From Arabic Text to Puzzles: LLM-Driven Development of Arabic Educational Crosswords

Kamyar Zeinalipour[1], Mohamed Zaky Saad[1], Marco Maggini[1], Marco Gori[1],

[1]University of Siena, DIISM, Via Roma 56, 53100 Siena, Italy

Correspondence: kamyar.zeinalipour2@unisi.it

Abstract

We present an Arabic crossword puzzle generator from a given text that utilizes advanced language models such as `GPT-4-Turbo`, `GPT-3.5-Turbo` and `Llama3-8B-Instruct`, specifically developed for educational purposes, this innovative generator leverages a meticulously compiled dataset named *Arabic-Clue-Instruct* with over 50,000 entries encompassing text, answers, clues, and categories. This dataset is intricately designed to aid in the generation of pertinent clues linked to specific texts and keywords within defined categories. This project addresses the scarcity of advanced educational tools tailored for the Arabic language, promoting enhanced language learning and cognitive development. By providing a culturally and linguistically relevant tool, our objective is to make learning more engaging and effective through gamification and interactivity. Integrating state-of-the-art artificial intelligence with contemporary learning methodologies, this tool can generate crossword puzzles from any given educational text, thereby facilitating an interactive and enjoyable learning experience. This tool not only advances educational paradigms but also sets a new standard in interactive and cognitive learning technologies. The model and dataset are publicly available. [1] [2] [3]

1 Introduction

Crossword puzzles, traditionally enjoyed for their challenge and entertainment value, are increasingly recognized for their educational potential. These puzzles facilitate learning in multiple disciplines, such as history, science, and linguistics, and are particularly effective in enhancing vocabulary and spelling skills (Orawiwatnakul, 2013; Bella and Rahayu, 2023; Dzulfikri, 2016). Their capacity to engage and educate simultaneously makes them invaluable in pedagogical contexts.

In language acquisition and the mastery of specialized terminology, crossword puzzles stand out as exceptional tools. They offer an interactive learning experience that promotes both the retention of technical jargon and general language skills (Nickerson, 1977; Yuriev et al., 2016; Sandiuc and Balagiu, 2020). This dynamic learning approach supports the cognitive development of learners by improving critical thinking abilities and memory retention (Kaynak et al., 2023; Dzulfikri, 2016; Mueller and Veinott, 2018; Zirawaga et al., 2017; Bella and Rahayu, 2023; Zamani et al., 2021; Dol, 2017).

The integration of advanced technologies, such as Natural Language Processing (NLP), further enhances the effectiveness of educational crossword puzzles. The advent of Large Language Models (LLMs) has particularly revolutionized the creation of Arabic educational crosswords, providing sophisticated, context-appropriate clues that significantly enrich the learning experience.

This paper introduces a novel application that leverages LLMs to produce tailored educational crossword puzzles from given text in a chosen category for Arabic learners. The application creates high-quality clues and answers, integrating user-provided texts or keywords through fine-tuning techniques. which can be utilized by educators to develop more engaging and effective instructional methodologies for learners.

Moreover, to support the development and assessment of this tool, a new dataset has been compiled named *Arabic-Clue-Instruct*, which will be disseminated to the scientific community. This dataset, accessible in an open-source format, consists of categorized texts in Arabic, each corresponding with clues and keywords in various learning categories. This initiative aims to streamline the creation of educational crosswords and facilitate their adoption in educational settings.

The structure of the paper is as follows. Section 2 provides a detailed review of the relevant literature. Section 3 contains the data set collection methodology and the curation process. The computational

[1] https://github.com/KamyarZeinalipour/Arabic-Text-to-Crosswords

[2] https://huggingface.co/Kamyar-zeinalipour/Llama3-8B-Ar-Text-to-Cross

[3] https://huggingface.co/datasets/Kamyar-zeinalipour/Arabic-Clue-Instruct

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 202–218
January 20, 2025.

methodologies used are also elucidated. Section 4 discusses the results of our experimental evaluations. Finally, Section 5 offers concluding remarks on the results and implications of our study. and Section 6 outlines the limitations of this study.

2 Related Works

The field of crossword puzzle generation has attracted significant academic interest due to its complex nature. Researchers have experimented with a wide array of approaches including the use of vast lexical databases and embracing the potential of modern Large Language Models (LLMs) (Zeinalipour et al., 2024a,d), which accommodate innovative methods like fine-tuning and zero/few-shot learning.

Pioneer works in this domain have been diverse and innovative. For example, Rigutini and his colleagues. successfully employed advanced natural language processing (NLP) techniques to automatically generate crossword puzzles directly from sources based on the Internet, marking a fundamental advancement in this field (Rigutini et al., 2008, 2012; Zeinalipour et al., 2024c). In addition, Ranaivo and his colleagues. devised an approach based on NLP that utilizes text analytics alongside graph theory to refine and pick clues from texts that are topic-specific (Ranaivo-Malançon et al., 2013). Meanwhile, aimed at catering to Spanish speakers, Esteche and his group developed puzzles utilizing electronic dictionaries and news articles for crafting clues (Esteche et al., 2017). On another front, Arora and his colleagues. introduced SEEKH, which merges statistical and linguistic analyzes to create crossword puzzles in various Indian languages, focusing on keyword identification to structure the puzzles (Arora and Kumar, 2019). Recent advances have included the efforts of Zeinalipour and his colleagues, who demonstrated how large-scale language models can be used to develop crossword puzzles in lesser-supported languages, including English, Arabic, Italian and Turkish showcasing the broad applicability of computational linguistics in generating puzzles that are both engaging and linguistically rich (Zeinalipour et al., 2023b,c,a, 2024c,b). In (Zugarini et al., 2024) they suggest a method for creating educational crossword clues and text datasets in English. In the Arabic version of generating crossword puzzles from text (Zeinalipour et al., 2023b), a few-shot learning approach was utilized, employing large language models (LLMs) as they are. However, in our current project, we have introduced a dataset specifically designed for this task in Arabic. Additionally, we have developed open-source models fine-tuned to better accommodate and enhance performance for this specific application.

However, the unique characteristics and challenges presented by the Arabic language have remained largely unexplored in the context of the generation of crossword puzzles from a given text. This study breaks new ground by explicitly employing cutting-edge language modeling techniques to create Arabic crossword puzzles from a given text. This approach not only fills a gap within the scholarly literature, but also enhances the tools available for language-based educational initiatives, thus advancing the field of Arabic crossword puzzle generation.

3 Methodology

This paper outlines the development of an innovative tool: an automated system to generate educational crossword puzzles from given text in Arabic, driven by Large Language Models (LLM). We introduce the *Arabic-Clue-Instruct* dataset, which comprises a curated selection of Arabic texts across several educational fields, including Chemistry, Biology, Geography, Philosophy, and History, among others.Our research is not confined to the specific category utilized in this study; it can be generalized to encompass a broader range of categories. This generalizability is attributed to the inherent capabilities of LLMs.This dataset served as the foundation for creating tailored crossword clues for each subject area.

An exhaustive iterative procedure enabled us to enrich the dataset, ensuring a diversified representation of Arabic educational material and specific crossword clues to correspond with each thematic area.

Our primary objective was to engineer Arabic crossword puzzles directly from textual content, employing the *Arabic-Clue-Instruct* dataset. We accomplished this by initially extracting the main context and the suitable keywords then generating human-evaluated clues using GPT-4-Turbo 3.1 then fine-tuning a suite of LLMs to adeptly generate accurate and pertinent crossword clues. The models employed in this undertaking include `Llama3-8B-Instruct` and `GPT3.5-Turbo` .

The subsequent sections of this paper will explore

the detailed techniques utilized in the creation of this dataset and the customization of LLMs. This development is expected to enhance learning experiences and outcomes in Arabic education. The comprehensive methods employed in this project are visualized in Figure 1.

3.1 *Arabic-Clue-Instruct*

Data Collection Methodology We initiated our data acquisition by extracting the initial sections of Arabic Wikipedia articles, specifically focusing on the prominently bolded keywords included in the Introduction section that correspond to the article's primary focus and other significant terms. In addition to this keyword-focused examination, we acquire essential metadata for each article, including metrics such as view counts, relevance assessments, condensed narrative overviews, key headlines, associative terms, categorization, and URLs.[4] This procedure benefits from the consistent structural format of Arabic Wikipedia, especially exploiting the information-dense introductory segments to methodically outline and extract the core ideas needed for a comprehensive data repository.

Data Refinement Techniques To enhance data quality through precision filtering, our strategy involves several critical steps. Initially, article selection is guided by measurements of popularity and relevance. We then discard articles that are either excessively long or significantly short in content by excluding those with a word count of less than 50 to ensure there is more room for clue generation. Furthermore, any associations with keywords that are more than two words long are removed from consideration to enhance the quality of the generated crossword puzzle. In the final step of our filtering process, we exclude keywords that either fall outside of the 3 to 20-character limit or include special characters and numerals. These filtering techniques are fundamental in maintaining the integrity and applicability of our dataset.

Development of the Prompting Mechanism. Creating a specific prompt was crucial for generating Arabic crossword clues from the given text using GPT-4-Turbo. These clues, which will be part of the main dataset for fine-tuning, relied heavily on Wikipedia articles to maintain topical relevance. This prompt was meticulously designed to facilitate the formulation of clues that were not only informative but also engaging, by weaving essential details and contextual background derived from the articles into the clues. The prompt was designed using the latest prompt engineering best practices and after trying different structures and prompting in Arabic and in English. The prompt utilized in this study is illustrated in Figure 2.

Formulation of Educational Arabic Clues. Influenced by the SELF-INSTRUCT framework (Wang et al., 2022), our approach leverages Large Language Models to automate the creation of educational crossword clues in Arabic from the given text. Our strategy is distinctive in its thorough integration of contextual information with the generated clues. For this purpose, we utilized `GPT-4-Turbo` an advanced version of LLMs known for its superior efficiency and capabilities. The culmination of our methodology involves employing carefully curated content, keywords, categories, and prompts as the foundational elements for crafting custom Arabic educational clues that meet our specific requirements.

Overview of the *Arabic-Clue-Instruct* Dataset We initiated our study by downloading approximately 211,000 articles from Arabic Wikipedia. This number was reduced to 14,497 suitable entries following a stringent content filtration process. These pages span 20 distinct thematic categories and form the basis of our curated dataset which features a mix of textual content and associated keywords.The data set was divided into 14,000 for training and 497 for testing.
To enrich the dataset further, we utilized the capabilities of `GPT-4-Turbo`, generating a diverse set of at least three clues per Wikipedia entry based on the length of the text. This effort resulted in a compilation of 54,196 unique clues, as detailed in Table 1.

A deeper analysis of the dataset reveals variability in context length, ranging from 75 to 1000 words, with most texts falling between 40 and 200 words. The clue-generation process typically yields clues between 20 and 30 characters in length. The character count for keywords is restricted to between 2 and 20 characters, and the majority of them are between 5 to 15 characters long. The distribution of text word counts, keyword occurrences, and clue character lengths is depicted in Figure 3. Figure 4 shows the distribution of data across various categories. Notably, the categories of "Ge-

[4] https://en.wikipedia.org/wiki/Wikipedia:Lists_of_popular_pages_by_WikiProject Wikipedia: Lists of popular pages by WikiProject

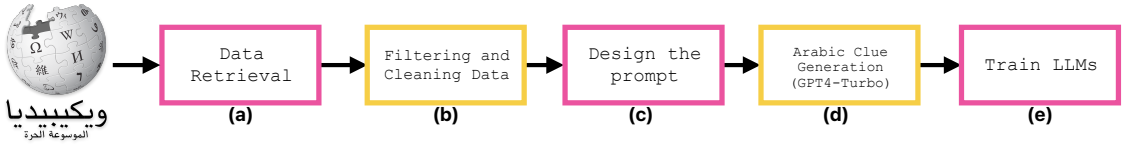

Figure 1: The methodology employed in this study is illustrated in this figure and includes the following steps: (a) Gathering data from Arabic Wikipedia. (b) Refining and filtering the data to enhance quality by eliminating content that is either too brief or excessively detailed. (c) Developing prompts for creating educational Arabic crossword clues derived from the educational content. (d) Employing `GPT4-Turbo` to generate Arabic crossword clues using the refined data and specifically crafted prompts. (e) Fine-tuning Large Language Models (LLMs) to more effectively produce Arabic clues tailored to the given context.

Arabic-Clue-Instruct		
Total Content-Keyword Pairs	Number of Distinct Categories	Total Generated Clues
14,497	20	54,196

Table 1: Overview of the *Arabic-Clue-Instruct* Dataset

```
You are an Arabic Crossword Puzzle Author
expert. your expertise involves taking any
paragraph and generate clues based on the
Keywords and the Category. your task is to
generate a clever and accurate clues which uses
wordplay, metaphors, puns, and other creative
techniques to make the clues engaging and
thought-provoking.This task is designed for
crossword puzzle authors looking to engage
their audience with entertainment clues.

based on a given keywords, a category and a
keyword-related context.
                    KEYWORDS: {keyword}
                    CATEGORY: {category}
                    CONTEXT: {text}
Follow these steps:
1. Analyze the context to identify the most
relevant and informative connections between
'{keyword}' and '{category}'. Extract key facts
from  the context, considering the length and
complexity.

2. Create {no} clues from these keyfacts in
Arabic (MAX 4 WORDS) that would be clue for the
{keyword}, making sure not to include the
keyword in the clues.

3. '{keyword}' should be the answer to the
clue. Avoid using the '{keyword}' or its
derivatives directly in the clues. Instead, use
descriptive phrases, synonyms, or creative
wordplay that hints at '{keyword}' without
revealing it outright.

4. Provide clues in the following structure:
Clues: <clue1> , <clue2> , <clue3> ...
```

Figure 2: Prompt used in the study.

ography", "Science", and "Society" dominate the dataset, whereas topics such as "Games", "Languages", and "Education" are comparatively underrepresented.

Evaluating quality of the *Arabic-Clue-Instruct* Dataset Producing accurate and engaging Arabic educational crossword clues is inhibited by the absence of a reference corpus, making it difficult to draw comparisons using standard measures, such as ROUGE scores. As typical metrics for comparison are not available, our evaluation strategy needed to adapt uniquely to the task requirements. Specifically, effective clues should represent contextually accurate paraphrases of text information. To accommodate this, we adopted an extractive method, using the ROUGE-L score to gauge the adequacy of clues in reflecting the input context. By comparing input sentences to the generated clues, the evaluation aimed to attain high scores to ensure strict adherence to the original text, minimizing irrelevant content and avoiding clues that merely replicate the input or improperly introduce the target keyword. Results indicated a substantial connection between the context and the clues, with an average ROUGE-L score of 0.0278, detailed results are provided in Table 2.

Considering that the ROUGE score merely compares the similarity between the n-grams of the generated clues and the reference text, it is not a reliable metric and does not provide any assessment of the semantic quality of the generated clues. However, it gives some thoughts about the generated clues.

In addition, the integrity of the generated clues was further examined through human evaluations conducted by experts in the Arabic language. A randomly selected subset of clues was assessed, consisting of 200 articles, each containing a maximum of three clues. This evaluation used a five-level criteria system, analogous to the methodology

Candidates	ROUGE-1	ROUGE-2	ROUGE-L
Text vs GPT4-Turbo clues	0.0281	0.0055	0.0278

Table 2: Mean ROUGE Scores for GPT4 vs Text

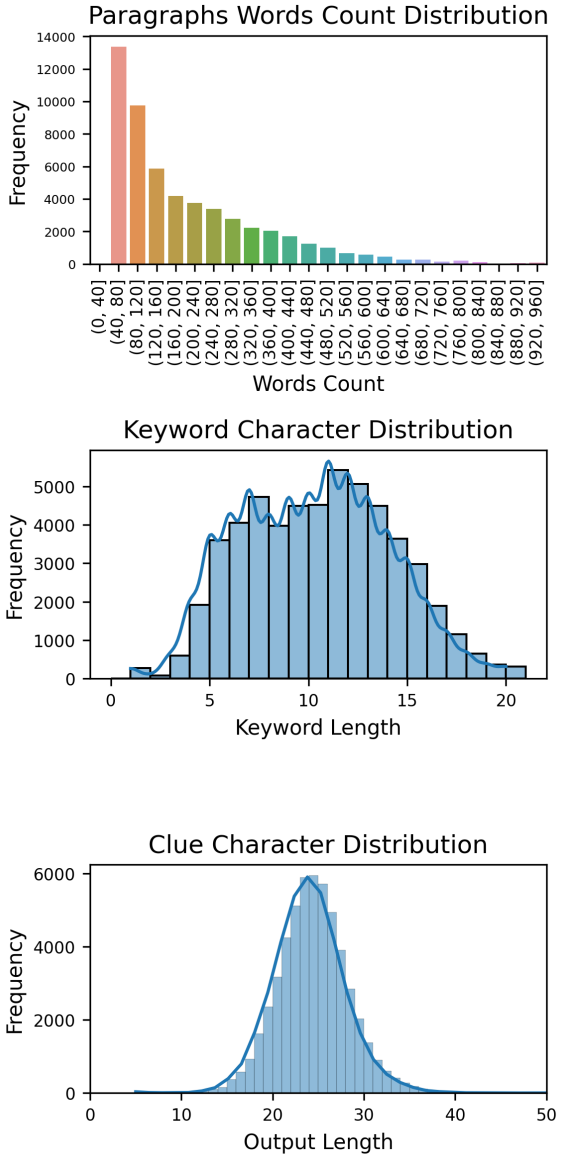

Figure 3: Word and Character Length Distributions for Contexts, Outputs, and Keywords.

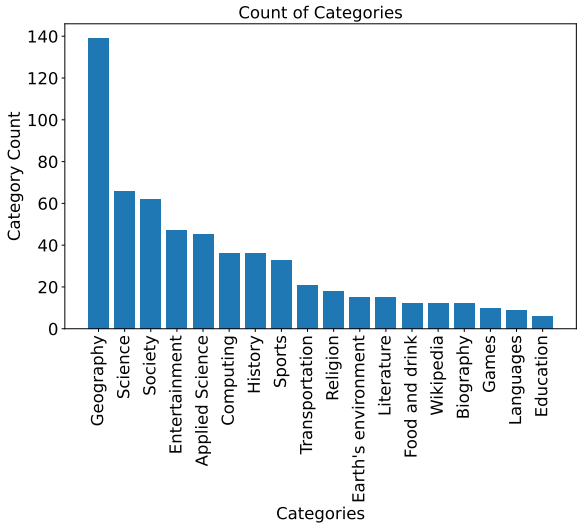

Figure 4: Bar Plot Showing the Frequency of Twenty Categories within the Dataset.

used by (Wang et al., 2022), which detailed the parameters as follows:

- `RATING-A`: The clue is coherent and valid, aligning correctly with the given context, answer, and specified category.
- `RATING-B`: This clue, while generally acceptable, exhibits slight discrepancies mainly characterized by a tenuous link to the category.
- `RATING-C`: Here, the clue relates directly to the answer but retains either a vague connection to the context or seems too broad.
- `RATING-D`: The clue fails by being irrelevant or incorrect in relation to the answer or the context.
- `RATING-E`: The clue is deemed unacceptable because it directly contains the answer or a variation of it.

The evaluation was made by a Native Arabic speaker who followed the criteria of evaluating based on the criteria mentioned above, As "Rating-A" was given to the good clues without any mistakes, "Rating-B" was given to the clues with minor mistakes but the clue is generally accepted, 'Rating-C' was given to the clue which has mistakes or the

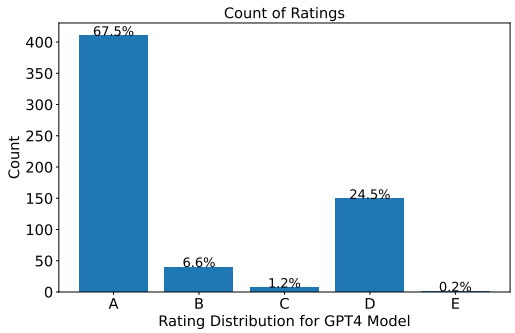

Figure 5: Bar Plot Showing the Frequency of GPT4 Ratings

connection isn't direct to the article but can still be understandable, "Rating-D" was given to clues that were entirely incorrect due to nonsensical words or irrelevance to the answer, and "Rating-E" to clues that contained the answer itself, making them unacceptable regardless of other qualities. The data and code used for evaluating the human annotation user interface (UI) are accessible on GitHub.[5]

The distribution of the evaluation outcomes is depicted in Figure 5 which also contains the percentage for each rating. These illustrate that the majority of the generated clues were of high quality with over 67.5% rated as 'A' and only a small fraction rated as 'C', 'D', or 'E'.
A qualitative analysis is available in Appendix B. AppendixA provides examples of different ratings for various clues generated by the `GPT-4-Turbo`.
By utilizing both quantitative metrics and qualitative assessments, the study aimed to robustly validate the educational utility and contextual accuracy of the clues created for Arabic educational crosswords.

3.2 Enhancing Arabic Educational Crossword Puzzle Generation Using LLMs

.

The study focus was the generation of crossword puzzle clues derived from Arabic texts, utilizing the advanced capabilities of Large Language Models (LLMs). The selected models for this task included `GPT-4-turbo`, `Llama3-8B-Instruct`, and `GPT-3.5-Turbo` , all of which are noted for their text generation skills and Arabic language support (Brown et al., 2020; Touvron et al., 2023). These models excel in decoding and formulating complex language structures, making them ideal for our objectives.
The process started with a crucial stage of model fine-tuning using the *Arabic-Clue-Instruct* dataset, rich in relevant content for our needs. This fine-tuning was critical to better equip the models to generate Arabic clues and to refine their ability to mirror the linguistic intricacies specific to the Arabic language used in educational contexts.
We applied meticulous parameter optimization to the fine-tuning phase to decrease task-specific errors. Our strategy had a two-fold purpose: to bolster the model's understanding of the Arabic educational material and to ensure an accurate representation of the language in the clues. The Arabic language presents unique challenges due to its complex grammar and vocabulary, making this an intensive task.
By tailoring these sophisticated LLMs with a dedicated dataset, we tried to enhance their capability to produce crossword clues from given Arabic text that is not just linguistically coherent but also contextually relevant to educational settings.

4 Experimental Results

In this section, we will detail the comprehensive experiments conducted in this study. Initially, we begin by describing the training setup utilized for training the LLMs using the *Arabic-Clue-Instruct*. This includes training parameters and the computational resources employed. Subsequently, we present the performance evaluations of the models, highlighting their effectiveness through automated metrics, with a particular focus on the ROUGE score. We will break down the results to show how the various configurations of the models compare and identify key areas where performance improvements are observed. Following this, we delve into an in-depth analysis of the human evaluations conducted to assess the models' performance. This analysis will include the methodology adopted for human evaluation, criteria for assessment, and qualitative feedback from human reviewers. We will provide insights into aspects such as relevance, coherence, and the overall quality of the generated content, which are often beyond the scope of automated metrics. Lastly, to illustrate the practical application of our proposed system, we provide an example of a crossword puzzle generated using the trained models. Through this detailed examination, we aim to present a holistic understanding of the experiments and their outcomes, demonstrating the

[5] https://github.com/KamyarZeinalipour/HumanAnnotation-UI-Ar-Text-To-Cross

robustness and versatility of the proposed system.

4.1 Training Setup

For the `GPT-3.5-Turbo`, the training regimen was meticulously designed, employing a batch size of 16 and a learning rate of 0.01 across three training epochs. Similarly, `Llama3-8B-Instruct` was fine-tuned using LORA (Hu et al., 2021) with parameters set to $r = 32$ and $\alpha = 64$ over the course of three training epochs, maintaining a total batch size of 128. The initial learning rate for `Llama3-8B-Instruct` was configured at 3×10^{-4}. For the inference phase, model distribution sampling was employed to generate clues for `Llama3-8B-Instruct`, with a temperature parameter set to 0.1. Additionally, the parameters for top-p and top-k sampling were adjusted to 0.95 and 50, respectively. The entirety of the experimental setup was conducted on a server equipped with dual NVIDIA A6000 GPUs, utilizing DeepSpeed (Rasley et al., 2020) and FlashAttention 2 (Dao, 2023) technologies.

4.2 Evaluation Results with the Automatic Metrics

We assessed the similarity between various sets of clues generated by different models, as presented in Table 3, and those generated by the `GPT-4-Turbo` model on a test set containing 200 educational contexts. This assessment was performed using ROUGE scores. The results reveal that the fine-tuned `Llama3-8b-Instruct` and `GPT-3.5-Turbo` model achieves a higher similarity to `GPT-4-Turbo`. In contrast,`Llama3-8B-Instruct` base model display significantly lower similarity exhibiting minimal overlap. These findings underscore the effectiveness of fine-tuning in aligning `GPT-3.5-Turbo` and `Llama3-8B-Instruct` using the *Arabic-Clue-Instruct* data set was increase the capability of models for generating the clues from Arabic educational text.

4.3 Evaluation Results with the human evaluator

Since the ROUGE score is not a reliable metric for assessing semantic quality, we conducted a human evaluation based on the judgment framework detailed in a previous section 3.1 , A human evaluation was conducted on both the generated and base models using a data set of 200 Arabic contexts, each containing 3 clues. The resulting ratings are illustrated in Figure 6 and summarized in Table 4, which shows the percentage distribution of each model's ratings. We employed the same 5-level rating system detailed in Section 3.

The ratio of each rating can be seen in Table 4. The provided table offers a comparative evaluation of language models' performance in generating Arabic clues from given text. Notably, both `GPT-3.5-Turbo` and `Llama3-8B-Instruct` models are assessed based on their base and fine-tuned configurations. Fine-tuning demonstrates a notable impact on model performance, as evidenced by the improvement of `GPT-3.5-Turbo FT` over its base counterpart. `Llama3-8B-Instruct` emerges as the top performer, achieving a remarkable 78.86% in category "A" from Originally being 36.02% in the Base Model which exceeds the `GPT-3.5-Turbo` in improving the performance after the fine tuning. The distribution of ratings showcases `Llama3-8B-Instruct` dominating category "A". These results underscore the efficacy of fine-tuning in enhancing model capabilities, particularly highlighted by `Llama3-8B-Instruct`'s performance with 8B parameters. Furthermore, after fine-tuning with the introduced dataset, the models' capability to generate Arabic clues from given text has significantly increased. This improvement also highlights the quality of the *Arabic-Clue-Instruct* dataset. A qualitative analysis is available in Appendix B. In Appendix A, you can find various examples of generated clue-answer pairs based on the provided text, along with their corresponding human ratings.

Generating Crossword Schema We investigated a methodology for generating Arabic crossword clues from categorized texts. This approach enables the creation of custom clues tailored to specific educational objectives. Following clue generation, educators can curate and select the most appropriate clues for constructing crosswords. An illustrative example of an Arabic crossword puzzle produced using this system is presented in Figure 7.

5 Conclusion

In this paper, we present a system for generating crossword clues from Arabic text. We created a dataset named *Arabic-Clue-Instruct* that includes text, keywords, categories, and related crossword clues in Arabic, making it the first of its kind in this context. Utilizing this dataset, we fine-tuned two large language models (LLMs), namely

Model	Model name	ROUGE-1	ROUGE-2	ROUGE-L
Base LLMs	GPT-3.5	0.0152	0.0011	0.0148
	LlaMa3-8b	0.0063	0.0009	0.0063
Fine-tuned LLMs	GPT-3.5	**0.0405**	**0.0045**	**0.0405**
	LlaMa3-8b	0.0354	0.0030	0.0354

Table 3: Mean ROUGE Scores for Various Comparisons with gpt4 clues

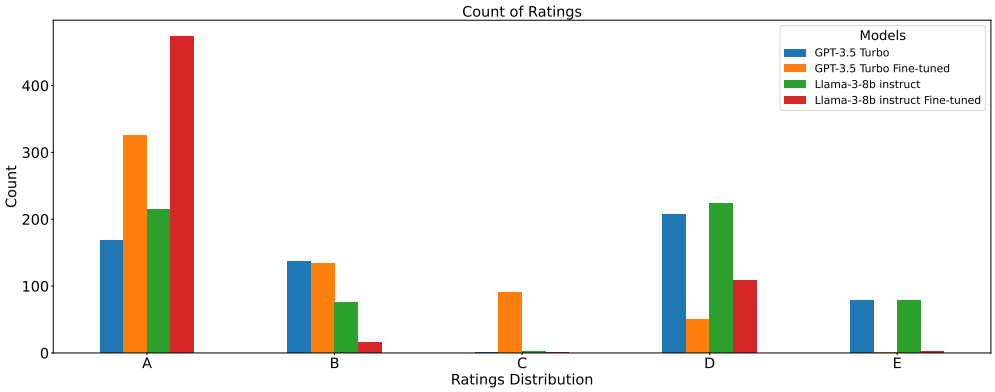

Figure 6: Bar Plot Displaying the Frequency of Ratings from Human Evaluation.

	GPT3.5-Turbo	GPT3.5-Turbo FT	Llama3-8B	Llama3-8B FT
# params	-	-	8B	8B
		Ratings %		
A	28.47	54.33	36.02	**78.86**
B	23.22	22.33	12.79	2.66
C	0	15.00	0.34	0
D	35.08	8.33	37.54	18.13
E	13.22	0	13.29	0.33

Table 4: Assessing the percentage of human evaluation for the clues generated by LLMs.

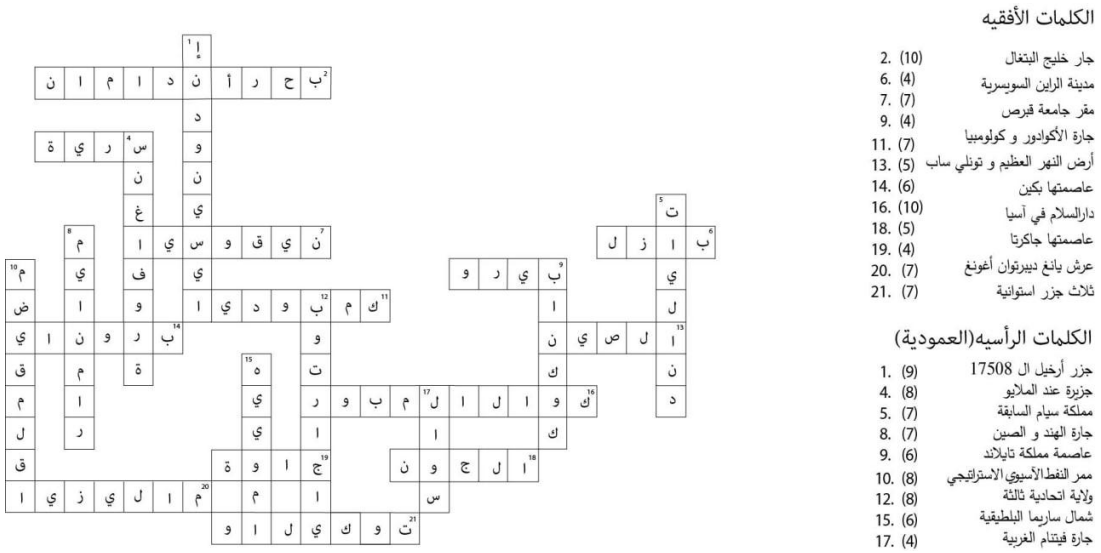

Figure 7: Crossword crafted using the proposed system.

`GPT-3.5-Turbo` and `Llama3-8B-Instruct`. Our results demonstrate a significant improvement in the models' capability to generate crossword clues from given text after fine-tuning. We have made the *Arabic-Clue-Instruct* dataset, along with the fine-tuned models, publicly available. These tools can be especially useful for students and teachers to generate educational crossword puzzles from Arabic text, as they can implement this crossword puzzle generator as a supplementary learning tool across various subjects by integrating the puzzles directly into lesson plans or as homework assignments. For future work, we aim to develop models capable of generating various types of crossword clues, including fill-in-the-blank crossword clues, and to expand our approach by adding more diverse datasets and experimenting with methodologies for languages with limited resources.

6 Limitations

The Arabic crossword puzzle generator, while innovative, has several limitations. The *Arabic-Clue-Instruct* dataset, despite its 50,000 entries, may not fully represent all Arabic dialects or recent language trends, potentially limiting its effectiveness across diverse Arabic-speaking regions.

Additionally, the tool's reliance on pre-defined categories might restrict its capacity to create puzzles for new or interdisciplinary topics, which could limit educators in customizing content for specific educational needs. Furthermore, varying levels of technological literacy among users might present challenges in using the tool effectively, potentially widening digital divides.

To address these limitations, future developments could focus on expanding and diversifying the dataset, enhancing the flexibility of puzzle generation, and ensuring the tool is accessible and user-friendly for a broad audience.

References

Bhavna Arora and NS Kumar. 2019. Automatic keyword extraction and crossword generation tool for indian languages: Seekh. In *2019 IEEE Tenth International Conference on Technology for Education (T4E)*, pages 272–273. IEEE.

Yolanda Dita Bella and Endang Mastuti Rahayu. 2023. The improving of the student's vocabulary achievement through crossword game in the new normal era. *Edunesia: Jurnal Ilmiah Pendidikan*, 4(2):830–842.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Sunita M Dol. 2017. Gpbl: An effective way to improve critical thinking and problem solving skills in engineering education. *J Engin Educ Trans*, 30(3):103–13.

Dzulfikri Dzulfikri. 2016. Application-based crossword puzzles: Players' perception and vocabulary retention. *Studies in English Language and Education*, 3(2):122–133.

Jennifer Esteche, Romina Romero, Luis Chiruzzo, and Aiala Rosá. 2017. Automatic definition extraction and crossword generation from spanish news text. *CLEI Electronic Journal*, 20(2).

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Serap Kaynak, Sibel Ergün, and Ayşe Karadaş. 2023. The effect of crossword puzzle activity used in distance education on nursing students' problem-solving and clinical decision-making skills: A comparative study. *Nurse Education in Practice*, 69:103618.

Shane T Mueller and Elizabeth S Veinott. 2018. Testing the effectiveness of crossword games on immediate and delayed memory for scientific vocabulary and concepts. In *CogSci*.

RS Nickerson. 1977. Crossword puzzles and lexical memory. In *Attention and performance VI*, pages 699–718. Routledge.

Wiwat Orawiwatnakul. 2013. Crossword puzzles as a learning tool for vocabulary development. *Electronic Journal of Research in Education Psychology*, 11(30):413–428.

Bali Ranaivo-Malançon, Terrin Lim, Jacey-Lynn Minoi, and Amelia Jati Robert Jupit. 2013. Automatic generation of fill-in clues and answers from raw texts for crosswords. In *2013 8th International Conference on Information Technology in Asia (CITA)*, pages 1–5. IEEE.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Leonardo Rigutini, Michelangelo Diligenti, Marco Maggini, and Marco Gori. 2008. A fully automatic crossword generator. In *2008 Seventh International Conference on Machine Learning and Applications*, pages 362–367. IEEE.

Leonardo Rigutini, Michelangelo Diligenti, Marco Maggini, and Marco Gori. 2012. Automatic generation of crossword puzzles. *International Journal on Artificial Intelligence Tools*, 21(03):1250014.

Corina Sandiuc and Alina Balagiu. 2020. The use of crossword puzzles as a strategy to teach maritime english vocabulary. *Scientific Bulletin" Mircea cel Batran" Naval Academy*, 23(1):236A–242.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Elizabeth Yuriev, Ben Capuano, and Jennifer L Short. 2016. Crossword puzzles for chemistry education: learning goals beyond vocabulary. *Chemistry education research and practice*, 17(3):532–554.

Peyman Zamani, Somayeh Biparva Haghighi, and Majid Ravanbakhsh. 2021. The use of crossword puzzles as an educational tool. *Journal of Advances in Medical Education & Professionalism*, 9(2):102.

Kamyar Zeinalipour, Achille Fusco, Asya Zanollo, Marco Maggini, and Marco Gori. 2024a. Harnessing llms for educational content-driven italian crossword generation. *https://arxiv.org/abs/2411.16936*.

Kamyar Zeinalipour, Achille Fusco, Asya Zanollo, Marco Maggini, and Marco Gori. 2024b. Harnessing llms for educational content-driven italian crossword generation. *arXiv preprint arXiv:2411.16936*.

Kamyar Zeinalipour, Tommaso Iaquinta, Giovanni Angelini, Leonardo Rigutini, Marco Maggini, and Marco Gori. 2023a. Building bridges of knowledge: Innovating education with automated crossword generation. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1228–1236. IEEE.

Kamyar Zeinalipour, Yusuf Gökberk Keptiğ, Marco Maggini, Leonardo Rigutini, and Marco Gori. 2024c. A turkish educational crossword puzzle generator. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 226–233.

Kamyar Zeinalipour, Yusuf Gökberk Keptiğ, Marco Maggini, and Marco Gori. 2024d. Automating turkish educational quiz generation using large language models. *https://arxiv.org/abs/2406.03397*.

Kamyar Zeinalipour, Mohamed Saad, Marco Maggini, and Marco Gori. 2023b. Arabicros: Ai-powered arabic crossword puzzle generation for educational applications. In *Proceedings of ArabicNLP 2023*, pages 288–301.

Kamyar Zeinalipour, Asya Zanollo, Giovanni Angelini, Leonardo Rigutini, Marco Maggini, Marco Gori, et al. 2023c. Italian crossword generator: Enhancing education through interactive word puzzles. *arXiv preprint arXiv:2311.15723*.

Victor Samuel Zirawaga, Adeleye Idowu Olusanya, and Tinovimbanashe Maduku. 2017. Gaming in education: Using games as a support tool to teach history. *Journal of Education and Practice*, 8(15):55–64.

Andrea Zugarini, Kamyar Zeinalipour, Surya Sai Kadali, Marco Maggini, Marco Gori, and Leonardo Rigutini. 2024. Clue-instruct: Text-based clue generation for educational crossword puzzles. *arXiv preprint arXiv:2404.06186*.

A Example of clues and answers generated from text using different models.

In this section, we present a selection of evaluated examples to illustrate the output generated by each model employed for clue generation, specifically: `GPT-4`, `GPT-3.5-Turbo`, `GPT-3.5-Turbo Fine-Tuned`, `Llama3-8B-Instruct`, and `Llama3-8B-Instruct Fine-Tuned`. The examples were carefully chosen to cover a range of ratings (A, B, C, D, and E) determined by expert evaluators in Arabic language proficiency. The examples were provided in 8 , 10 and 12, followed by their translation in 9 , 11 and 13

B Human Observations and Analysis of LLMs on Clue Generation

Each model demonstrates specific failure patterns in generating clues from the given text. The `GPT-4` model, for instance, often struggles when keywords appear in brackets within the text. This limitation extends to instances where place names, movie titles, or other notable phrases are enclosed in brackets or quotations, as seen in 8. Such failure patterns include misinterpretations of context, along with occasional hallucinations.

The `GPT-3.5-Turbo` model displays different challenges. For example, it frequently attempts to offer letter counts as hints, a typical feature of Arabic crossword puzzles, but consistently fails to provide accurate counts. In Arabic crosswords, clues are conventionally labeled as "horizontal" (afqi) or "vertical" (r'asī), followed by the respective clues. Many D-labeled clues generated by this model simply state "horizontal:" (afqī:) without following with a proper clue. When the "keyword" is contained in brackets, as in 12, the model sometimes generates clues that inadvertently include the answer. Similarly, `Llama-8b` encounters difficulty when Latin characters are present within the text, an issue clearly illustrated in Examples 8, 10, and 12.

While the base models `GPT-4`, `GPT-3.5-Turbo`, and `Llama-8b` exhibit specific failure patterns, the fine-tuned versions `GPT-3.5-Turbo Fine Tuned` and `Llama-8b-Fine Tuned` show a different issue. Although these fine-tuned models generate clues using the information within the text, they also tend to incorporate additional details from their broader knowledge base. This reliance on external knowledge, as observed in Example 3 (13), can lead to inaccuracies or unintended context, thereby complicating the model's ability to remain strictly within the provided text.

(من مواليد فيرونيكا إيفيت بينيت؛ 10 أغسطس 1943 - 12 يناير (Ronnie Spector :بالإنجليزية) روني سبيكتور 2022) هي مغنية أمريكية. كانت سبكتور هي المغنية الرئيسية في فرقة الروك / البوب الصوتية للفتيات "الرونتس"، والتي قدمت سلسلة من الأغاني خلال أوائل إلى منتصف الستينيات مثل «كن حبيبي»، و «حبيبي، أنا احبك». بعد ذلك، بدأت سبيكتور مسيرتها المنفردة وأصدرت منذ ذلك الحين خمسة ألبومات (سيرين في عام 1980، وشئ ما سيحدث في عام 2003، وآخر نجوم الروك في عام 2006، والقلب الإنجليزي في عام 2016) ومسرحية ممتدة واحدة (هي محادثات مع قوس قزح في عام 1999).
في عام 1986، عادت سبيكتور إلى الحياة المهنية عندما ظهرت في أغنية موسيقى البوب روك التي رشحها إدي موني وقد غنت وتعاونت مع عدة Billboard Hot 100. بعنوان «خذني إلى المنزل الليلة» والتي احتلت المركز الرابع في أعمال أخرى. يُطلق على سبيكتور «الفتاة السيئة للروك أند رول» الأصلية. في عام 2007، تم إدخال روني وفرقتها الرونتس في قاعة مشاهير الروك آند رول.

Model	Keyword	Clue	Rating
GPT4	روني سبيكتور	صوت الرونتس الرئيسي	**A** اللغز الذي تم إنشائه صحيح و لغويا و مرتبط بالنص بشكل كبير
GPT4	روني سبيكتور	الفتاة السيئة للروك	**C** كان من اللممكن ان تأخذ تقييم اعلى و لكن المودل لم يستطع ان يفهم ان هذا عبارة عن لقب و لم يستطع تكملته
GPT 3.5 Turbo	روني سبيكتور	مغنية أمريكية شهيرة في مجال الروك والبوب (٦ حروف)	**B** حاول المودل ان يعطي عدد الحروف كما هو متعارف عليه في الكلمات المتقاطعة باللغة العربية و لكن اعطى عدد الحروف بشكل خاطئ
GPT 3.5 Turbo	روني سبيكتور	اسم الفرقة التي كانت تقودها روني سبيكتور (٦ حروف)	**E** قام بإعطاء الإجابة في اللغز
GPT 3.5 Turbo Fine tuned	روني سبيكتور	صوت الرونتس الرئيسي	**A** قام بإعطاء نفس نتيجة ال GPT4
GPT 3.5 Turbo Fine tuned	روني سبيكتور	"نجمة "خذني إلى المنزل	**C** لم يتمكم المودل من التعرف على الإسم الكامل للفيلم و لكنه قريب جدا
Llama8b	روني سبيكتور	مغنية أمريكية في فرقة "الرونتز" واشتهرت بأغانيها في الستينيات	**A**
Llama8b	روني سبيكتور	مطربة روك أند رول أمريكية اشتهرت بأغنية "كن حبي"и	**B** اللغز صحيح بشكل كامل و لكن تم إنشاء معه حروف غير مفهومة
Llama8b Fine tuned	روني سبيكتور	نجمة الرونتس الأصلية	**A**
Llama8b Fine tuned	روني سبيكتور	قاعة مشاهير 2007	**D** اللغز ليس له علاقة بالكلمة المفتاحية و لكنه أتى بالتاريخ من السياق

Figure 8: Example of Ronnie Spector article in Arabic

Ronnie Spector (born Veronica Yvette Bennett; August 10, 1943 – January 12, 2022) was an American singer. Spector was the lead vocalist of the girl group rock/pop vocal band "The Ronettes," which released a series of hits in the early to mid-1960s, including "Be My Baby" and "Baby, I Love You." Later, Spector began her solo career and went on to release five albums (including *Siren* in 1980, *Something's Gonna Happen* in 2003, *The Last of the Rock Stars* in 2006, and *English Heart* in 2016) and one extended play (*She Talks to Rainbows* in 1999).
In 1986, Spector returned to the spotlight when she appeared in Eddie Money's pop-rock song "Take Me Home Tonight," which reached No. 4 on the Billboard Hot 100. She sang and collaborated with several other acts. Spector is often referred to as the "original bad girl of rock and roll." In 2007, Ronnie and her band The Ronettes were inducted into the Rock & Roll Hall of Fame.

Model	Keyword	Clue	Rating
GPT4	Ronnie Spector	The lead voice of The Ronettes.	**A** The puzzle created is accurate, linguistically correct, and closely related to the text.
GPT4	Ronnie Spector	The bad girl of rock	**C** It could have received a higher rating, but the model could not understand that this was her title and failed to complete it.
GPT 3.5 Turbo	Ronnie Spector	Famous American singer in rock and pop (6 letters)	**B** The model tried to provide the number of letters as commonly done in Arabic crossword puzzles but gave the wrong count.
GPT 3.5 Turbo	Ronnie Spector	The band led by Ronnie Spector (6 letters)	**E** It provided the answer in the clue
GPT 3.5 Turbo Fine tuned	Ronnie Spector	The lead voice of the Ronettes	**A** Gave same answer as GPT4
GPT 3.5 Turbo Fine tuned	Ronnie Spector	Star of "*Take Me Home"*	**C** The model couldn't provide the whole name of the movie
Llama8b	Ronnie Spector	An American singer in the band *The Ronettes*, famous for her songs in the 60s.	**A**
Llama8b	Ronnie Spector	An American rock and roll singer famous for the song *Be My Baby*.	**B** The clue is totally correct but a weird alphapets were generated also
Llama8b Fine tuned	Ronnie Spector	The original star of The Ronettes	**A**
Llama8b Fine tuned	Ronnie Spector	2007 Hall of Fame	**D** The Clue is irrelevant from the context

Figure 9: Example of Ronnie Spector article translated in English

أنا الحقير سلسلة أفلام من إنتاج إستديو إليمونيشن للترفيه ومن توزيع شركة يونيفرسل ستوديوز، تضم السلسلة أربعة :أفلام لغاية الآن
أنا الحقير الذي أنتج سنة 2010 والذي حقق دخلا قدر ب543 مليون دولار جعله يصنف كالفيلم السادس والثلاثين من .حيث أكبر ايرادات
وتلاه أنا الحقير 2 سنة 2013 حيث حقق ايرادات بلغت 970 مليون دولار كما صنف كسابع فيلم في قائمة اعلى أفلام .الرسوم المتحركة دخلا
لقيت أفلام السلسلة نجاحا كبيرا حيث انه بعد إصدار الجزأين الاولين لقيت المخلوقات الصفراء شهرة كبيرة من طرف الجمهور دفع الشركة لإنتاج فيلم التوابع سنة 2015 والذي صنف كثاني اعلى فيلم تحقيقا للإيرادات وكان هدف الفيلم .التعريف بالمخلوقات ومن اين اتت وما هو غرضها
وفي سنة 2017 تم إنتاج الجزء الرابع من سلسلة الافلام أنا الحقير 3 وتلقى هذا الفيلم اعجابا من المشاهدين فاق نسبة الجزأين السابقين جعله يحتل المرتبة الرابعة في نفس القائمة بمدخول فاق المليار دولار. كما انه من المتوقع إنتاج جزء . خامس من سلسلة الأفلام سنة 2020 تحت عنوان التوابع 2 وجزء سادس سنة 2024 بعنوان أنا الحقير 4

Model	Keyword	Clue	Rating
GPT4	أنا الحقير	سلسلة إليمونيشن الشهيرة	**A**
GPT4	أنا الحقير	سينما الرسوم المتحركة	**C** اللغز في سياق الفقرة و لكن غير صحيح
GPT 3.5 Turbo	أنا الحقير	أفلام رسوم متحركة تتبع مغامرات مخلوقات صفراء (٧ حروف) - أنا الحقير	**E** اللغز صحيح و لكن المودل قام بإنشاء اللغز و الجواب
GPT 3.5 Turbo	أنا الحقير	سلسلة أفلام تحقق إيرادات ضخمة وشهرة عالمية (٨ حروف) - أنا الحقير	**E** اللغز صحيح و لكن المودل قام بإنشاء اللغز و الجواب
GPT 3.5 Turbo Fine tuned	أنا الحقير	مخلوقات صفراء مشهورة	**C** قام بأخذ السياق و لكن اللغز خاطئ
GPT 3.5 Turbo Fine tuned	أنا الحقير	توابع وتحقيقات	**C** قام بأخذ السياق و لكن اللغز خاطئ
Llama8b	أنا الحقير	(HIGHEST-GROSSING FILM)	**D** قام بإنشاء لغز باللغة الإنجليزية
Llama8b	أنا الحقير	(FOLLOW-UP FILM) - "تلاه أناالحقير ٢ سنة ٢٠١٣" (After I Am Groot 2 in 2013	**E** قام بإنشاء لغز باللغة الإنجليزي و قام بإعطاء الكلمة المفتاحية في اللغز ايضا
Llama8b Fine tuned	أنا الحقير	سلسلة أفلام إستديوهات إليمونيشان	**A**
Llama8b Fine tuned	أنا الحقير	ثلاثية من الرسوم الكرتونية	**A**

Figure 10: Example of Despicable Me article

Despicable Me is a film series produced by Illumination Entertainment and distributed by Universal Studios, currently consisting of four films:
Despicable Me, produced in 2010, generated a revenue of $543 million, making it the 36th highest-grossing film.
It was followed by Despicable Me 2 in 2013, which earned $970 million, ranking as the seventh highest-grossing animated film.
The films in the series have enjoyed significant success. After the release of the first two parts, the yellow minions became immensely popular with audiences, leading the company to produce Minions in 2015. This film became the second-highest grossing animated film, with the goal of introducing the origins and purpose of the minions.
In 2017, the fourth installment, Despicable Me 3, was released and received even more audience acclaim than the previous two, reaching fourth place on the same list with over a billion dollars in revenue.
A fifth film in the series, Minions 2, was expected in 2020, and a sixth installment, Despicable Me 4, is anticipated in 2024.

Model	Keyword	Clue	Rating
GPT4	Despicable Me	Illumination's famous series	**A**
GPT4	Despicable Me	Animated cinema	**C** The puzzle is in the context of the paragraph but is incorrect.
GPT 3.5 Turbo	Despicable Me	Animated films following the adventures of yellow creatures (7 letters) - Despicable Me	**E** The puzzle is correct, but the model created both the puzzle and the answer.
GPT 3.5 Turbo	Despicable Me	A film series achieving massive revenue and global fame (8 letters) - Despicable Me	**E** The puzzle is correct, but the model created both the puzzle and the answer.
GPT 3.5 Turbo Fine tuned	Despicable Me	Famous yellow creatures	**C** It took the context into account, but the puzzle is incorrect.
GPT 3.5 Turbo Fine tuned	Despicable Me	Minions and investigations	**C** It took the context into account, but the puzzle is incorrect.
Llama8b	Despicable Me	(HIGHEST-GROSSING FILM)	**D** Generated Clue in English
Llama8b	Despicable Me	(FOLLOW-UP FILM) - *Followed by Despicable Me 2 in 2013*	**E** It created a clue in English and also provided the keyword in the puzzle.
Llama8b Fine tuned	Despicable Me	Illumination Studios film series	**A**
Llama8b Fine tuned	Despicable Me	A trilogy of cartoons	**A**

Figure 11: Example of Despicable Me article translated in English

، من مواليد 27 أكتوبر 1957، لاعب كرة قدم إنجليزي سابق، ومدرب (Glenn Hoddle :بالإنجليزية) غلين هودل .كرة قدم إنجليزي سابق
لعب مع منتخب إنجلترا لكرة القدم وعدة أندية في إنجلترا وفرنسا , درب عدة أندية إنجليزية بعد اعتزالة وخلال الفترة بين . عامي 1996 - 1999 كان مدرباً لمنتخب إنجلترا

Model	Keyword	Clue	Rating
GPT4	غلين هودل	لاعب ومدرب كرة إنجليزي	**A**
GPT4	غلين هودل	منتخب إنجلترا وأندية فرنسية	**D** اللغز في سياق الفقرة و لكن غير صحيح
GPT 3.5 Turbo	غلين هودل	اسم لاعب كرة القدم الإنجليزي السابق (غلين هودل).	**E** اللغز صحيح و لكن المودل قام بإنشاء اللغز و الجواب
GPT 3.5 Turbo	غلين هودل	مدرب كرة القدم الإنجليزي السابق الذي درب منتخب إنجلترا (غلين هودل)	**E** اللغز صحيح و لكن المودل قام بإنشاء اللغز و الجواب
GPT 3.5 Turbo Fine tuned	غلين هودل	نجم إنجلترا السابق	**A**
GPT 3.5 Turbo Fine tuned	غلين هودل	أسطورة الكرة الإنجليزية	**C** في السياق و لكن ليس من الفقرة
Llama8b	غلين هودل	لاعب كرة قدم إنگليزي) سابق لعب مع منتخبي إنگلترا (والفرنسي	**A**
Llama8b	غلين هودل	(مواليد 1957 لاعب كرة قدم -ومدرب إنگليزى) - "1957 born English footballer and coach who played for several English clubs"	**C** قام بإنشاء اللغز صحيحا و لكن قام أيضا بإنشاء الترجمة بالرغم من انه ليس مطلوب منه
Llama8b Fine tuned	غلين هودل	مدرب الأزرق خلال التسعينات	**C** قام بإنشاء لغز من السياق و لكن قام بإضافة معلومة خارج السياق
Llama8b Fine tuned	غلين هودل	نجم كرة قدم إنجليزية سابق	**A**

Figure 12: Example of glenn hoddle article

Glenn Hoddle, born on October 27, 1957, is a former English football player and coach. He played for the England national football team and several clubs in England and France. After retiring, he managed several English clubs, and between 1996 and 1999, he was the coach of the England national team.

Model	Keyword	Clue	Rating
GPT4	Glenn Hoddle	English football player and coach	**A**
GPT4	Glenn Hoddle	England national team and French clubs	**D** The puzzle is in the context of the paragraph but is incorrect.
GPT 3.5 Turbo	Glenn Hoddle	The name of the former English football player (Glenn Hoddle).	**E** The puzzle is correct, but the model created both the puzzle and the answer.
GPT 3.5 Turbo	Glenn Hoddle	The former English football coach who managed the England national team (Glenn Hoddle).	**E** The puzzle is correct, but the model created both the puzzle and the answer.
GPT 3.5 Turbo Fine tuned	Glenn Hoddle	Former England star	**A**
GPT 3.5 Turbo Fine tuned	Glenn Hoddle	English football legend	**C** In context, but not from the paragraph.
Llama8b	Glenn Hoddle	A former English football player who played with the England and French national teams	**A**
Llama8b	Glenn Hoddle	(1957-born English footballer and coach) 1957" --born English footballer and coach who played for several English clubs"	**C** It created the puzzle correctly but also provided the translation, even though it was not required.
Llama8b Fine tuned	Glenn Hoddle	The Blues' coach during the 1990s	**C** It created a puzzle from the context but added information outside of it.
Llama8b Fine tuned	Glenn Hoddle	Former English football star	**A**

Figure 13: Example of glenn hoddle article translated in English