english-grammar-multiple-choice-generation

References

[1] H. G. Widdowson, Teaching Language as Communication, Oxford University Press, Oxford, 1978.

[2] H. G. Widdowson, Explorations in Applied Linguistics, Oxford University Press, Oxford, 1979.

[3] W. Gan, Z. Qi, J. Wu, J. C.-W. Lin, Large language models in education: Vision and opportunities, in: 2023 IEEE International Conference on Big Data (BigData), 2023, pp. 4776–4785. doi:10.1109/BigData59044.2023.10386291.

[4] Q. Wang, R. Rose, N. Orita, A. Sugawara, Automated generation of multiple-choice cloze questions for assessing english vocabulary using gpt-turbo 3.5, 2024. URL: https://arxiv.org/abs/2403.02078. arXiv:2403.02078.

[5] P. Chomphooyod, A. Suchato, N. Tuaycharoen, P. Punyabukkana, English grammar multiple-choice question generation using text-to-text transfer transformer, Computers and Education: Artificial Intelligence 5 (2023) 100158. URL: https://www.sciencedirect.com/science/article/pii/S2666920X23000371. doi:https://doi.org/10.1016/j.caeai.2023.100158.

[6] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, Y. Yu, Texygen: A benchmarking platform for text generation models, 2018. URL: https://arxiv.org/abs/1802.01886. arXiv:1802.01886.

[7] D. H. Hymes, On communicative competence, in: J. B. Pride, J. Holmes (Eds.), Sociolinguistics. Selected Readings, Penguin, Harmondsworth, 1972, pp. 269–293.

[8] Q. Liu, Y. Huang, Z. Liu, S. Huang, J. Chen, X. Zhao, G. Lin, Y. Zhou, W. Luo, Sc-ques: A sentence completion question dataset for english as a second language learners, in: C. Frasson, P. Mylonas, C. Troussas (Eds.), Augmented Intelligence and Intelligent Tutoring Systems, Springer Nature Switzerland, Cham, 2023, pp. 678–690.

[9] L. F. Bachman, Fundamental Considerations in Language Testing, Oxford University Press, Oxford, 1990.

[10] J. E. Purpura, Assessing Grammar, Cambridge Language Assessment, Cambridge University Press, 2004.

[11] G. Fulcher, G. Fulcher, Practical Language Testing, 1st ed., Routledge, 2010. doi:10.4324/980203767399.

[12] V.-M. Pho, T. André, A.-L. Ligozat, B. Grau, G. Illouz, T. François, Multiple choice question corpus analysis for distractor characterization, in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 4284–4291. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/692_Paper.pdf.

[13] E. Sumita, F. Sugaya, S. Yamamoto, Measuring non-native speakers' proficiency of english by using a test with automatically-generated fill-in-the-blank questions (2005). doi:10.3115/1609829.1609839.

[14] J. Brown, G. A. Frishkoff, M. Eskénazi, Automatic question generation for vocabulary assessment, in: HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada, The Association for Computational Linguistics, 2005, pp. 819–826. URL: https://aclanthology.org/H05-1103/.

[15] S. Smith, A. P.V.S, A. Kilgarriff, Gap-fill tests for language learners: Corpus-driven item generation, 2010. URL: https://api.semanticscholar.org/CorpusID:61531901.

[16] M. Majumder, S. K. Saha, A system for generating multiple choice questions: With a novel approach for sentence selection, in: H. Chen, Y. Tseng, Y. Matsumoto, L. Wong (Eds.), Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications, NLP-TEA@ACL/IJCNLP, Beijing, China, July 31, 2015, Association for Computational Linguistics, 2015, pp. 64–72. URL: https://doi.org/10.18653/v1/W15-4410. doi:10.18653/V1/W15-4410.

[17] M. Majumder, S. K. Saha, A system for generating multiple choice questions: With a novel approach for sentence selection, in: H. Chen, Y. Tseng, Y. Matsumoto, L. Wong (Eds.), Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications, NLP-TEA@ACL/IJCNLP, Beijing, China, July 31, 2015, Association for Computational Linguistics, 2015, pp. 64–72. URL: https://doi.org/10.18653/v1/W15-4410. doi:10.18653/V1/W15-4410.

[18] T. Goto, T. Kojiri, T. Watanabe, T. Iwata, T. Yamada, Automatic generation system of multiple-choice cloze questions and its evaluation, Knowledge Management & E-Learning: An International Journal 2 (2010) 210–224. URL: https://api.semanticscholar.org/CorpusID:15482954.

[19] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: C. E. Brodley, A. P. Danyluk (Eds.), Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001, Morgan Kaufmann,

2001, pp. 282–289.

[20] S. K. Bitew, J. Deleu, A. S. Doğruöz, C. Develder, T. Demeester, Learning from partially annotated data: Example-aware creation of gap-filling exercises for language learning, in: E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, T. Zesch (Eds.), Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2023, Toronto, Canada, 13 July 2023, Association for Computational Linguistics, 2023, pp. 598–609. URL: https://doi.org/10.18653/v1/2023.bea-1.51. doi:10.18653/V1/2023.BEA-1.51.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[22] S. Matsumori, K. Okuoka, R. Shibata, M. Inoue, Y. Fukuchi, M. Imai, Mask and cloze: Automatic open cloze question generation using a masked language model, IEEE Access 11 (2023) 9835–9850. URL: http://dx.doi.org/10.1109/ACCESS.2023.3239005. doi:10.1109/access.2023.3239005.

[23] P. Chomphooyod, A. Suchato, N. Tuaycharoen, P. Punyabukkana, English grammar multiple-choice question generation using text-to-text transfer transformer, Comput. Educ. Artif. Intell. 5 (2023) 100158. URL: https://doi.org/10.1016/j.caeai.2023.100158. doi:10.1016/J.CAEAI.2023.100158.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. URL: https://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[25] V. Slavuj, L. Nacinovic Prskalo, M. Brkic Bakaric, Automatic generation of language exercises based on a universal methodology: An analysis of possibilities, Bulletin of the Transilvania University of Brasov. Series IV: Philology and Cultural Studies 14 (63) (2022) 29–48. doi:10.31926/but.pcs.2021.63.14.2.3.

[26] A. Malafeev, Language exercise generation, International Journal of Conceptual Structures and Smart Applications 2 (2014) 20–35. doi:10.4018/IJCSSA.2014070102.

[27] D. Perrett, D. March, An evidence-based approach to distractor generation in multiple-choice language tests, 2019. doi:10.13140/RG.2.2.22779.16165.

[28] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[29] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[30] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909.

[31] Meta, Introducing Meta Llama 3: The most capable openly available LLM to date, https://ai.meta.com/blog/meta-llama-3/, April 2024.

[32] K. Tirumala, D. Simig, A. Aghajanyan, A. Morcos, D4: Improving llm pretraining via document de-duplication and diversification, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 53983–53995. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/a8f8cbd7f7a5fb2c837e578c75e5b615-Paper-Datasets_and_Benchmarks.pdf.

[33] O. et al., Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[34] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, 2023. URL: https://arxiv.org/abs/2305.14314. arXiv:2305.14314.

[35] Z. Fu, W. Lam, A. M.-C. So, B. Shi, A theoretical analysis of the repetition problem in text generation, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 12848–12856. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17520. doi:10.1609/aaai.v35i14.17520.

[36] Z. Xu, S. Jain, M. Kankanhalli, Hallucination is inevitable: An innate limitation of large language models, 2024. URL: https://arxiv.org/abs/2401.11817. arXiv:2401.11817.

A. Error analysis

Thanks to the human evaluation we conducted a small error analysis on the errors made by the model. By analyzing the exercises that the annotator marked as incorrect we found out that the major issue is the coherence of the exercise sentence. More precisely, 75% of the wrong exercises has a meaningless or absurd exercise sentence. This behaviour is directly related to the hallucinations suffered by LLMs[36]. The second prevailing error is the ambiguity between the key and the distractors. The model does not possess a deep understanding of what a distractor is. In fact some generated distractors are interchangeable with the key.

Despite these limitations, the model is very effective in producing exercises that are not trivial (plausibility error rate at 1%) and negligibly affected by bias and stereotypes.

grammar topic	CS	Acc	Amb	P
articles	1.00	-	-	-
comparison adjectives	0.64	0.36	-	-
conditional statements	1.00	-	-	-
future simple	1.00	-	-	-
modal verbs	0.85	-	0.15	-
infinitive and gerund verbs	0.50	0.12	0.38	-
passive tenses	0.83	-	0.17	-
past continuous	0.60	-	0.40	-
past perfect	0.50	-	0.38	0.12
past simple	0.40	-	0.60	-
personal pronouns	0.56	0.11	0.33	-
possessive adjectives	1.00	-	-	-
prepositions	0.67	-	0.33	-
present continuous	0.50	-	0.50	-
present perfect	1.00	-	-	-
present simple	1.00	-	-	-
quantifiers	0.75	0.25	-	-
relative clauses	0.64	0.18	0.18	-
WH- question	0.80	-	0.20	-
average	**0.75**	**0.05**	**0.19**	**0.01**

Table 2
Results of the error analysis on the wrong exercises. **CS** stands for Common Sense, **Acc** for Acceptability, **Amb** for Ambiguity and **P** for Plausibility.

B. Prompts

In this section, the prompts used in our work are presented. utilize the Llama3 chat template format, but to make the text more readable we use three placeholders: `#SYSTEM`, `#USER` and `#ASSISTANT`.

B.1. Fine-Tuning prompt

The prompt used to fine-tune the model has the same structure for all the grammar topics. The only varying parts are the name of the grammar topic and the number of distractors required. These parts are highlighted by the brackets and change depending on the dataset items. The prompt used is the following.

```
Write a multiple-choice gap exercise on {grammar_topic} with {n_distractors}
    distractors.
```

Listing 1: Fine-tuning prompt.

A training example is created by concatenating to the prompt the desired JSON representation of the exercise. We decided to use this format because it is easier to use at inference time. An example of training data is the following.

```
# USER
Write a multiple-choice gap exercise on comparisons with 3 distractors.
# ASSISTANT
{
 "filled_text": "Thanks to high technology, doctors can better assess
     patients' conditions.",
 "gapped_text": "Thanks to high technology, doctors can ___ assess patients'
     conditions.",
 "solution": "better",
 "distractors": ["best", "good", "well"]
```

Listing 2: Example from the Fine-Tuning dataset.

B.2. Baseline prompt

To test the performances of the baseline Llama3 we utilize its instruction-tuned version, Llama3-Instruct that can follow direction given by the user. This model is not able to answer correctly using the prompt described above. Therefore, we construct an alternative one in which all the useful information is given to the model. We include the structure of the exercise, the roles of each component with their constraints and the desired format of the output. The results are the following.

```
# SYSTEM
You are an english teacher creating multiple-choice-gap exercises.
# USER
Write one exercise on {grammar_topic}.
It must contains the:
 - sentence: the body exercise text that must contain the tag <GAP> instead
     of the solution
 - solution: the that correctly fill the gap
 - distractor: a word related to the solution, but different
The distractor must be such that if substituted to the solution, the sentence
     is wrong.
Do not generate any exaplanation.
The output must be a JSON object with the following structure:
{"sentence": str, "solution": str, "distractor": list[str]}
```

Listing 3: Prompt used to the generation of exercises with the base Llama3 model.

C. Ethical Considerations

This section outlines the ethical considerations of the system we developed.

Bias and Fairness The dataset used in this study is obtained from a publicly available source, ensuring that all data was collected with appropriate consent. To protect personal information, we removed all sensitive data such as phone numbers, email addresses and URLs. Since humans created this data, we assume that proper names or any reference to existing entities are invented. Moreover, those that contain preferences such as films, books, etc., we assume do not reflect real preferences of the users. We suppose that events or situations described in the exercises are not related to existing facts. Finally, since the data have been created by professional creators we assume that any possible bias or stereotype in the dataset is not intended and it is a coincidence.

Accuracy and Reliability The accuracy of the generated exercises is paramount. We employ both automated validation tools and human expert reviews to ensure the correctness and reliability of the content. Any inaccuracies identified are promptly rectified. We acknowledge the potential for bias in LLM-generated content. However, the human evaluation highlights a negligible presence in the generated outputs.

Transparency We strive for transparency by documenting the sources of our training data and explaining the model architecture. All the techniques used to manipulate the data and the steps done are described step by step highlighting all the important aspects.

Educational Impact We assess the impact of LLM-generated exercises on learning outcomes. We aim to enhance personalized learning while preventing over-reliance on automated systems. The content is designed to be inclusive and accessible to all students.