Automatic Generation of School Activities for Teaching English to Children

Lorenna Marinho Lucena[1], **Matheus Lisboa Oliveira dos Santos**[1], **Claudio E. C. Campelo**[1]

[1]Systems and Computing Department – Federal University of Campina Grande (UFCG)
58429-900 – Campina Grande – PB – Brazil

lorenna.lucena@ccc.ufcg.edu.br, matheus.lisboa@copin.ufcg.edu.br

campelo@dsc.ufcg.edu.br

Abstract. *Activity sheets are essential tools for developing basic skills, integrating theoretical and practical knowledge. For children, they should be playful, creative and interactive, facilitating the association of ideas. Creating these sheets is a challenge, requiring technical and pedagogical skills, especially in language teaching. This article proposes the use of Large Language Models (LLMs) with fine-tuning to generate English as a Second Language (ESL) exercises for children. Different models were evaluated and compared against fine-tuned GPT-3.5. The approach was evaluated with the help of ESL teaching experts, based on criteria such as naturalness, usefulness, diversity and personalization. The results indicated that the fine-tuning approach is promising.*

1. Introduction

Activity sheets have been shown to be essential in the learning process, allowing students to practice, apply and generalize the content they have learned, as well as making it possible to assess students. Studies in this area highlight this relevance, such as the research by [Karpicke and Roediger III 2008], which demonstrates that active retrieval, such as tests and recall exercises, is more effective for long-term knowledge retention than passive methods, such as re-studying through simple reading, reinforcing that school exercises and strategies that require active retrieval of knowledge not only assess, but also enhance learning.

In the learning process for children, activity sheets need to have a more playful content, such as stories, characters and attractive images with simple colors and symbols that arouse the child's interest, in order to capture their focus on the activity. The Double Coding Theory [Paivio 2013] reinforces this idea, pointing out that the combination of visual and verbal stimuli enhances learning, since multiple sensory channels are activated simultaneously in the brain. In light of this, the importance of well-planned activities that integrate elements that facilitate learning from children's first contact with formal education is evident.

When learning English as a Second Language (ESL), it is essential to establish a clear correspondence between the words of the new language and the vocabulary of the student's mother tongue, in order to facilitate the understanding of meanings. In this sense, the use of visual examples, such as images and drawings, is essential to connect and contextualize the content to children's daily lives. In addition, studies show that

DOI: 10.5753/sbie.2025.12773

the exposure of words in various contexts, immediate and concrete, allows children to learn new terms intuitively, without relying on translation, i.e. understanding precedes naming; first the child identifies the object or action through visual aids or situations and only then internalizes the word that represents it [Krashen 1982], which also highlights the importance of designing appropriate activity sheets with elements that help in this process.

1.1. Problem

As pointed out in [da Silva and do Rosário Sabota Silva 2019], traditional textbooks often present content in a rigid way, with rigid structures and a lack of stimulus for communication and interaction between students. In view of this limitation, it is proposed that complementary activities be developed that integrate cross-cutting content. To this end, the active role of the teacher in adapting and creating resources in line with teaching perspectives that value diversity and student autonomy is essential.

However, designing tasks based on appropriate approaches according to the students' different levels of proficiency, building creative layout structures and searching for images and drawings is a job that requires time and a lot of effort to develop different activities that deal with the same content in a way that ensures that the student is exposed to repetition of concepts. In addition, it requires computer skills such as information search and retrieval strategies to find images, the use of photoshop technologies and drawing tools which, in the end, do not guarantee a good result and which take up many hours. Education professionals often have to resort to cutting and pasting to build their activity sheets, which end up not being what they envisioned, due to time constraints and the challenges of the teaching routine filled with commitments and class hours.

The expansion of research into, and use of Large Language Models (LLMs) has led to the proposal of various approaches based on question generation. However, most of them do not include activities aimed at children that contain images, drawings, colored elements and symbols organized in creative layouts. Although these models can create content quickly and diversely, they do not fully meet the needs of developing educational activities aimed at children. All data and code used in this study are publicly available[1].

1.2. Proposed Approach

This study evaluates the potential of LLMs in creating ESL activity sheets for literate children in early elementary school and proposes a model to generate well-structured and pedagogically relevant worksheets. For the training and evaluation of the model, we produced a dataset of manually crafted activity sheets based on examples extracted from textbooks. Finally, an evaluation was conducted by professionals in the field to assess the usefulness of the generated sheets.

2. Related Work

Several studies have explored the application of generative models and Natural Language Processing (NLP) techniques for the development of automated question generation and personalized teaching systems. [Rao et al. 2022] propose a model capable of generating three types of questions: complex, multiple-choice, and gap-filling. The model generates

[1]https://github.com/lorennavictoria/llms-question-generator

question-answer pairs from input texts, saving time and effort compared to traditional methods. The questions produced are grammatically correct and can be used in educational contexts, such as quizzes for practicing specific content. The system also facilitates automatic correction, allowing anyone with the answer sheet to check the solutions.

In another study, [Sonderegger 2022] discusses the integration of generative models, such as GPT-3, into social robots to improve interactive learning. The author pointed out that systems based on rules or information retrieval are limited, while generative models allow for more fluid and personalized responses. A robotic tutor was developed using GPT-3 within the ICAP framework. The model explains a topic, asks a theoretical question, evaluates and corrects the answer and simulates a dialog. Practical application in modules such as explanation, quiz and dialogue demonstrated GPT-3's ability to generate dynamic and contextualized content.

In the context of language learning, [Kwon 2023] proposes two approaches based on generative language models. The first, GPTChat, is a non-restricted chatbot that generates examples of vocabulary and phrases based on free queries from users, using few-shot prompting techniques to improve the relevance of responses. However, evaluations have revealed challenges, such as the generation of uneven or out-of-context content. The second approach, GPTutor, overcomes these limitations by structuring the interaction in three stages: the definition of interlocutors, the user profile, and the topic of conversation. In this way, it generates contextualized dialogues that have proven to be more relevant and suitable for beginners compared to manual databases.

In [Kriangchaivech and Wangperawong 2019], a model based on transformers is proposed to generate questions from the content available on Wikipedia. The model uses attention to produce questions that are grammatically correct and relevant to the context being worked on, although the questions generated are simpler compared to those created by humans.

In our approach, LLMs are used to generate activity sheets only designed to teach ESL to children. This is done from a data set made up of school activities structured in HTML and CSS, which includes auxiliary textual descriptions, well-designed prompts, and the application of fine-tuning in the model to ensure the generation of coherent exercises.

This work addresses a critical gap in the literature by integrating key dimensions such as structure, visual organization, and personalization, which are often overlooked by other approaches. Unlike existing solutions, we focus on the automatic generation of worksheets specifically for children learning ESL.

3. Methodology

3.1. Structure of the activities

Given that the target audience is children, most of the sheets contain images, blocks and colors to facilitate assimilation of the content. In this sense, instead of generating an image for each activity, a detailed textual description of the visual elements is produced. So, instead of generating complete images, descriptions of the graphic elements are created, which can then be used by text-to-image models. In this way, an organizational skeleton is built for activity sheets, ensuring that exercises can be easily adjusted or expanded.

This approach does not compromise the quality of the activities, since using the textual description allows us to reorganize the elements as necessary, since they are purely text, and in this way we can use a single model to generate both the structure and the description of the images. In addition, we used a format for representing activity sheets based on the chain-of-thought technique [Li et al. 2024], which helps the model to better understand and structure them. This approach takes the form of organizing the sheets into files structured into three main parts:

1. **Generation Parameters:** A set of attribute-value pairs that control the creation of sheets. These parameters allow the user to specify details such as the school year, the topic of the activity, the task to be carried out by the student, the learning objectives and the width of the sheet. This control makes it possible to generate personalized activities for different student profiles.
2. **Textual Explanation:** Each activity contains a detailed explanation in JSON format, structured in three main attributes:
    - **Task:** A description of what the student will learn and the actions required to carry out the activity. This field improves understanding of the model and adds explainability to the generation process.
    - **Layout:** A detailed textual description of the organization of the elements in the activity. This allows the model to understand how the components should be distributed on the sheet.
    - **Answer:** A detailed step-by-step with the expected answer. Studies show that LLMs respond more accurately when instructions are presented in a sequential and enumerated manner [Zhou et al. 2022]. Therefore, the expected response was structured in this way, reinforcing the chain-of-thought approach and contributing to better interpretation of the model.
3. **The structure of the sheet under development:** To ensure the organization and flexibility of the data set, each activity sheet was structured using HTML and CSS, allowing the generation of an adaptable layout. In addition, this approach makes it easier to separate the visual structure from the textual content of the activities, allowing for dynamic adjustments without compromising their integrity.

3.2. Dataset

The dataset was structured to cover activities from the second to the fifth year of elementary school, excluding the first year. This decision was made because, at this early stage of schooling, children are still in the process of becoming literate, making the use of playful activities such as games and storytelling more relevant than written exercises requiring reading, interpretation and text production. Within the years selected, the activity sheets have been designed so that the model can generate activities that help teach verbs, verb tenses, and vocabularies. In addition, the exercises have been designed to encourage the development of different cognitive skills, such as observing images, writing detailed answers, connecting items, and organizing words in a sentence, while maintaining the development and learning objectives of this age group.

The sheets were based on and adapted from Primary 1 textbooks [Passos and Silva 2014], ensuring alignment with the educational guidelines and learning objectives established [Council 2022]. In addition, new sheets were developed

with the support of English teaching experts, ensuring that the exercises were pedagogically appropriate for the age group. The process of constructing the dataset as a whole involved research, adaptations, and the manual formulation of tasks, resulting in a total of 60 sheets making up the dataset. Although this number may seem small, it proved to be sufficient, since LLMs have a high capacity for generalization, making it possible to identify patterns and generate new activities based on a small number of examples. The activity sheets were structured to include multiple images, descriptions containing eye-catching colors, and different formats and blocks that organize the information in a clear and intuitive way, making them more engaging and in line with active teaching methodologies.

The consultancy process was carried out to ensure that the vocabulary was age-appropriate and aligned with the educational objectives. The words were selected based on the children's familiarity, avoiding terms that, despite being written similarly in English and Portuguese, are not part of the children's daily lives, such as the word ravioli, as well as content considered inappropriate, such as whisky, wine, beer, casino, and poker. Instead, the activities prioritized high-frequency, relatable words such as apple, pencil, dog, and cat to reinforce vocabulary retention through real-world connections. This approach not only supported effective language learning, but also ensured that the content remained engaging, relevant and appropriate for the young learners. This selection was essential to ensure that the activities helped students make direct connections between the new vocabulary and its use in everyday life, which is a significant contribution to the teaching process.

There was no dataset that met the specific structural requirements for proper model fitting, to enable the automated generation of school activities within the proposed approach. It was therefore necessary to build a suitable dataset for this purpose. The strategy adopted was to automate the creation of questions using the language models themselves, considering that studies show that LLMs tend to perform better when trained with data generated by themselves [Zelikman et al. 2022].

For the initial construction of the dataset, GPT 4o was used without prior adjustments. The model received detailed prompts containing specific instructions on the task, the structure of the sheets, and illustrative examples to guide generation. At first, some examples of manually-generated sheets were provided, and then the model was instructed to infer new sheets from reference images, maintaining structural and pedagogical coherence with the examples provided. Each sheet generated went through a manual validation process to ensure that the content was appropriate, that the expected answers were correct and that the educational objectives were faithful.

The results of this sheet generation process were promising, since the model managed to structure the skeleton of the sheet properly. However, some adjustments were necessary to improve the organization and ensure the usability of the data in subsequent training. Among the main corrections was the need to remove interactive codes, which were generated even with explicit instructions to avoid them. In addition, it was found that the distribution of the worksheet elements did not always occupy the space on the sheet optimally. These modifications were essential to make the structure of the sheets more robust and improve the quality of the data set, providing better model results after the fine-tuning.

During the process of automatically generating the activities, specific limitations of the model were identified in the creation of activities of the type "connect items" and "associate words with images". The model failed to shuffle the alternatives, even when this instruction was explicitly included in the prompt - for example: "If the activity is about linking items, make sure that the items are shuffled in the columns." This limitation compromises the pedagogical validity of the activities, since the predictability of the answers reduces the cognitive challenge proposed to the student.

In addition, there was a need to improve the descriptions for activities, as initial model outputs lacked sufficient detail for effective text-to-image conversion. For instance, descriptions like "Tom playing ball"—lacking visual context (e.g., "a boy playing ball in an open field") — proved inadequate for generating meaningful images. Additionally, the wording of activities was refined to ensure clarity, explicitly guiding students on the tasks to be performed. These improvements aimed to make instructions more precise and actionable, addressing gaps in both visual and instructional coherence.

While constructing the activities and defining the generation parameters, a recurring impasse was identified between the fields of learning objective and theme, especially in the activities aimed at teaching vocabulary. In many cases, these two elements overlapped semantically, making them difficult to distinguish clearly. For example, in an activity aimed at teaching the names of fruit, both the theme and the learning objective could be defined as "fruit", which created redundancy in the parameterization and made it difficult to diversify the activities.

3.3. Prompt

For the automated generation of school activities, we followed a fine-tuning technique with the aim of guiding the model in the creation of coherent structures. The prompt can be divided into three parts:

- In the first part of the prompt, it is defined that the generation of the activity must include a structured textual description in JSON format, made up of the fundamental attributes: `task`, `layout` and `response` explained at the beginning of this section.
- In the second part, parameters such as school year, topic, learning objectives, sheet width and type of task to be completed by the student are set, ensuring that the generated activity is personalized.
- In the third part of the prompt, the instructions are directed at generating the structure of the activity using in HTML and CSS, with a focus on creating a clear, creative layout that is compatible with the content of the task.

In the third part, the model is instructed to organize all the elements of the activity in a completely non-interactive way, considering that the proposal is for the exercises to be solved using only pen and paper. The prompt also establishes specific rules, such as the use of the `alt` attribute on images to visually describe the content presented, ensuring accessibility and contextualization, and the inclusion of an additional attribute called `exp` within the tags in the HTML, designed to explain the pedagogical function of each component of the activity. It is also requested that explanatory comments be added throughout the activity sheet generated in HTML and CSS, in order to justify the layout decisions adopted. Although the model was trained with prompts in English, it is multilingual and can accept prompts in other languages.

3.4. Evaluation metrics

One of the difficulties encountered in the literature is the dependence on manual evaluation, due to the machine's difficulty in identifying a good activity sheet, since automatic metrics cannot fully capture the quality, relevance and pedagogical suitability of the sheets generated [Neo and Neo 2024]. In this sense, to evaluate the school activities generated, we use criteria identified in the literature, proposed by Faraby [Al Faraby et al. 2024], which measure the pedagogical value of the activities. The criteria adopted are as follows:

- **Naturalness**: assesses whether the activity is written in a fluid and comprehensible way, whether it can be easily understood by the student, whether it is adequately aligned with the proposed context and also considers whether the activity can be completed.
- **Utility**: analyzes whether the activity effectively contributes to learning, addressing information that is relevant and central to the topic being worked on. The activity should promote the acquisition of knowledge.
- **Diversity and Controllability**: this criterion considers both the variety of activities that the model is capable of generating (diversity) and its ability to adjust the type and complexity of activities according to defined parameters (controllability).
- **Personalization**: refers to the model's ability to adapt activities to the specific needs of each student, taking into account their current level of knowledge and individual preferences.

The evaluation was carried out by a control group composed of five professionals in the field of teaching ESL, who have been previously instructed on the meaning of each criterion. For the analysis, a Likert Scale was used, with scores from 1 to 5, where 1 represents "totally disagree" and 5 represents "totally agree". This approach aims to quantify, in an organized way, the quality and applicability of the activity generated.

4. Results and discussions

All inferences were made with the same base prompt, described in Section 4.3, and the same parameters used to customize the activity. In this scenario, 4th grade students were asked to complete an activity focused on teaching verbs in Past Tense, with the topic "Sports" and requiring the student to write the correct answer. The results show considerable variations in both the structure of the activities and their pedagogical suitability.

GPT 3.5 with fine-tune showed the best performance among the other models evaluated, generating a complete gap-filling activity with verbs in the past tense and clear instructions, as well as good visual distribution. The descriptions of the images were well contextualized with the theme, and the verbs provided ("swam", "played", "rode", "did") were correctly conjugated, as shown in Figure 1. The model showed a good understanding of the parameters passed, as it generated a coherent activity.

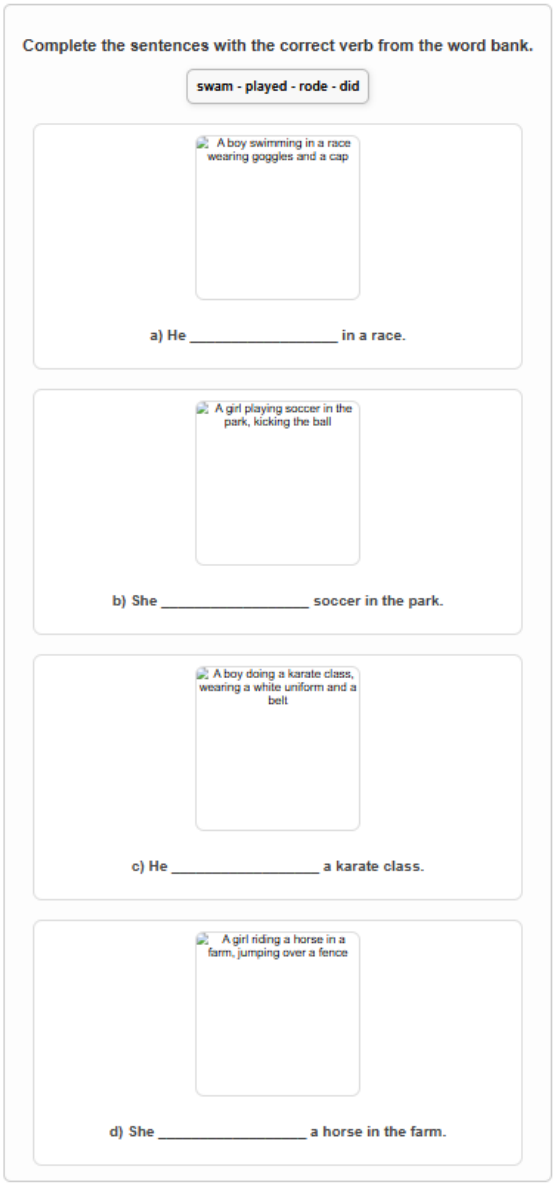

Figure 1. Inference result with GPT 3.5 with fine tuning

In addition to the main model evaluated, other language models were also considered for comparison, including GPT-4, DeepSeek-V2 (14B), Llama 3 (8B) and GPT-3.5 itself. It is important to note that none of these models underwent a specific fine-tuning process for the task in question, and were evaluated in their standard form (out-of-the-box).

The comparison between model size and performance revealed that a greater number of parameters does not always guarantee better results. Although GPT-3.5 is estimated as a smaller model than DeepSeek and GPT-4, it outperformed both by generating a complete, coherent and well-structured activity, aligned with the given context, when adjusted. However, GPT-3.5 without tuning performed the worst, demonstrating that tuning plays a more crucial role in the quality of results than model size alone. The evaluation criteria are shown in Table 1 and the summary of the results can be seen in Table 2.

Table 1. Evaluation Criteria

Criteria	Description
1	Good organization of activity elements
2	Correct use of the requested verb tense (Past Tense).
3	Clear instructions on what the student needs to do.
4	Well-described images related to the content.
5	Compliance with the requested parameters.

Table 2. Model evaluation by criterion

Model	1	2	3	4	5
LLaMA 3 (8B)		X			
DeepSeek (14.8B)	X	X	X		
GPT-4	X	X	X		
GPT-3.5 (without fine-tune)					
GPT-3.5 (fine-tune)	X	X	X	X	X

5. Evaluation of the quality of the activity sheets generated

The quality of the generated sheets was evaluated by English teachers using a 5-point Likert scale. The evaluation process was structured in a form, with the image of each worksheet accompanied by 8 evaluation questions. The evaluation questions contained in the form were:

- Q1 Does the activity adequately address the subject?
- Q2 Is the activity statement clear and understandable?
- Q3 Does the activity provide all the necessary information for students to be able to solve it correctly?
- Q4 Is the layout of the activity well-structured and easy to understand?
- Q5 Is the level of the activity suitable for elementary school students?
- Q6 Would you use this activity in the classroom if it included images with the characteristics described?
- Q7 Is the activity well contextualized within the theme specified?
- Q8 Is the activity creative and does it stimulate learning in a playful way?

A total of 16 sheets were evaluated, 8 of which were generated by the GPT 3.5 fine-tuned model and 8 of which came from textbooks or were prepared manually with the help of professionals in the field. Participants were not informed of the origin of the sheets, which were presented at random. To ensure a fair comparison, the sheets not generated by the model were adapted to follow an equivalent structure to the sheets produced by the fine-tuned GPT 3.5 (i.e., the original images were removed and replaced with corresponding textual descriptions).

The 16 sheets were distributed in two forms. Each form had 8 sheets, 4 of which came from the model and the other 4 did not. The process of allocating sheets was done

in such a way that each sheet had at least two evaluations from two different english teachers. Figure 2 shows the distribution of teacher evaluations for each of the eight evaluation questions in relation to the sheets generated by the model and those adapted from teaching materials. Each boxplot represents the distribution of all the scores given when that specific question was asked to the different evaluators, regardless of the sheet analyzed.

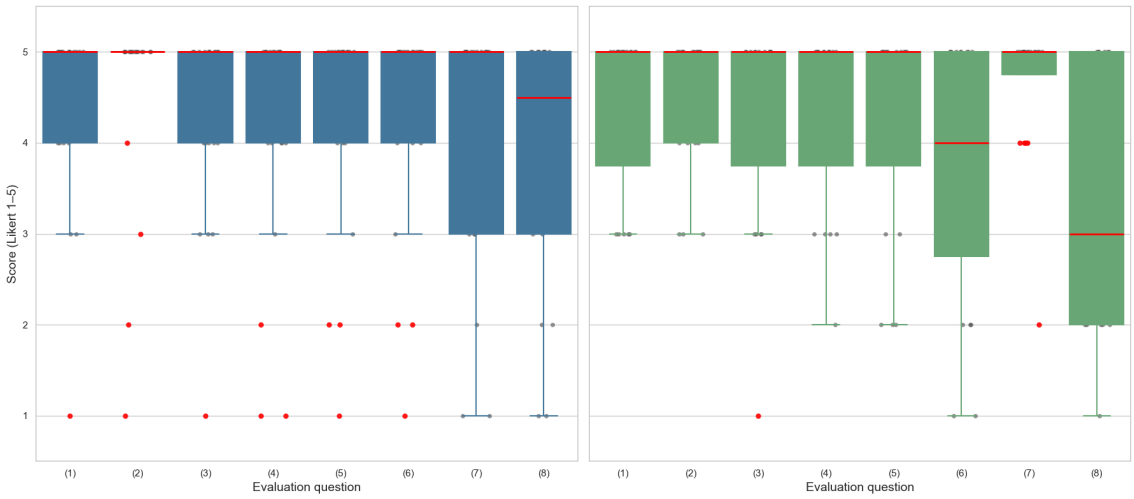

Figure 2. Distribution of human-assigned scores for the generated sheets across different evaluation questions. The left side shows the score distribution for sheets automatically generated, while the right side displays the scores for sheets produced manually by humans.

For sheets generated by the model (Figure 2, left side), there was a positive trend in most of the criteria evaluated, with medians situated at levels 4 and 5 on the Likert scale, which indicates that the generated sheets are generally well structured and meet the parameters requested. Moreover, it suggests that the model can produce activities for a variety of topics.

The mean and standard deviation for each of the sheets generated by the model are shown in Table 3. As it can be seen, the Q1, that evaluates the relevance of the sheet's content to the subject matter, was also rated as mostly positive, with a mean of 4.4 and a standard deviation of 1.05, suggesting that the model adequately understands and applies the topics requested when generating the sheets, meeting the personalization criterion.

Table 3. Statistics of answers to the evaluation questions for the activity sheets generated by the fine-tuned model.

Criteria	Number of ratings	Mean	Standard Deviation
Q1	20	4.40	1.046
Q2	20	4.50	1.147
Q3	20	4.25	1.070
Q4	20	4.10	1.334
Q5	20	4.25	1.251
Q6	20	4.25	1.251
Q7	20	4.10	1.410
Q8	20	3.90	1.410

In general, the results of the evaluations suggest that the proposed fine-tuned model is capable to produce sheets that satisfactorily meet the established criteria, especially in relation to clarity, adequacy of content, and structure. This indicates that the model has the potential to act as an auxiliary tool in the development of teaching materials.

With regard to the evaluations of the human-generated sheets, Figure 2 shows that the appropriateness of the subject matter (Q1) has a median of 5, indicating that most of the sheets appropriately address the proposed content. As for the assessment of necessary information, this shows greater variability, including some lower ratings, indicating that not all traditional activity sheets provide sufficient data for adequate resolution. The structuring of the layout, although with a high median, also shows considerable dispersion, demonstrating different levels of visual organization among the materials evaluated.

The variability observed in response to playfulness and classroom use may be the result of the masking done on the sheets to mix them up between the sheets generated by the model combined with the rigid nature of the teaching materials. This information for each of the sheets generated by the model is presented in Table 4. In this table, the second evaluation question had 19 evaluations because one of the participants left the question blank.

Table 4. Statistics of answers to the evaluation questions for the activity sheets not generated by the fine-tuned model.

Criteria	**Quantity of ratings**	**Average**	**Standard Deviation**
Q1	20	4.45	0.89
Q2	19	4.47	0.77
Q3	20	4.25	1.12
Q4	20	4.25	0.97
Q5	20	4.15	1.14
Q6	20	3.65	1.46
Q7	20	4.65	0.75
Q8	20	3.40	1.31

Questions 6 and 8 aim to evaluate creativity and playful stimulus. Figure 2 (left side) shows the presence of outliers for Q6 and elongated boxplot for Q8, indicating greater variability among evaluators's scores. However, in Figure 2 (right side), it is also possible to notice elongated boxplots for both Q6 and Q8, suggesting that even professionally prepared materials struggle to be rated homogeneously. This is reinforced by the standard deviations observed in Tables 3 and 4, for Q6 and Q8.

From the Tables 3 and 4, we can deepen the analysis. The sheets generated by the model have higher and more consistent averages compared to those produced by humans. In the Table 3, the averages range from 3.90 to 4.50, with only Q8 falling below 4.0. In contrast, Table 4 shows a wider range, from 3.40 to 4.65, with two items (Q6 and Q8) below 4.0. This suggests that the model tends to produce sheets with better general acceptance. The dispersion of responses, measured by the standard deviation, also shows a certain difference. In the sheets generated by the model, the standard deviations range from 1.046 to 1.410, indicating that the answers tend to be more concentrated around the average. For the sheets that were produced by professionals, the dispersion is wider, ranging from 0.75 to 1.46. As already identified, in both tables, Q8, which assesses whether the activity is creative and has a playful stimulus, stands out negatively, with low averages of 3.90 and 3.40 and high dispersion of 1.41 and 1.31. This indicates that this item, regardless of the origin of the activity sheet, presents problems to be worked on.

6. Conclusion and Future Work

Despite the promising results, there are still significant challenges to be overcome in the automatic generation of activity sheets. One of the main challenges is the spatial organization of the visual elements, with occurrences of poorly optimized layouts – such as insufficient space for images, overlapping and misaligned components. In addition, there are few effective mechanisms to guide generation according to specific pedagogical criteria. This resulted in sheets that did not have their elements arranged in an order that made them useful, such as activities about connecting items and associating words and images with their elements without being scrambled.

To overcome these problems, we intend to expand the database with a greater number and variety of questions for executing a longer fine-tuning. In addition, refining the approach to insert more explanations in the dataset would help increase the controllability of sheet generation, to make the activity more useful and natural. Finally, we plan to evaluate this approach on state-of-the-art open-source LLMs to make a model available to the public for use.

References

Al Faraby, S., Adiwijaya, A., and Romadhony, A. (2024). Review on neural question generation for education purposes. *International Journal of Artificial Intelligence in Education*, 34(3):1008–1045.

Council, B. (2022). Diretrizes para o ensino de inglês nos anos iniciais. Acessado em: 31 mar. 2025.

da Silva, M. D. R. and do Rosário Sabota Silva, B. (2019). Perspectivas criticas de ensino de linguas: Produção e adaptação de materiais didativos. V CEPE, Universidade Estadual de Goiás, CCSEH.

Karpicke, J. D. and Roediger III, H. L. (2008). The critical importance of retrieval for learning. *science*, 319(5865):966–968.

Krashen, S. (1982). Principles and practice in second language acquisition.

Kriangchaivech, K. and Wangperawong, A. (2019). Question generation by transformers. *arXiv preprint arXiv:1909.05017*.

Kwon, T. (2023). *Interfaces for personalized language learning with generative language models*. PhD thesis, Columbia University.

Li, Z., Liu, H., Zhou, D., and Ma, T. (2024). Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*, 1.

Neo, A. and Neo, G. (2024). Revisão sobre a geração automática de questões na educação: Técnicas, conjuntos de dados e métricas de avaliação.

Paivio, A. (2013). *Imagery and verbal processes*. Psychology Press.

Passos, C. and Silva, Z. (2014). *Eu gosto m@is: 2º, 3º, 4º e 5º anos do Ensino Fundamental*. IBEL, São Paulo, 2 edition. Coleção didática.

Rao, P. R., Jhawar, T. N., Kachave, Y. A., and Hirlekar, V. (2022). Generating qa from rule-based algorithms. In *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, pages 1697–1703. IEEE.

Sonderegger, S. (2022). How generative language models can enhance interactive learning with social robots. *International Association for Development of the Information Society*.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. (2022). Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.