ArabIcros: AI-Powered Arabic Crossword Puzzle Generation for Educational Applications

Kamyar Zeinalipour and **Mohamed Zaky Saad** and **Marco Maggini** and **Marco Gori**
DIISM, University of Siena
Via Roma 56, Siena, Italy
{kamyar.zeinalipour2, marco.maggini, marco.gori}@unisi.it
m.zakyanwarzakymo@student.unisi.it

Abstract

This paper presents the first Arabic crossword puzzle generator driven by advanced AI technology. Leveraging cutting-edge large language models including GPT4, GPT3-Davinci, GPT3-Curie, GPT3-Babbage, GPT3-Ada, and BERT, the system generates distinctive and challenging clues. Based on a dataset comprising over 50,000 clue-answer pairs, the generator employs fine-tuning, few/zero-shot learning strategies, and rigorous quality-checking protocols to enforce the generation of high-quality clue-answer pairs. Importantly, educational crosswords contribute to enhancing memory, expanding vocabulary, and promoting problem-solving skills, thereby augmenting the learning experience through a fun and engaging approach, reshaping the landscape of traditional learning methods. The overall system can be exploited as a powerful educational tool that amalgamates AI and innovative learning techniques, heralding a transformative era for Arabic crossword puzzles and the intersection of technology and education.

1 Introduction

Combining traditional puzzle constructs with educational components, pedagogical crosswords foster interactive learning experiences by integrating vocabulary, history, sciences, and other subjects. Intriguingly, they effectively strengthen students' vocabulary and spelling abilities due to the puzzles' requirement for accurate spelling (Orawiwatnakul, 2013; Dzulfikri, 2016; Bella and Rahayu, 2023). These puzzles are particularly significant for language acquisition and learning specific technical terms (Nickerson, 1977; Sandiuc and Balagiu, 2020; Yuriev et al., 2016). Moreover, they enhance problem-solving, critical thinking skills, and memory retention, thereby making the learning process enjoyable and productive (Kaynak et al., 2023; Dol, 2017; Mueller and Veinott, 2018; Dzulfikri, 2016; Zirawaga et al., 2017; Bella and Rahayu, 2023; Zamani et al., 2021; Yuriev et al., 2016).

Creating Arabic educational crosswords can be challenging due to the required wordplay expertise. However, with the help of innovations in natural language processing, Large Language Models (LLMs) are now able to generate high-quality Arabic crossword clues. LLMs are pre-trained on a mix of sources like books, academic articles, and web content and this wide spectrum of content enables them to create challenging and engaging crossword clues. This aids puzzle designers and improves the solver's experience, enabling even beginners to design personalized puzzles.

The results show that the proposed approach can be effectively employed to generate Arabic educational crossword puzzles, introducing an innovative system using LLMs to generate top-quality clues and answers. By inputting text passages or keywords, the system generates clue-answer pairs, based on techniques like fine-tuning and few-shot learning used for generation. We also present models to filter inappropriate clue-answer pairs for puzzle construction optimization, propose an advanced algorithm for designing Arabic educational crossword layouts, and provide a comprehensive dataset of curated Arabic clue-answer pairs. These advances simplify the creation of Arabic pedagogical crosswords and expand their potential for their broader exploitation.

This paper is structured as follows; section 2 explores relevant literature; section 3 discusses the collected Arabic dataset; section 4 outlines the research methodology, section 5 presents the findings, and, finally, section 6 summarizes the overall outcomes.

2 Related works

The generation of crosswords represents a complex task that has been addressed by some research works. These studies have utilized a variety of tools, including traditional dictionaries and thesauri, or

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 288–301

have engaged in the linguistic analysis of text content derived from the web.

Rigutini et al. (Rigutini et al., 2008, 2012) pioneered the first fully automated crossword creator system in 2008. The proposed system leverages natural language processing techniques to generate crossword clues by scraping related documents from the web, extracting relevant text segments, and using part-of-speech tagging, dependency parsing, and WordNet-based similarity measures. This approach produces clues based on specific ranking criteria.

An alternative methodology for crossword construction using natural language processing is documented in (Ranaivo-Malançon et al., 2013). This approach consists of a four-stage process, which includes initial data retrieval of a targeted topic-specific text compilation, extraction of complete sentences, determination of the dependency syntactic structure of each sentence, and removal of words from stop-lists. The extracted information undergoes a transformation into a graph representation for depth-first pre-order search. This framework integrates pre-processing, candidate identification, clue formation, and answer selection.

Esteche et al.'s study (Esteche et al., 2017) delved into the creation of Spanish language crossword puzzles from news articles. The system is based on a twofold procedure: initially pivotal terms are identified and their meanings are isolated from a trusted online dictionary. Subsequently, these definitions are employed as hints for the assembly of compelling crossword puzzles.

In a related study, Arora et al. (Arora and Kumar, 2019) discuss a software tool that uses NLP techniques to identify crucial keywords for creating crossword puzzles in various Indian languages. Their proposed framework, SEEKH, combines statistical and linguistic methods to highlight significant keywords useful for crossword creation.

Despite significant research, accurately generating comprehensive and unique clue-answer sets from linguistic corpora remains a challenge, particularly for the unique linguistic nuances of Arabic. To address these issues, we propose an innovative methodology using LLMs to create intricate educational clues. As a pioneering attempt, our technique successfully generates Arabic crossword puzzles, filling a gap unaddressed by previous methods. By generating intellectually stimulating and original crossword puzzles, this novel approach enhances learners' deep understanding of the subjects by providing comprehensive answers. Hence, the proposed work not only brings novelty to Arabic crossword generation, but also offers a groundbreaking solution in the realm of educational tools.

3 Dataset

Given the scarcity of data for Arabic crossword puzzles, a clue-answer pair dataset was gathered manually. The dataset encompasses the period from 2020 to 2023.

During the initial stage of data collection, we pursued all accessible crossword puzzles, encompassing web-based games, journals, and magazines, ensuring that the training set comprised accurate clue-answer pairs sourced from original Arabic crossword puzzles. We had a collection of crossword images, and we needed to extract the text contained within these images to build a dataset for obtaining the text from these images. To accomplish this, we initially utilized optical character recognition (OCR) as a tool. However, it's important to note that the OCR process was predominantly supervised by humans who used it to facilitate the extraction. Additionally, human validation was employed to evaluate both spelling errors within the journals and the overall quality of the clue-answer pairs. This meticulous process resulted in a catalog of 57,706 entries from two different sources. One of them was the Al-Joumhouria Journal, from which we manually extracted 5,661 Clue and Answer pairs. The other source was the Al-Ghad Electronic Journal, where we utilized the OCR tool to assist in the extraction process. In the end, this yielded 25,908 unique pairs with answers varying in length from 1 to 21 characters, with the majority of the data falling within a specific answer's character length range from 2 to 9 (see Fig. 1).

The structure of the pairs is recurrent. For instance, some of the pairs are synonyms or antonym definitions, that define the answer by means of one or more synonyms or antonyms. An example of this category includes "موتي" with the answer "حتفي". Some others were general information, such as for example "دولة عربية" with the answer "مصر". Another structure can be a word but the letters are not in order, as for example "جميل مبعثرة" with the answer "ل م ي ج". Finally, the definition can give the word and requires part of it for the answer, as for instance "نصف نادر" with the answer "در".

A meticulous pre-processing step was carried out

on the data to refine it for fine-tuning. This involved the elimination of Arabic accents, redundant pairs, and markers suggesting a reversal in crosswords—an idiosyncrasy of Arabic. The aim of this study was to pave the way for further research by making this processed dataset publicly accessible, encouraging other scholars to contribute to this field. [1]

4 Methodology

The proposed system includes several components, such as mechanisms to generate clue-answer pairs using user-provided text or keywords, and a crossword schema generator as depicted in Figure 2. Users can input any instructional text to extract relevant clue-answer pairs or insert a list of chosen keywords to generate clues. After combining both clue-generation methods, the quality of the generated pairs is evaluated using specific validation modules. Users can then review and select their preferred clue-answer sets, which are employed in the final step by a separate module for creating the crossword layout.

4.1 Path (a): Generating clue-answer pairs from input text

In our system, we employ zero-shot and few-shot learning to create clue-answer pairs. This process involves segmenting the text into paragraphs, keyword extraction, generating potential clues, and rigorously validating the resulting pairs. More details on these stages are provided later in the paper. Our experiments are based on the models GPT3.5-Turbo and GPT4 (Brown et al., 2020) (OpenAI, 2023). We use dynamic experimental approaches, including both customized English and Arabic prompts, to assess prompt language strategies' effectiveness across models.

4.1.1 Keyword extraction

Our Few-Shot Learning Framework begins with prompt construction, involving the incorporation of extensive educational text that includes potential crossword keywords. These keywords, chosen to match possible answers from the provided text, enhance precision as the LLM is prompted with well-curated information. The process concludes by inputting the educational text and the tailored prompt to the LLM, enabling it to utilize its few-shot learning experiences to extract potential keywords from the input paragraph. This mechanism allows the LLM to extrapolate potential keywords effectively, resulting in a more comprehensive analysis.

4.1.2 Generating crossword clues from the extracted keywords

In this stage, we harness the power of few-shot learning once more. By utilizing the keywords identified in the previous phase along with the input text, we generate relevant crossword clues. Additional information, including an example of valid paragraph, keywords, and clues, was also input into the LLM along with the target text and previously generated keywords that needed crossword clues. This strategy enabled the LLM to craft unique clues by leveraging the supplied text and initial keywords. This systematic approach significantly improves the precision and relevance of the generated crossword clues, ensuring each clue aligns with the context of the provided text and identified keywords.

4.1.3 Path(a) Validation

To enhance the quality and appropriateness of our generated keyword-clue pairs, a method to exclude low-quality and inappropriate pairings is applied in several discrete stages. The first step utilized a filter system to eliminate answers containing more than three words, which are typically unsuitable for crossword puzzles. Our empirical research has shown that the LLM can occasionally produce clues by drawing upon its innate knowledge rather than relying solely on the provided text. Additionally, in instances where the generated clues did not effectively capture relevant keywords, we took steps to address this issue. To enhance the quality of our output and ensure the creation of appropriate clue-answer pairs, we employed a zero-shot learning approach, effectively filtering out undesired clues.

4.2 Path (b): Generating clues based on provided answers

There may be scenarios where we need to generate crossword clues using provided answers without a full-text context. To face this task, we deployed a holistic approach that started with fine-tuning different language models using the introduced Section 3, each specifically designed for this task. We further enriched this scheme by using data from these fine-tuned models to create various classi-

[1]The dataset is available at https://huggingface.co/datasets/Kamyar-zeinalipour/AR_CW

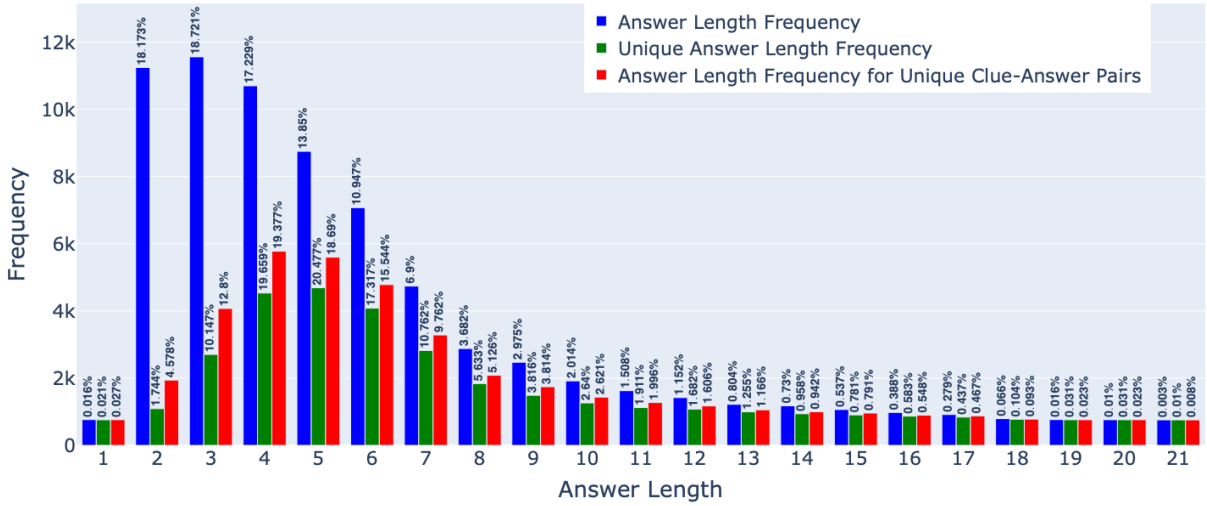

Figure 1: The introduced dataset entries are visually presented in terms of answer length distribution. The blue bars represent all the clue-answer pairs, while the green bars depict the frequency of unique answers. Additionally, the red bars indicate the frequency of unique answer-clue pairs.

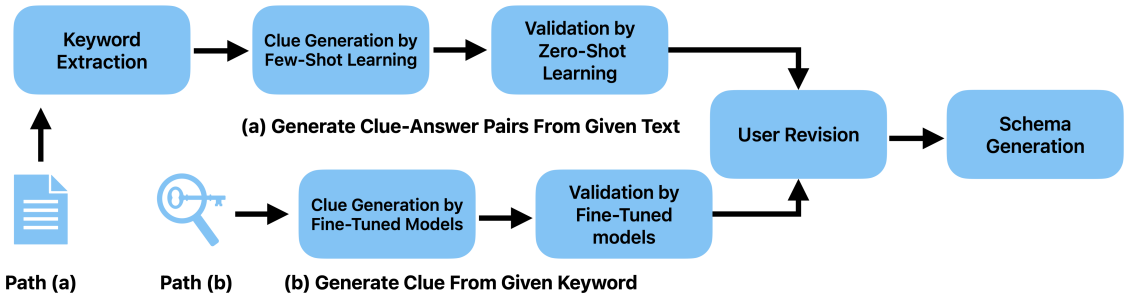

Figure 2: Overall system architecture. Path (a) Clue-answer generation from input text. Path (b) Clue generation from the given answers.

fiers. These classifiers aim to differentiate between high-quality generated clue-answer pairs and less suitable alternatives.

4.2.1 Fine-tuning LLMs to generate clues from provided answers

In the pursuit of crafting crossword clues from given answers and textual information, our research delved into the optimization of language models. This refinement process was informed by the dataset meticulously outlined in Section 3. Our evaluation encompassed a spectrum of models, notably the robust Lamma2 13B and the efficient Llama2 7B, distinguished by their substantial 13 billion and 7 billion parameters, respectively. We also examined the 1.5 billion-parameter GPT2-XL model, recognized for its versatility, and the T5 Base model, endowed with 350 million parameters as expounded in (Brown et al., 2020).

This section encapsulates our methodical approach to model selection, emphasizing the diversity of parameters and architectures considered in our quest to enhance the generation of crossword clues. The subsequent analysis and results, detailed in the following sections, shed light on the efficacy and performance of these fine-tuned language models in the context of crossword clue generation.

4.2.2 Path(b) Validation

The design of the overall system focuses on enhancing the overall quality of the generated clue-answer pairs. We incorporated a filtering process into the system pipeline to enhance the quality and usability of the generated pairs. Using the data obtained from the fine-tuned language models, we created a classifier capable of distinguishing between effective and unsuitable clues.

For this purpose, several models were fine-tuned, including GPT3-DaVinci with 175 billion parameters, GPT3-Curie with 13 billion parameters, GPT3-

Babbage with 1.3 billion parameters, GPT3-Ada with 350 million parameters (Brown et al., 2020), and BERT-base-Arabic with 110 million parameters (Raffel et al., 2020; Safaya et al., 2020). These models provided important insights into their respective capabilities and aided in validating the generated clues.
Our primary objective was to use these models with their varying parameter counts to comprehensively evaluate their effectiveness in filtering and validating the generated clues. This methodology aimed to ensure only high-quality and contextually relevant clues were retained, thereby improving the overall precision and functionality of our system.

4.3 Schema Generator

The algorithm for creating educational crossword puzzles follows a streamlined approach using input parameters such as the answer list, workspace dimensions, and termination criteria. Initially, a central answer is placed randomly followed by strategically adding surrounding answers. This cycle of adding and occasionally removing the recently added answers or entirely resetting is repeated until an optimal solution is obtained. The quality of the crossword is evaluated through a comprehensive scoring process. Each solution's merit is determined by the following scoring formula:

$$\text{Score} = (\text{FW} + 0.5 \cdot \text{LL}) \cdot \text{FR} \cdot \text{LR} \quad (1)$$

The variables exploited in this formula correspond to the following metrics:

- Filled Words (FW): This represents the count of the added words, signaling the puzzle's completeness.
- Linked Letters (LL): This counts the instances of letter-sharing between intersecting words, indicating the puzzle's coherence.
- Filled Ratio (FR): This metric, calculated as the filled letters count divided by the area of the smallest covering rectangle, showcases the efficiency of the crossword's space utilization.
- Linked Letters Ratio (LR): By dividing LL by the total letter count, LR highlights the extent of letter linkage and word-relations within the puzzle.

These four criteria collectively contribute to the evaluation and selection of the optimal solution during the algorithm execution.
The algorithm makes use of a variety of stopping criteria to guide its decision-making and determine when to end the crossword construction. These criteria are as follows:

- Minimum Number of Answers: The algorithm stops once it has added a preset minimum count of answers to the grid, ensuring an adequate crossword complexity.
- Minimum Filled Ratio Threshold: A certain threshold of the filled ratio, when met or surpassed, triggers the algorithm to stop, preventing the overabundance of empty spaces and maintaining appealing aesthetics.
- Grid Rebuilding Limit: The algorithm ceases to operate if the grid's reconstruction exceeds a set count, avoiding getting stuck in inefficient solutions and encouraging exploration of other possibilities.
- Maximum Time Duration: Upon reaching the allowed maximum time duration, the algorithm finishes, ensuring the process is time-efficient and the resources are optimally utilized.

This method allows the algorithm to identify the highest-scoring solution, enabling efficient production of high-quality crosswords given its input parameters. Furthermore, the algorithm can prioritize a list of "preferred answers," increasing their chances of inclusion, thereby ensuring that the crossword design aligns with specific objectives or preferences.

5 Experiments

In this section, we detail the empirical evaluation of the proposed system, focusing on individual elements and their roles within the overall framework.

5.1 Experimental Evaluation: Path (a)

This paper's experimental dataset aims to rigorously assess our system's output quality in relation to various language prompts. We conducted an in-depth investigation using two prompt types, categorized as English and Arabic. Two different models, GPT4 and GPT3.5 Turbo, were used for evaluation. The comprehensive list of prompts can be found within the paper's Appendix B. This provides comprehensive evaluations of linguistic aspects, leading to robust, multifaceted findings. The

system underwent thorough evaluation using 100 educational selected Wikipedia paragraphs to examine performance in different language contexts. Performance markers were established based on empirical evidence. Evaluation guidelines, created under expert supervision, ensured robust results. Detailed criteria for evaluation are in Appendix A, and cumulative findings are presented in Table 1. GPT4 and GPT3.5-Turbo models performed impressively in English prompts, achieving keyword extraction accuracy of 95.05% and 92% respectively. They similarly excelled in Arabic prompts with accuracies of 94.32% and 97.38%.

In clue generation, these models demonstrated their value in retrieving meaningful information. In English prompts, GPT4 and GPT3.5-Turbo reached accuracies of 94.62% and 55.33%, respectively, while GPT4 and GPT3. marking respective accuracies of 93.23% and 37.78% in Arabic prompts.

The evaluation of clue-answer pairs yielded satisfactory results. In English, the GPT4 and GPT3.5-Turbo models exhibited accuracies of 87.76% and 89.04% and maintained substantial accuracy of 84.01% and 89.32% in Arabic prompts.

In the final evaluation, which included system-wide validation and acceptability of potentially generated clues and answers, both models upheld their performance. it means we analyze the clue-answer pairs that align with the validation part of the system, and then culminate in the calculation of the proportion of generated clues and answers that successfully pass the criteria established through human oversight which is the total performance of the model. It was overall 78.95% and 74.6% for the GPT4 model for English and Arabic prompts, respectively, while the GPT3.5-Turbo model had a total performance of 46.68% and 68.83% for English and Arabic prompts respectively.

Figure 3 provides a practical illustration of this system component's functionality. It sequentially depicts the transformation from initial text to final crossword clue-answer pairs, demonstrating input paragraphs (a), keyword extraction (b), clue generation (c), and clue-answer pair validation (d). This visual representation clarified the system's operational process, elucidating its capability to turn text into precise crossword clues and their corresponding answers. Comprehensive translations for the content depicted in Figure 3 can be found in the paper's Appendix C.

5.2 Experimental Evaluation: Path (b)

This section details experimental tests on clue generation and validation from keywords using three distinct models, GPT3-DaVinci, GPT3.5-Turbo, and GPT3-Curie. These were designed and optimized based on concepts discussed in Section 4.2.1, with a specific emphasis on forming clues from identified keywords.

In the preparation phase, a subset of the dataset discussed in Section 3, specifically 25,908 unique clue-answer pairs, was selected. Afterwards, each refined model produced 2,000 clues which were evaluated using human judgement based on the criteria presented in Appendix A.

In conclusion of our evaluation, Table 2 presents the results, highlighting the performances of GPT3-DaVinci, GPT3.5-Turbo, and GPT3-Curie. These models successfully generated satisfactory clues 41.9%, 81%, and 21.35% of the time, respectively. Observations indicate that GPT3.5-Turbo significantly outperforms the other models in the task of clue generation from the given keywords. For a thorough assessment of the generated clues, a detailed review identifying acceptable and unacceptable cases was undertaken. Each clue-answer pair was carefully examined and categorized, Tables 3 and 4 present illustrative clues generated by distinct fine-tuned models. Table 3 demonstrates instances of well-constructed clues, while Table 4 highlights cases of unacceptable clue generation. Detailed translations for these clues can be located in the Appendix C. This meticulous evaluation facilitated performance analysis of the algorithm, notably its ability to generate captivating crossword puzzles.

Several classifiers were developed in this study. Coupled with various language models, they enabled the distinction between suitable and unsuitable clue-answer pairings. The results from the evaluation of the test set are shown in Table 5.

The process utilized a dataset of 6,000 human-evaluated instances from previous steps to build several classifiers. The dataset was divided, with 80% used for training, and the remaining 20% for testing classifier performance. The analysis revealed that the dataset consists of 52% acceptable clues and 48% unacceptable ones. The system's effectiveness was gauged through the accuracy of four distinct classifiers - GPT3-DaVinci, GPT3-Curie, GPT3-Babbage, GPT3-Ada, and Bert in discerning between satisfactory and unsatisfactory clues. Notably, GPT3-DaVinci topped the list

Table 1: Assessment outcomes of the clue-answer pairs generated from the provided Text.

Model	*System Part*	*English Prompt*	*Arabic Prompt*
GPT4	Keyword Extractor	95.05 %	94.32%
	Clues Generator	94.62 %	93.23 %
	Validator	87.76 %	84.01 %
	Total performance	78.95 %	74.6 %
GPT3.5-Turbo	Keyword Extractor	92 %	97.38%
	Clues Generator	55.33 %	37.78 %
	Validator	89.04 %	89.32 %
	Total performance	46.68 %	68.83 %

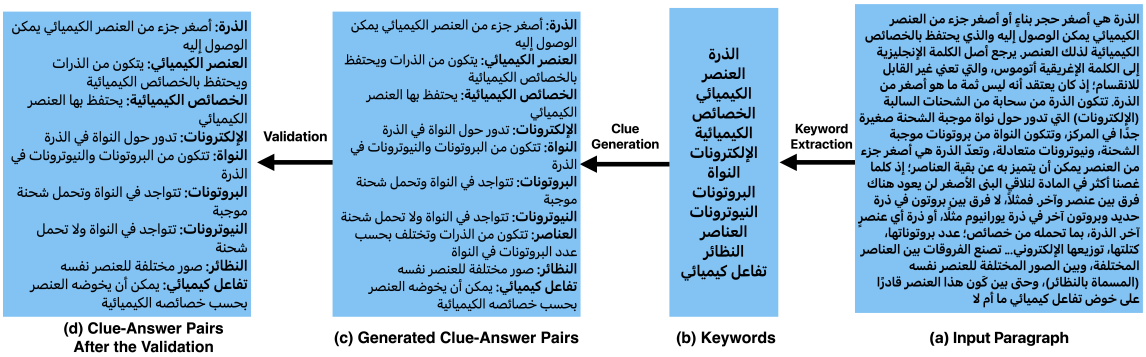

Figure 3: A comprehensive collection of clue-answer pairs generated by the introduced system from a given text, providing illustrative examples.

Table 2: Assessment outcomes of the clues generated from the provided keyword.

Model	*% of acceptable clues*
GPT3-DaVinci	41.9
GPT3-Curie	21.35
GPT3.5-Turbo	81

Table 3: Acceptable clues from given keywords using various models.

Model	*Clue-Answer pair*
GPT3-DaVinci	نجوم : في السماء ليلا
GPT3-Curie	كروم : من المعادن
GPT3.5-Turbo	قوة : قدرة

with an exceptional 85.74% accuracy, followed by GPT3-Curie at 81.29%. GPT3-Babbage showed decent results with 78.69% accuracy, while GPT3-Ada and Bert had fair performances with 79.19% and 71.42% accuracy, respectively. These results underscore the commendable performance of these classifiers in identifying agreeable clues.

Table 4: Unacceptable clues from given keywords using various models.

Model	*Clue-Answer pair*
GPT3-DaVinci	زرافة : من الحشرات
GPT3-Curie	مثلث : مثنى مثلث
GPT3.5-Turbo	عمة : اخت والد او والدة

5.3 Schema Generation

Our algorithm for schema generation envisages a spectrum of educational crosswords utilizing a group of generated clue-answer pairs. Illustrated in Figure 4 is a comprehensive Arabic educational crossword about physics, crafted by the proposed system. The clue-answer pairs are procured either from a text (path (a), refer to Figure 3) or directly produced from a keyword (path (b), denoted by examples marked with a $\star$), as observed in Table 3.

6 Conclusions

The work featured in this paper focuses on multiple innovative offerings, among which is the introduction of a comprehensive dataset for Arabic

Table 5: Classifier performance on distinguishing acceptable Clue-Answer pairs

Model	*accuracy %*	*precision %*	*recall %*	*F1 Score*
GPT3-Dvinci	85.74	83.39	85.26	0.8431
GPT3-Curie	81.29	78.86	79.89	0.7937
GPT3-Babbage	78.69	75.17	78.54	0.7682
GPT3-Ada	79.19	77.48	75.75	0.7660
Bert-base-Arabic	71.42	67.91	70.04	0.6896

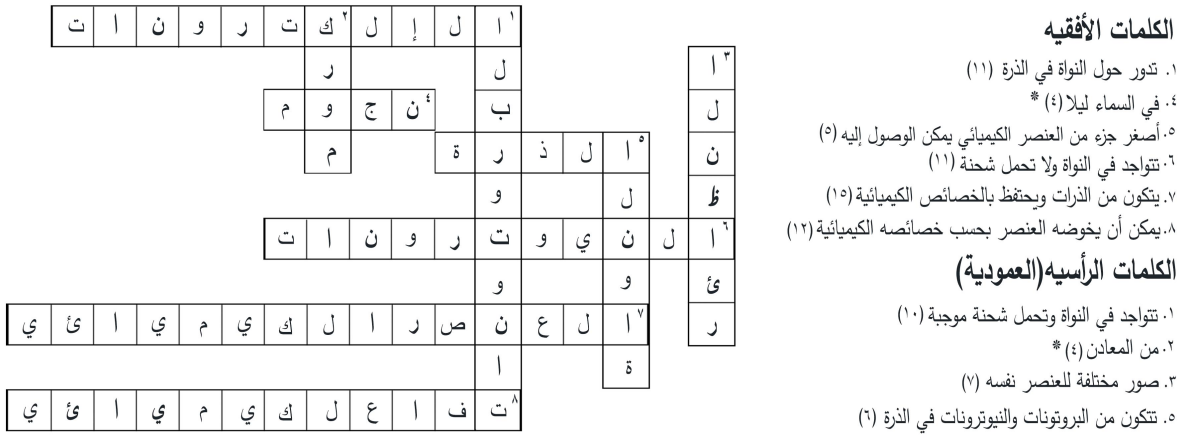

Figure 4: An illustrative Arabic educational crossword generated through the proposed system.

clue-answer pairs. In addition to this, we have also formulated a ground-breaking method employing large language models that generate educational Arabic crossword puzzles influenced by the provided texts or given keywords.

To uphold stringent quality standards in our methodology, our approach integrates human oversight in conjunction with specific guidelines (see Appendix A). In the process of generating clue-answer pairs from textual data, we conducted experiments using two distinct models: GPT-4 and GPT3.5-Turbo, while employing prompts in both English and Arabic languages. We conducted various types of evaluations considering different parts of the system and overall performance:

- Keyword Extraction: Notably, when paired with Arabic prompts, GPT3.5-Turbo exhibited exceptional performance, successfully generating high-quality keywords with an impressive accuracy rate of 97.38%.

- Crossword Clue Generation: GPT4, when prompted in English, consistently produced relevant and well-suited crossword clues, achieving a commendable success rate of 94.62%.

- Validation Component: Within our system, the validation step was a critical component. GPT3.5-Turbo, when prompted in Arabic, demonstrated superior performance in this role, boasting an impressive validation accuracy rate of 89.32%.

- Total Performance: GPT4 displayed remarkable proficiency in this role, surpassing expectations with an impressive validation accuracy rate of 78.95% when prompted in English.

In our quest to generate clues from provided keywords, we engaged in the fine-tuning process using a curated dataset (refer to Section 3). We fine-tuned three distinct models, namely GPT3-DaVinci, GPT3.5-Turbo, and GPT3-Curie. We rigorously tested the performance of each model by generating clues for a carefully chosen set of 2000 educational-related keywords. Notably, the fine-tuned GPT3.5-Turbo outperformed the others, consistently producing high-quality clues with a remarkable success rate of 81%.

Utilizing the data generated through the evaluation

of fine-tuned models, we construct classifiers to distinguish between acceptable and non-acceptable clues for a specified keyword. The most effective model in this task was GPT3-Davinci, achieving an impressive accuracy rate of 85.74%.

Our process to produce educational crossword layouts is both efficient and diverse. We hope that these findings will enrich the learning process and foster interactive learning. The developed system can be integrated into current teaching methods to enhance educational practices. As a future course of action, we plan on venturing into the development of more advanced models for more direct clue and answer pair generation and examine specialized models for different clue types. We also intend to implement this system in actual classrooms and evaluate its impact. Our goal is to revolutionize the creation of educational crossword puzzles and usher in an era of unique teaching practices.

Acknowledgments

The funding for this paper was provided by the TAILOR project and the HumanE-AI-Net projects, both supported by the EU Horizon 2020 research and innovation program under GA No 952215 and No 952026, respectively.

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.

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).

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.

OpenAI. 2023. Gpt-4 technical report.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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.

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.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. *arXiv preprint arXiv:2007.13184*.

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.

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.

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.

A Appendix

This study entailed developing a classifier to distinguish optimal and sub-optimal crossword clue-answer pairs. Crossword puzzles necessitate linguistic acumen, innovation, and adherence to construction guidelines for quality clues and answers. Such a classifier auto-evaluates the clue-answer quality, aiding puzzle designers and improving puzzle-solving experiences. This provides insight into key aspects of language and puzzle architecture.

The development of a robust framework for determining acceptable and unacceptable crossword clue-answer pairs is crucial to the effectiveness of a classifier. This provides the groundwork upon which our classifier can effectively discriminate between high-quality clues and ill-fit ones. Rigorous adherence to these guidelines facilitates accuracy in quality evaluation by the classifier and ultimately enhances the appeal and satisfaction derived from crossword puzzles.

Let us now probe into the salient features of the guideline for assessing crossword clue-answer quality:

- Coherence and Relevance: An ideal pair of clues and answers should display an evident and significant association between the two. The clue should offer adequate context or prompts that guide solvers toward the desired solution. The answer should be linear to the clue and sound logical within the subject matter or theme of the given puzzle.

- Wordplay and Creativity: A finely constructed crossword clue frequently employs wordplay, ingenious nuances, or concealed connotations that provoke and fascinate solvers. Seek clues that necessitate unconventional thinking, dual meanings, or linguistic resourcefulness. An effective clue-answer duo will enthrall the solvers, enhancing the puzzle's intrigue and pleasure.

- Unambiguity and Specificity: Clues should be unequivocal and clear-cut, presenting solvers with a distinct and exact solution. Refrain from clues that allow for multiple interpretations or result in various potential answers. The aim is to propose a single accurate answer that correlates directly with the intended meaning of the clue.

- Linguistics and Grammar: Both the clue and the answer should conform to correct grammar, syntax, and language norms. It's essential to verify that the language utilized in the clue-answer duo is grammatically accurate, coherent, and appropriate for a crossword puzzle.

- Universal Knowledge and Equity: Clues should be based on general knowledge or facts that a wide spectrum of solvers would reasonably be anticipated to understand. Refrain from using excessively obscure or specialized references, which only a small subset of solvers would recognize. An optimal clue-answer match should maintain a balance between challenge and fairness, accommodating a varied assortment of puzzle aficionados.

Adhering to these guidelines, we can construct a dataset capable of building a dependable classifier to differentiate between well-formulated crossword clue-answer pairs and those that are nonsensical or inappropriate. This classifier holds the potential to transform the process of creating, evaluating, and solving crossword puzzles. It offers crucial insights into the art of crafting puzzles that are both engaging and intellectually challenging.

B Appendix

The following prompts were employed for (Keyword Generation, Clue Generation, and Clue Verification) in both the Arabic and English versions:

English Keyword Extraction Prompt:

"Objective: Your task is to extract keywords (maximum 2 words) from a given text to create short crossword definitions. Please follow these steps to achieve the objective:

Keyword extraction: Extract the most important keywords from the text.

Validate keywords: Check if the keywords are well explained in the given text.

Final keywords: Remove all the keywords that are not well-defined in the text, based on the previous step.

Text: {text}

Here is an example Text:

الفقرة: ألأسد حيوان من الثدييات من فصيلة السنوريات وأحد السنوريات الأربعة الكبيرة المنتمية لجنس النمور ، وهو يُعد ثاني أكبر السنوريات في العالم بعد البر، حيث تفوق كتلة الذكور الكبيرة منه ٢٥٠ كيلوغراما (٥٥٠ رطلًا). تعيش معظم الأسود البرية المتبقية اليوم في إفريقيا جنوب الصحراء الكبرى، ولا تزال جمهرة واحدة صغيرة مهددة بالانقراض تعيش في آسيا بولاية غوجرات في شمال غربي الهند. كان موطن الأسود شاسعًا جدًا في السابق، حيث كانت تتواجد في شمال إفريقيا، الشرق الأوسط، وآسيا الغربية، حيث انقرضت منذ بضعة قرون فقط. وحتى بداية العصر الحديث (الهولوسين، منذ حوالي ١٠,٠٠٠ سنة)، كانت الأسود تُعتبر أكثر ثدييات اليابسة الكبرى انتشارا بعد الإنسان، حيث كانت توجد في معظم أنحاء إفريقيا، الكثير من أنحاء أوراسيا من أوروبا الغربية وصولا إلى الهند، وفي الأمريكيتين، من يوكون حتى البيروآ

Below are the legitimate keywords extracted from the provided text:

الكلمات المفتاحية: الأسد، الثدييات، فصيلة السنوريات، الأسود البرية، إفريقيا، الهند، شمال إفريقيا، الشرق الأوسط، آسيا الغربية، انتشار، الإنسان ، نمور ، ذكور

Use the following output format:

Keywords: <Final keywords>"

English Clue Generation Prompt:

"Your objective is to create short crossword clues for a list of keywords based on the given text:

Keywords: {keywords}
Text: {text}

Follow these steps to achieve the task:

Identify the part of the text that contains information about each provided keyword.

Generate short Arabic crossword clues (maximum 4 words) for all the keywords, using just the information from the text.

Here is an example Text:

الفقرة: أسد حيوان من الثدييات من فصيلة السنوريات وأحد السنوريات الأربعة الكبيرة المنتمية لجنس النمور. وهو يُعد ثاني أكبر السنوريات في العالم بعد البر، حيث تفوق كتلة الذكور الكبيرة منه ٢٥٠ كيلوغراما (٥٥٠ رطلًا). تعيش معظم الأسود البرية المتبقية اليوم في إفريقيا جنوب الصحراء الكبرى، ولا تزال جمهرة واحدة صغيرة مهددة بالانقراض تعيش في آسيا بولاية غوجرات في شمال غربي الهند. كان موطن الأسود شاسعًا جدًا في السابق، حيث كانت تتواجد في شمال إفريقيا، الشرق الأوسط، وآسيا الغربية، حيث انقرضت منذ بضعة قرون فقط. وحتى بداية العصر الحديث (الهولوسين، منذ حوالي ١٠,٠٠٠ سنة)، كانت الأسود تُعتبر أكثر ثدييات اليابسة الكبرى انتشارا بعد الإنسان، حيث كانت توجد في معظم أنحاء إفريقيا، الكثير من أنحاء أوراسيا من أوروبا الغربية وصولا إلى الهند، وفي الأمريكيتين، من يوكون حتى البيرو.

Below is a list of valid keywords for the provided text:

الكلمات المفتاحية: أسد، حيوان، ثدييات، سنوريات، سنوريات الأربعة، جنس النمور، ببر، الزكور الكبيرة، إفريقيا، صحراء الكبرى، أمريكيتين

Here is a compilation of valid clue-answer pairs corresponding to the provided keywords and text:

Keyword: أسد
Clue: حيوان ثديي من السنوريات

Keyword: حيوان
Clue: ينتمي لفصيلة السنوريات

Keyword: ثدييات
Clue: نوع من الحيوانات

Keyword: سنوريات
Clue: تشمل الأسد

Keyword: سنوريات الأربعة
Clue: مجموعة من السنوريات الكبيرة

Keyword: جنس النمور
Clue: يعتبر الأسد منه

Keyword: ببر
Clue: السنورية الأكبر في العالم

النمور ، وهو يُعد ثاني أكبر السنوريات في العالم بعد البر، حيث تفوق كتلة الذكور الكبيرة منه ٢٥٠ كيلوغراما (٥٥٠ رطلًا). تعيش معظم الأسود البرية المتبقية اليوم في إفريقيا جنوب الصحراء الكبرى، ولا تزال جمهرة واحدة صغيرة مهددة بالانقراض تعيش في آسيا بولاية غوجرات في شمال غربي الهند. كان موطن الأسود شاسعًا جدًا في السابق، حيث كانت تتواجد في شمال إفريقيا، الشرق الأوسط، وآسيا الغربية، حيث انقرضت منذ بضعة قرون فقط. وحتى بداية العصر الحديث (الهولوسين، منذ حوالي ١٠,٠٠٠ سنة)، كانت الأسود تُعتبر أكثر ثدييات اليابسة الكبرى انتشارا بعد الإنسان، حيث كانت توجد في معظم أنحاء إفريقيا، الكثير من أنحاء أوراسيا من أوروبا الغربية وصولا إلى الهند، وفي الأمريكيتين، من يوكون حتى البيرو.

الكلمات المفتاحية التي تم إستخراجها كالأتي الكلمات المفتاحية: الأسد، الثدييات، فصيلة السنوريات، الأسود البرية، إفريقيا، الهند، شمال إفريقيا، الشرق الأوسط، آسيا الغربية، انتشار، الإنسان ، نمور ، ذكور

شكل النتيجة النهائية: »الكلمات المفتاحية«

Arabic Clue Generation Prompt:

هدفك هو إنشاء الغاز قصيرة للعبة الكلمات المتقاطعة مناسبة للكلمات المفتاحية الأتية استنادا الى الفقره بحيث ان يكون كل كلمة مفتاحية يوجد لها اللغز خاص بها سأقوم بتزويدك بمثال بعد طريقة اتمام المهمة

الكلمات المفتاحية: الكلمات المفتاحيه

الفقرة: الفقره

استخدم هذه الطريقة لإتمام المهمة:

قم بالتعرف على الاجزاء التي تحتوي على الكلمات المفتاحية في الفقرة قم بإنشاء لغز لكل الكلمات المفتاحية بإستخدام المعلومات في الفقرة تأكد من انه لا يوجد اي كلمات مساعدة للوصول إلى الكلمة المفتاحية لهذا اللغز في اللغز الذي تم إنشاءه قم بإنشاء اللغز بحيث يدل فقط على الكلمة المفتاحية و لا يتواجد في اللغز نفسه تأكد من ان اللغز اجابته كلمة مفتاحية واحده فقط تأكد من ان لكل من الكلمات المفتاحية يوجد له لغز اذا وجد لغز مناسب

مثال للمطلوب:

- الفقرة كالأتي الفقرة: أسد حيوان من الثدييات من فصيلة السنوريات وأحد السنوريات الأربعة الكبيرة المنتمية لجنس النمور. وهو يُعد ثاني أكبر السنوريات في العالم بعد البر، حيث تفوق كتلة الذكور الكبيرة منه ٢٥٠ كيلوغراما (٥٥٠ رطلًا). تعيش معظم الأسود البرية المتبقية اليوم في إفريقيا

Keyword:الزكور الكبير
Clue: تتجاوز وزنها ٢٥٠ كيلوغرام

Keyword: إفريقيا

Clue: مكان عيش معظم الأسود البرية

Keyword: صحراء الكبرى
Clue: تقع إلى جنوب إفريقيا

Keyword: أمريكيتين

Clue: تتواجد الأسود فيهما

Use the following format:

Keyword: <Keyword>

Clue: <Crossword Clue>

English Prompt for Hallucination Verification:

"Please assess the quality of the crossword clues based on the given text.

Text: {text}

Clues: {clues}

To accomplish this task, follow these steps:

Check Clue in the text: Verify Whether the content of each clue is present in the text.

If a content clue is found in the text, print True; otherwise, print False.

Use the following format for each clue:

Check Clue in the text:

<Check Clue in the text>"

Arabic Keyword Generation Prompt:

الهدف: استخراج كلمات مفتاحية (تتكون من كلمتين على الأكثر) من الفقرة التالية لإستخدام هذه الكلمات المفتاحية لإنشاء تعريفات قصيرة من اجل لعبة الكلمات المتاقطعة تتأكد من استخراج اهم الكلمات المفتاحية من الفقرة ثم قم بعمل فحص لهذه الكلمات المتقاطعة اذا كان تم شرحها بشكل جيد و واضح في الفقرة واذا لم تجد شرح وافي لكلمة من الكلمات المفتاحية فقم بالتخلص منها

الفقرة:

{text}

مثال للمطلوب: هذه الفقرة التي قمت بإستخراج منها الكلمات المفتاحية الفقرة: ألأسد حيوان من الثدييات من فصيلة السنوريات وأحد السنوريات الأربعة الكبيرة المنتمية لجنس

تأكد من القيام بالسابق لكل لغز منفرد و طباعة النتيجة لكل لغز
قم بالتعامل مع كل لغز على حدى
استخدم الصيغة الأتية للنتيجة النهائية فقط قم بطباعة اللغز:النتيجةْ بدون اي شرح او اي شئ اخر
الصيغة النهائية:
اللغز: النتيجة

C Appendix

In the upcoming section, you will find English translations of the Arabic content within this paper. These translations have been included to improve understanding for readers who may have limited proficiency in Arabic, ultimately ensuring greater accessibility to the content. The translation for the Figure 3 content is as follows:

Input paragraph:

الذرة هي أصغر حجر بناءٍ أو أصغر جزء من العنصر الكيميائي يمكن الوصول إليه والذي يحتفظ بالخصائص الكيميائية لذلك العنصر. يرجع أصل الكلمة الإنجليزية إلى الكلمة الإغريقية أتوموس، والتي تعني غير القابل للانقسام؛ إذ كان يعتقد أنه ليس ثمة ما هو أصغر من الذرة. تتكون الذرة من سحابة من الشحنات السالبة (الإلكترونات) التي تدور حول نواة موجبة الشحنة صغيرة جدًا في المركز، وتتكون النواة من بروتونات موجبة الشحنة، ونيوترونات متعادلة، وتعدّ الذرة هي أصغر جزء من العنصر يمكن أن يتميز به عن بقية العناصر؛ إذ كلما غصنا أكثر في المادة لنلاقي البنى الأصغر لن يعود هناك فرق بين عنصر وآخر. فمثلاً، لا فرق بين بروتون في ذرة حديد وبروتون آخر في ذرة يورانيوم مثلًا، أو ذرة أي عنصرٍ آخر. الذرة، بما تحمله من خصائص؛ عدد بروتوناتها، كتلتها، توزيعها الإلكتروني... تصنع الفروقات بين العناصر المختلفة، وبين الصور المختلفة للعنصر نفسه (المسماة بالنظائر)، وحتى بين كَون هذا العنصر قادرًا على خوض تفاعل كيميائيماأملا.

English translation of the input paragraph:

The atom is the smallest building block or the smallest part of an element that can be reached and retains the chemical properties of that element. The English word 'atom' is derived from the Greek word 'atomos,' which means indivisible, as it was believed that there was nothing smaller than the atom. An atom consists of a cloud of negative charges (electrons) orbiting around a very small

جنوب الصحراء الكبرى، ولا تزال جمهرة واحدة صغيرة مهددة بالانقراض تعيش في آسيا بولاية غوجرات في شمال غربي الهند. كان موطن الأسود شاسعًا جدًا في السابق، حيث كانت تتواجد في شمال إفريقيا، الشرق الأوسط، وآسيا الغربية، حيث انقرضت منذ بضعة قرون فقط. وحتى بداية العصر الحديث (الهولوسين، منذ حوالي ١٠,٠٠٠ سنة)، كانت الأسود تُعتبر أكثر ثدييات اليابسة الكبرى انتشارا بعد الإنسان، حيث كانت توجد في معظم أنحاء إفريقيا، الكثير من أنحاء أوراسيا من أوروبا الغربية وصولا إلى الهند، وفي الأمريكيتين، من يوكون حتى البيرو.

ـ الكلمات المفتاحية كالأتي الكلمات المفتاحية: أسد، حيوان، ثدييات، سنوريات، سنوريات الأربعة، جنس النمور، ببر، الزكور الكبيرة، إفريقيا، صحراء الكبرى، أمريكيتين
هذه النتيجة:
الكلمة المفتاحية: أسد اللغز: حيوان ثديي من السنوريات
الكلمة المفتاحية: حيوان اللغز: ينتمي لفصيلة السنوريات
الكلمة المفتاحية: ثدييات اللغز: نوع من الحيوانات
الكلمة المفتاحية: سنوريات اللغز: تشمل الأسد
الكلمة المفتاحية: سنوريات الأربعة اللغز: مجموعة من السنوريات الكبيرة
الكلمة المفتاحية: جنس النمور اللغز: يعتبر الأسد منه
الكلمة المفتاحية: ببر اللغز: السنورية الأكبر في العالم
الكلمة المفتاحية: الزكور الكبيرة اللغز: تتجاوز وزنها ٢٥٠ كيلوغرام
الكلمة المفتاحية: إفريقيا اللغز: مكان عيش معظم الأسود البرية
الكلمة المفتاحية: صحراء الكبرى اللغز: تقع إلى جنوب إفريقيا
الكلمة المفتاحية: أمريكيتين اللغز: تتواجد الأسود فيهما
شكل النتيجة النهائية :
اللغز: اللغز
الكلمة المفتاحية: الكلمة المفتاحية

Arabic Prompt for Hallucination Verification:

قم بتقييم جودة الألغاز على حسب الفقرة الأتية
الفقرة: الفقرة
الألغاز: اللغز
لتقوم بهذه المهمة قم بالأتي:
قم بفحص اللغز في الفقرة. إذا كانت الفقرة تحتوي على كل من الألغاز. قم بطباعة صحيح و إذا لم تجده قم بطباعة خطأ

positively charged nucleus at the center. The nucleus is composed of positively charged protons and neutral neutrons. The atom is the smallest unit of an element that can be distinguished from the rest of the elements. As we delve deeper into matter, we find smaller structures, but there is no longer a distinction between one element and another. For example, there is no difference between a proton in an iron atom and another proton in a uranium atom, or in the atom of any other element. Atoms, with their properties, such as the number of protons, mass, and electronic distribution, create differences between different elements, between different forms of the same element (called isotopes), and even whether this element is capable of undergoing a chemical reaction or not.

Extracted keywords (English and Arabic):

Atom - ذرة
Chemical Element - العنصر الكيميائي
Chemical Properties - الخصائص الكيميائية
Electrons - الإلكترونات
Nucleus - النواة
Protons - البروتونات
Neutrons - النيوترونات
Elements - العناصر
Isotopes - النظائر
Chemical Reaction - تفاعلكيميائي

Generated Clue-Answer Pairs (English and Arabic):

الذرة: أصغر جزء من العنصر الكيميائي يمكن الوصول إليه
Atom: The smallest unit of a chemical element that can be reached.

العنصر الكيميائي: يتكون من الذرات ويحتفظ بالخصائص الكيميائية
Chemical Element: Composed of atoms and retains chemical properties.

الخصائص الكيميائية: يحتفظ بها العنصر الكيميائي
Chemical Properties: Characteristics retained by the chemical element.

الإلكترونات: تدور حول النواة في الذرة
Electrons: Orbits around the nucleus in the atom.

النواة : تدور حول النواة في الذرةتتكون من البروتونات والنيوترونات في الذرة
Nucleus: Composed of protons and neutrons in the atom.

البروتونات: تتواجد في النواة وتحمل شحنة موجبة
Protons: Located in the nucleus and carry a positive charge.

النيوترونات: تتواجد في النواة ولا تحمل شحنة
Neutrons: Located in the nucleus and carry no charge.

العناصر: تتكون من الذرات وتختلف بحسب عدد البروتونات في النواة
Elements: Composed of atoms and vary based on the number of protons in the nucleus.

النظائر: صور مختلفة للعنصر نفسه
Isotopes: Different forms of the same element.

تفاعل كيميائي: يمكن أن يخوضه العنصر بحسب خصائصهالكيميائية
Chemical Reaction: Can be undergone by the element based on its chemical properties.

Next, we will include the English translations for the examples we presented as both acceptable and unacceptable generated clues from the fine-tuned models in Tables 3 and 4.

Acceptable Examples(Arabic and English):

نجوم : في السماء ليلا
Stars: in the sky at night

كروم : من المعادن
Ores: from minerals

قوة: قدرة
Strength: capability

Unacceptable Examples(Arabic and English):

زرافة: من الحشرات
Giraffe: from the insects

مثلث : مثنى مثلث
Triangle: plural triangle

عمة : اخت والد او والدة
Aunt: sister of a parent or a parent's sister