Harnessing LLMs for Educational Content-Driven Italian Crossword Generation

Kamyar Zeinalipour[1,*,†], Achille Fusco[2,†], Asya Zanollo[1,†], Marco Maggini[1] and Marco Gori[1]

[1]University of Siena, DIISM, Via Roma 56, 53100 Siena, Italy
[2]IUSS Pavia, Piazza della Vittoria 15, 27100 Pavia (PV)

Abstract

In this work, we unveil a novel tool for generating Italian crossword puzzles from text, utilizing advanced language models such as `GPT-4o`, `Mistral-7B-Instruct-v0.3`, and `Llama3-8b-Instruct`. Crafted specifically for educational applications, this cutting-edge generator makes use of the comprehensive *Italian-Clue-Instruct* dataset, which comprises over 30,000 entries including diverse text, solutions, and types of clues. This carefully assembled dataset is designed to facilitate the creation of contextually relevant clues in various styles associated with specific texts and keywords. The study delves into four distinctive styles of crossword clues: those without format constraints, those formed as definite determiner phrases, copular sentences, and bare noun phrases. Each style introduces unique linguistic structures to diversify clue presentation. Given the lack of sophisticated educational tools tailored to the Italian language, this project seeks to enhance learning experiences and cognitive development through an engaging, interactive platform. By meshing state-of-the-art AI with contemporary educational strategies, our tool can dynamically generate crossword puzzles from Italian educational materials, thereby providing an enjoyable and interactive learning environment. This technological advancement not only redefines educational paradigms but also sets a new benchmark for interactive and cognitive language learning solutions.

Keywords

Large Language Models, Italian Educational Puzzles, Interactive Learning, Italian Educational Crosswords

1. Introduction

While traditionally valued for their challenge and entertainment, crossword puzzles are increasingly recognized for their educational benefits. They provide an interactive learning environment that enhances the retention of both technical terms and general language skills, hence facilitating learning across various disciplines, improving language acquisition, and supporting cognitive development, through critical thinking and memory retention [1, 2, 3, 4, 5, 6, 7, 3, 8, 9, 2, 10, 11].

The integration of Natural Language Processing (NLP) and Large Language Models (LLMs) has further enhanced their effectiveness by providing sophisticated, contextually relevant clues for educational crosswords.

This paper presents a novel tool that uses LLMs to generate tailored Italian educational crossword puzzles from texts, offering various clue types. By integrating user-provided texts or keywords and applying fine-tuning techniques, the tool produces high-quality clues and answers, offering educators a resource to develop more interactive and effective instructional methods.

Furthermore, a new dataset called [1] has been compiled and will be released to the scientific community.

The layout of this paper is organized in the following manner: Section 2 surveys the relevant literature in detail. Section 3 explains the methods used for dataset collection and curation. In Section 3, we describe the computational techniques employed in our study. Section 4 reports the results derived from our experimental analysis. Finally, Section 5 closes with conclusive insights and the broader implications of our research findings.

2. Related Works

Among the pioneering efforts in the field of crossword puzzle generation, Ranaivo et al. have formulated a distinctive strategy that merges text analytics with graph theory, allowing for the extraction and refinement of topic-specific clues through NLP [12]. Another notable contribution comes from Rigutini et al., who laid the groundwork by utilizing advanced NLP to automatically generate crossword puzzles from online sources, representing a seminal step in the field [13, 14].

In parallel, Esteche and his team have focused on Spanish-speaking audiences by creating puzzles with the aid of electronic dictionaries and news articles to formulate

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.
†These authors contributed equally.
kamyar.zeinalipour2@uisi.it (K. Zeinalipour); achille.fusco@iusspavia.it (A. Fusco); a.zanollo@student.unisi.it (A. Zanollo); marco.maggini@unisi.it (M. Maggini); marco.gori@unisi.it (M. Gori)
https://kamyarzeinalipour.github.io. (K. Zeinalipour)
0009-0006-3014-2511 (K. Zeinalipour); 0000-0002-5389-8884 (A. Fusco); 00000-0002-6428-1265 (M. Maggini); 0000-0001-6337-5430 (M. Gori)

[1]https://huggingface.co/datasets/Kamyar-zeinalipour/ita_cw_text

clues [15].
On a different front, Arora et al. developed SEEKH, a system that integrates statistical and linguistic analyses to generate crossword puzzles in multiple Indian languages. Their approach emphasizes the identification of keywords to structure the puzzles [16].
Recent progress in crossword puzzle generation has been notably advanced by the work of Zeinalipour et al. [17, 18, 19, 20], who demonstrated the use of large-scale language models to develop puzzles in languages with limited support, such as English, Italian and Arabic. Their research highlights the vast potential of computational linguistics in crafting puzzles that are both engaging and linguistically rich. Initially, they employed few-shot and zero-shot learning techniques to generate new crossword clues from text [18, 17].
Furthermore, Zugarini et al. [21] introduced a method for generating educational crossword clues from the provided text in English.
In their Italian crossword puzzle generation study [18], Zeinalipour et al. initially used few-shot learning with large language models as-is. However, our current project goes a step further by introducing a specially designed dataset for this task in Italian. Additionally, we have developed open-source models that have been fine-tuned to significantly enhance performance for this specific application.
The current research initiates a novel approach by utilizing state-of-the-art language modeling to develop Italian crossword puzzles from given texts. By doing so, it enriches the toolkit for language education, thereby pushing forward the development of Italian crossword puzzles.

3. Methodology

We have developed an automated system that generates educational Italian crossword puzzles using LLMs, with the *Italian-Clue-Instruct* dataset at its core. Our approach leverages the adaptability of LLMs, like GPT-4o, to create puzzles from text, with human validation for accuracy. Additionally, we fine-tuned models such as `Llama3-8b-Instruct` and `Mistral-7B-Instruct-v0.3` to improve clue accuracy and relevance.

A more detailed description of our methodology, illustrated in Figure 1, is provided in the following.

Italian-Clue-Instruct

Data Collection Methodology Initiating the data collection process, we began by extracting the introductory portions of Italian Wikipedia articles. We use Wikipedia API and Beautiful Soup to automatically extract the pages. The prominent focus was placed on the bolded keywords that highlight the primary topic and other significant terms within each article. Beyond keyword identification, we also gathered a variety of essential metadata. This included metrics such as view counts, relevance assessments, brief narrative summaries, central headlines, related terms, categorization, and URLs.[2] The uniform structure of the Italian Wikipedia significantly aids this process. By tapping into the introductory sections, which are particularly information-rich, we could systematically extract and outline the key concepts needed. This approach ensures a comprehensive data repository, capturing critical elements and insights from a diverse array of articles.

Data Enhancement To ensure the reliability and effectiveness of our data, we performed some filtering based on different criteria. The first filter was designed to prioritize the most important pages and those with the highest number of views. Firstly, articles were selected based on their popularity and relevance. To ensure a balanced and manageable dataset, we also discarded articles that were either too lengthy or too brief, specifically those with fewer than 50 words. Additionally, we removed keyword associations longer than two words to maintain the clarity and relevance of the crossword clues. Finally, we imposed restrictions on keywords to ensure they were between 3 and 20 characters in length and free of special characters or numerals. Multi-words expressions were also included as good keywords as they are quite common in crossword puzzles.

Formulation of Various Prompts Crafting specialized prompts was pivotal for producing Italian crossword clues from a given text using GPT-4o. The prompts were created to generate clues that were both informative and engaging, by incorporating crucial details and background context from the articles. Additionally, apart we aimed to elicit three specific types of clue varying in their syntactic structures:

- `definite determiner phrases`: nominal clues headed by a definite article and usually modified by adjectives, prepositional phrases (PPs) or relative clauses (RCs), like <*La repubblica asiatica con capitale Tashkent*, *Uzbekistan*> ('The Asian republic with Tashkent as capital', 'Uzbekistan'). Such clues are examples of definite descriptions which have been traditionally analyzed as carrying a uniqueness presupposition ([22]) when singular and a maximality presupposition [23] when plural. In the context of crosswords, clues of this kind refer to their solution as the single

[2] Wikipedia: Lists of popular pages by WikiProject

entity or the maximal plural entity satisfying the description.

- `bare noun phrases` [24]: the clue consists of a simple noun phrase (NP) with no determiner and typically modified by adjectives, PPs or RCs, for example <*Grande centro commerciale di lusso con sede a Londra, Harrods*> ('Luxury shopping mall based in London', 'Harrods'). In Italian, NPs are taken to denote a predicate that can be true of one or more individuals [22, 25].[3] Given the absence of the definite determiner, bare NP clues do not specify whether the referent of the solution uniquely satisfy the description [22], thus more than one solution could in principle be possible.
- `copular sentences` [26]: copular clues are clausal definitions structured as <*copula predicate*> with an elliptical subject as in <*è una salsa piccante tipica della Tunisia, Harissa*> ('(It) is a spicy sauce typical of Tunisia', 'Harissa'). Copulas, like Italian *essere* ('to be') connect a subject with a non-verbal predicate, such as an adjectival phrase (AP), a PP or another nominal phrase (NP/DP). In crossword puzzles, the solution targets the precopular position of such sentences, i.e. the elliptical subject. [4]

To accomplish this, we created three distinct prompts for each clue structure, and one prompt that does not specify the structure. This step allows us to test the syntactic sensitivity of the models employed and, more importantly it gives us the possibility of manipulating the structure to create variation not just with respect to the subject matter but also in the clue syntactic complexity. Moreover, generating clues with specific structures represents an interesting resource for the educational characterization of puzzles. Indeed, it is well-known from psycholinguistic research that different structures can elicitate different reactions in the processing which can be correlated with factors like age, linguistic disorders etc. and this can be exploited when creating puzzles specific for any solver's needs.

As for the prompt engeneering, the structure has been explicitated in one dedicated step of the prompt chain. For what regards the copular structure, which is widespread and widely used with different formulation, we include an example in the prompt (as shown in 9) to ensure that the required structure is given in output. It has been observed during the prompt trials that the validity of precise structures for clues strongly depends on the type of text given in input. The prompts used for clue generation in this study are presented in Figures 6, 7, 8 and 9, located in the Appendix.

Generation of Educational Italian Clues. Guided by the SELF-INSTRUCT framework [27], we devised a method to automate the generation of educational crossword clues in Italian, harnessing the power of LLMs. Central to our approach is the sophisticated GPT-4o[5], an enhanced version of LLMs, renowned for its efficiency. A key differentiator of our strategy is the integration of contextual information with the clues produced. To achieve this, we carefully curated the content and keywords from the Wikipedia text extracted in previous sections. We used four distinct types of prompts, each designed to generate different categories of clues: bare noun phrases, definite determiner phrases, and copular sentences. These prompts were crafted to create diverse types of clues, ensuring alignment with our specific objectives for educational content in Italian.

Overview of the *Italian-Clue-Instruct* Dataset Our research began with downloading 88,403 articles from the Italian Wikipedia, which we filtered down to 11,413 relevant entries. From this refined set, we selected 5,000 articles for clue generation, spanning 29 thematic categories. To enhance our dataset, we leveraged the capabilities of GPT-4o, generating a minimum of three diverse clues per Wikipedia article, depending on the text length. This effort resulted in a compilation of 15,000 unique clues.

The dataset's in-depth analysis demonstrates a variability in context length, ranging from 10 to 1512 tokens, with most texts falling between 100 and 600 tokens. Figure 2 showcases the token distribution for contexts and clues, which have been processed using the Llama3 tokenizer. Typically, the clue-generation process results in clues ranging from 4 to 55 tokens in length.
Figure 3 illustrates the spread of data across different categories. The dataset is notably dominated by the categories of "Entertainment", "Geography", and "History". In contrast, categories such as "Mathematics", "Architecture", and "Languages" are underrepresented.

Evaluating quality of the *Italian-Clue-Instruct* Dataset Producing accurate and engaging Italian 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.

[3] Bare NPs are known to denote also natural kinds [22]. However, given that NP clues occur in isolation, it is rather difficult to distinguish among the two senses, therefore we assume the more general reading of NPs as predicates. We leave this discussion to future analyses.

[4] Copular sentences are known to be differentiated between canonical and inverse structures [26]. Usually in crossword clues canonical structure are found more frequently, but inverse copular clues are not excluded. We leave the question open for further, purely linguistic research.

[5] https://openai.com/index/hello-gpt-4o/

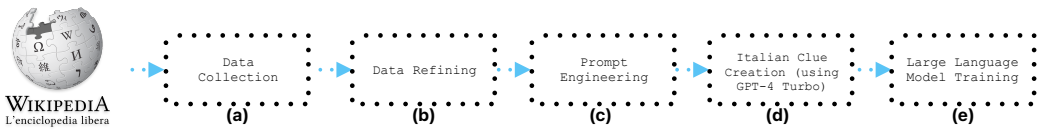

Figure 1: The methodology followed in this study comprises the following stages: (a) Gathering an extensive dataset from the Italian Wikipedia. (b) Refining and filtering the data by eliminating entries that are either too brief or excessively detailed, thereby optimizing its quality. (c) Developing specialized prompts intended to create educational Italian crossword clues derived from the curated dataset. (d) Utilizing GPT-4o to generate Italian crossword clues based on the processed data and crafted prompts. (e) Fine-tuning Large Language Models (LLMs) to enhance their performance in producing contextual and tailored Italian crossword clues. These systematic steps ensure the effective leveraging of advanced natural language processing technologies to create high-quality educational content in the form of Italian crossword clues.

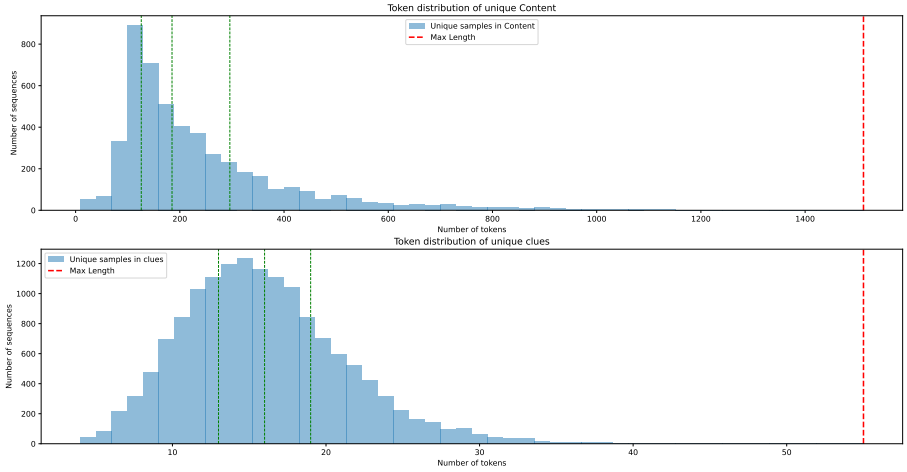

Figure 2: Token Distributions for Context and Clues of *Italian-Clue-Instruct*

Our evaluation strategy adapts 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 that we extracted from Wikipedia. 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-1, ROUGE-2, and ROUGE-L score of 0.159, 0.114, and 0.146 respectively.

Considering that the ROUGE score merely compares the similarity between the n-grams of the generated clues and the reference text from Wikipedia, it is not a reliable metric and does not provide any assessment of the semantic quality of the generated clues. However, it provides a general picture of the generated clues.

In addition, the integrity of the generated clues was further examined through human evaluations. A randomly chosen subset of clues was assessed, generated from a sample of 100 articles, with a maximum of three clues per article. To avoid repetitions, duplicate clues were removed. The evaluation employed a five-level criteria system, analogous to the methodology utilized by [27]. For the present evaluation, the following parameters were used:

- RATING-A: The clue is coherent and valid, align-

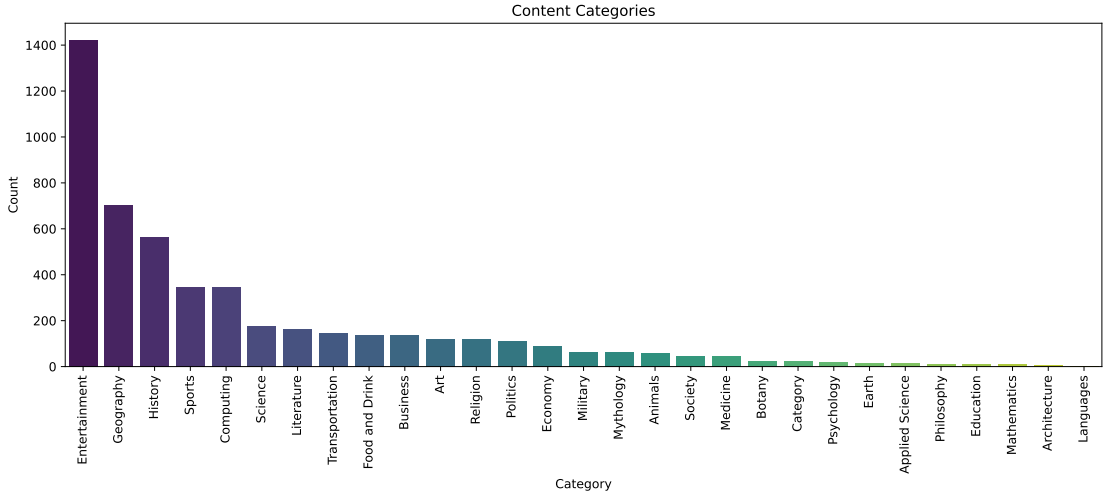

Figure 3: Bar Plot Showing the Frequency of Different Categories within the Dataset.

ing correctly with the given context, answer, and specified structure.

- `RATING-B`: This clue, while generally acceptable, exhibits slight discrepancies mainly due to suboptimal phrasing or structure.
- `RATING-C`: The clue relates directly to the answer but retains a vague connection to the context or provide information which, even if correct, is not properly conveyed.
- `RATING-D`: The clue is strictly referring to the context and fails to comprehensively identify the answer.
- `RATING-E`: The clue is deemed unacceptable because it is ungrammatical, it directly contains the answer or a variation of it, or doesn't identify the referent of the answer.

The evaluation was made by a native Italian speaker, master student of linguistics, and PhD student in linguistics, who followed the criteria described above. Please refer to Table 2 for examples of clues and their respective ratings.

The distribution of the evaluation outcomes is depicted in Figure 4, these illustrate that the majority of the generated clues were of high quality rated as 'A' and only a small fraction rated as 'C', 'D', or 'E'.

By utilizing both quantitative metrics and qualitative assessments, the study aimed to validate the educational utility and contextual accuracy of the clues created for Italian educational crosswords.

Enhancing LLMs for Italian text-based Educational Crossword Puzzle Generation To develop crossword

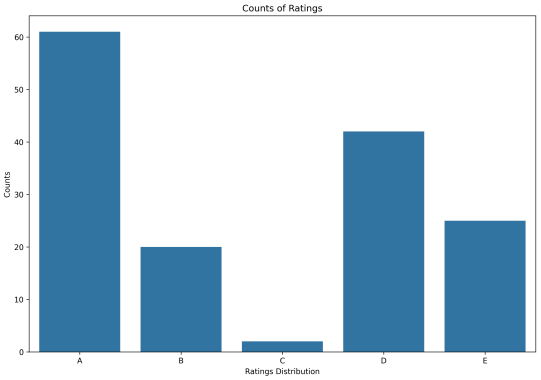

Figure 4: Bar Plot Showing the Frequency of GPT-4o Ratings

puzzle clues from Italian texts using advanced LLM functionalities, we employed three models: `GPT-4o` (for data generation), `Mistral-7B-Instruct-v0.3`, and `Llama3-8b-Instruct` known for their strong text generation and Italian language support. [28, 29].
We began the process by fine-tuning the models with the *Italian-Clue-Instruct* dataset, which was rich in relevant material. This calibration was vital to enhance the models' proficiency in generating Italian clues while accurately reflecting the Italian language's intricate grammar and vocabulary within educational contexts.
To further refine the models, we optimized the parameters during the fine-tuning phase. This effort aimed to reduce errors specific to our task and better align the output of the models with Italian educational materials. Ultimately, the specialized tuning of these LLMs with a

dedicated dataset was intended to foster their ability to generate high-quality crossword clues from Italian texts. The goal was to ensure that the resulting clues were not only linguistically sound but also relevant within an educational framework.

4. Experimental Results

This section offers a detailed overview of the experiments conducted in the study. It begins with the training setup for the *Italian-Clue-Instruct* LLMs, including key parameters and computational resources. The performance of the models is then evaluated using automated metrics, such as the ROUGE score, to compare configurations and identify areas for improvement. This is followed by an in-depth analysis of human evaluations, focusing on relevance, coherence, and content quality to provide insights beyond automated metrics. Additionally, an example of a generated crossword puzzle is presented to demonstrate practical usability. The goal is to highlight the robustness and versatility of the proposed approach.

Training Setup The models `Mistral-7B-Instruct-v0.3` and `Llama3-8b-Instruct` were fine-tuned using LORA [30], with parameters set to $r = 16$ and $\alpha = 32$, across three training epochs, maintaining a total batch size of 64. The full experimental setup was performed on a server equipped with four NVIDIA A6000 GPUs, utilizing DeepSpeed [31] and FlashAttention 2 [32]. For the initial learning rate was configured at 3×10^{-4}. During inference, model distribution sampling was applied to generate clues for both `Mistral-7B-Instruct-v0.3` and `Llama3-8b-Instruct`, with a temperature parameter set to 0.1. Additionally, the parameters for top-p and top-k sampling were set to 0.95 and 50, respectively. Among the three epoch checkpoints, the one with the minimum loss was selected, which, in our case, turned out to be the second checkpoint.

Evaluation Results with the Automatic Metrics We evaluated the resemblance between various sets of clues produced by different models (details shown in Table 1) and those generated by the `GPT-4o` model on a test set of 200 educational contexts. This evaluation was done using ROUGE scores. Our results indicate that the fine-tuned `Mistral-7B-Instruct-v0.3` and `Llama3-8b-Instruct` models exhibit a closer similarity to `GPT-4o`. On the other hand, the base `Llama3-8b-Instruct` model shows significantly lower similarity with minimal overlap. These outcomes highlight the efficacy of fine-tuning, demonstrating that using the *Italian-Clue-Instruct* dataset enhances the capability of `Mistral-7B-Instruct-v0.3` and `Llama3-8b-Instruct` models in generating clues from Italian educational texts.

Evaluation Results with the human evaluator Using a dataset of 100 Italian contexts, each containing 3 clues, a human evaluation was conducted on both the generated and base models. The results of this evaluation are depicted in Figure 5. The evaluation employed the 5-level rating system described in Section 3.

The table provided offers a comparative evaluation of the performance of language models in generating Italian clues from a given text. Specifically, the models `Mistral-7B-Instruct-v0.3` and `Llama3-8b-Instruct` are evaluated based on both their base and fine-tuned configurations. Upon fine-tuning, `Mistral-7B-Instruct-v0.3` displays a significant improvement, emerging as the top performer in category "A", and surpassing `Llama3-8b-Instruct` in terms of performance enhancement. These findings underscore the impact of fine-tuning on enhancing model capabilities, particularly highlighted by the performances of `Mistral-7B-Instruct-v0.3` and `Llama3-8b-Instruct`, which feature 7 and 8 billion parameters, respectively. Furthermore, fine-tuning with the introduced dataset significantly increased the models' ability to generate Italian clues from the given text, illustrating the quality and effectiveness of the *Italian-Clue-Instruct* dataset.

The methodology for generating Italian crossword clues from educational texts was explored, enabling customized clues. This would allow educators to select suitable clues matching their teaching needs. The selected clues could in turn be used to automatically generate a crossword schema as discussed Zeinalipour et al. [17]. Figure 10 in Appendix shows an example puzzle, demonstrating the system's application.

5. Conclusion

A novel system for generating crossword clues from Italian text is introduced, leveraging the newly developed *Italian-Clue-Instruct* dataset. This dataset, which includes text, keywords, categories, and related crossword clues in Italian, is pioneering in this field. By fine-tuning two large language models (LLMs), `Mistral-7B-Instruct-v0.3` and `Llama3-8b-Instruct`, using this dataset, we have achieved significant improvements in the models' ability to generate crossword clues from given text. The results highlight a substantial enhancement in model performance after fine-tuning. Both the *Italian-Clue-Instruct* dataset and the fine-tuned models are now publicly available, providing valuable tools for students and teachers to create educational crossword puzzles from Italian text.

Model	Model name	ROUGE-1	ROUGE-2	ROUGE-L
Base LLMs	Mistral-7B	0.342	0.176	0.261
	Llama3-8b	0.258	0.112	0.198
Fine-tuned LLMs	Mistral-7B	**0.611**	**0.458**	**0.556**
	Llama3-8b	0.552	0.403	0.501

Table 1
Mean ROUGE Scores for Various Comparisons with GPT-4o generated clues

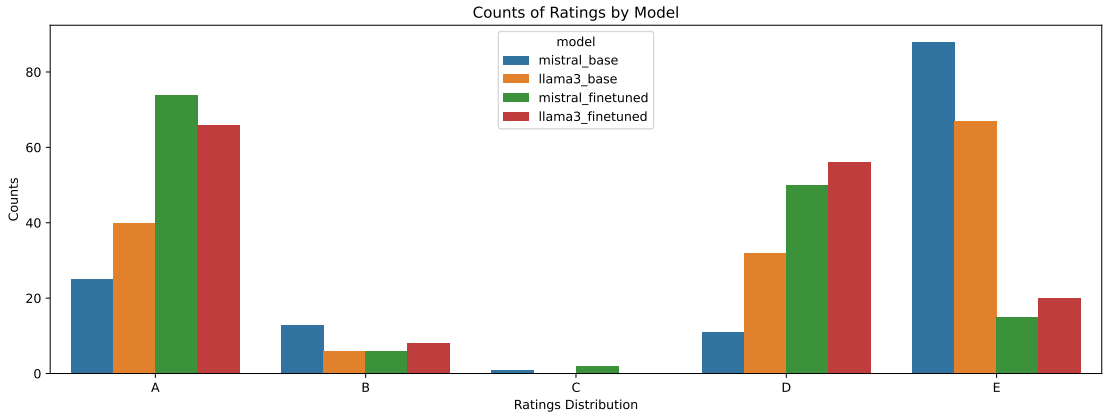

Figure 5: Bar Plot Showing the Frequency of the ratings after the evaluation.

Future research will aim to develop models capable of generating various types of crossword clues, including fill-in-the-blank clues.

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

[1] W. Orawiwatnakul, Crossword puzzles as a learning tool for vocabulary development, Electronic Journal of Research in Education Psychology 11 (2013) 413–428.

[2] Y. D. Bella, E. M. Rahayu, The improving of the student's vocabulary achievement through crossword game in the new normal era, Edunesia: Jurnal Ilmiah Pendidikan 4 (2023) 830–842.

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

[4] R. Nickerson, Crossword puzzles and lexical memory, in: Attention and performance VI, Routledge, 1977, pp. 699–718.

[5] E. Yuriev, B. Capuano, J. L. Short, Crossword puzzles for chemistry education: learning goals beyond vocabulary, Chemistry education research and practice 17 (2016) 532–554.

[6] C. Sandiuc, A. Balagiu, The use of crossword puzzles as a strategy to teach maritime english vocabulary, Scientific Bulletin" Mircea cel Batran" Naval Academy 23 (2020) 236A–242.

[7] S. Kaynak, S. Ergün, A. Karadaş, 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 (2023) 103618.

[8] S. T. Mueller, E. S. Veinott, Testing the effectiveness of crossword games on immediate and delayed memory for scientific vocabulary and concepts., in: CogSci, 2018.

[9] V. S. Zirawaga, A. I. Olusanya, T. Maduku, Gaming in education: Using games as a support tool to teach history., Journal of Education and Practice 8 (2017) 55–64.

[10] P. Zamani, S. B. Haghighi, M. Ravanbakhsh, The use of crossword puzzles as an educational tool, Journal of Advances in Medical Education & Professionalism 9 (2021) 102.

[11] S. M. Dol, Gpbl: An effective way to improve critical

thinking and problem solving skills in engineering education, J Engin Educ Trans 30 (2017) 103–13.
[12] B. Ranaivo-Malançon, T. Lim, J.-L. Minoi, A. J. R. Jupit, Automatic generation of fill-in clues and answers from raw texts for crosswords, in: 2013 8th International Conference on Information Technology in Asia (CITA), IEEE, 2013, pp. 1–5.
[13] L. Rigutini, M. Diligenti, M. Maggini, M. Gori, A fully automatic crossword generator, in: 2008 Seventh International Conference on Machine Learning and Applications, IEEE, 2008, pp. 362–367.
[14] L. Rigutini, M. Diligenti, M. Maggini, M. Gori, Automatic generation of crossword puzzles, International Journal on Artificial Intelligence Tools 21 (2012) 1250014.
[15] J. Esteche, R. Romero, L. Chiruzzo, A. Rosá, Automatic definition extraction and crossword generation from spanish news text, CLEI Electronic Journal 20 (2017).
[16] B. Arora, N. Kumar, Automatic keyword extraction and crossword generation tool for indian languages: Seekh, in: 2019 IEEE Tenth International Conference on Technology for Education (T4E), IEEE, 2019, pp. 272–273.
[17] K. Zeinalipour, T. Iaquinta, G. Angelini, L. Rigutini, M. Maggini, M. Gori, Building bridges of knowledge: Innovating education with automated crossword generation, in: 2023 International Conference on Machine Learning and Applications (ICMLA), IEEE, 2023, pp. 1228–1236.
[18] K. Zeinalipour, A. Zanollo, G. Angelini, L. Rigutini, M. Maggini, M. Gori, et al., Italian crossword generator: Enhancing education through interactive word puzzles, arXiv preprint arXiv:2311.15723 (2023).
[19] K. Zeinalipour, M. Saad, M. Maggini, M. Gori, Arabicros: Ai-powered arabic crossword puzzle generation for educational applications, in: Proceedings of ArabicNLP 2023, 2023, pp. 288–301.
[20] K. Zeinalipour, Y. G. Keptiğ, M. Maggini, L. Rigutini, M. Gori, A turkish educational crossword puzzle generator, in: International Conference on Artificial Intelligence in Education, Springer, 2024, pp. 226–233.
[21] A. Zugarini, K. Zeinalipour, S. S. Kadali, M. Maggini, M. Gori, L. Rigutini, Clue-instruct: Text-based clue generation for educational crossword puzzles, arXiv preprint arXiv:2404.06186 (2024).
[22] G. Chierchia, Reference to kinds across language, Natural language semantics 6 (1998) 339–405.
[23] G. Link, The logical analysis of plurals and mass terms: A lattice theoretical approach, Meaning, Use, and Interpretation of Language/Walter de Gruyter (1983).
[24] G. Longobardi, Reference and proper names: A theory of n-movement in syntax and logical form, Linguistic inquiry (1994) 609–665.
[25] Z. Roberto, Layers in the determiner phrase, Ph.D. thesis, PhD Thesis, University of Rochester (Published by Garland, 2000), 1995.
[26] A. Moro, Copular sentences, The Blackwell companion to syntax (2006) 1–23.
[27] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language model with self generated instructions, arXiv preprint arXiv:2212.10560 (2022).
[28] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[29] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
[30] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).
[31] J. Rasley, S. Rajbhandari, O. Ruwase, Y. He, 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, 2020, pp. 3505–3506.
[32] T. Dao, Flashattention-2: Faster attention with better parallelism and work partitioning, arXiv preprint arXiv:2307.08691 (2023).

A. Appendix

```
You are a crossword expert.
Generate concise and clever clues in Italian for educational crossword puzzles based on a specified Keyword and its relation to an
assigned Text. To execute this task properly, replicate the guidelines below:

KEYWORD: {keyword}
TEXT: {text}

Observe the following steps:
1. Substitute every pronoun in the text with full phrases expressing their referents.
2. Split the text into small independent sentences that could be understood out of context.
3. Pinpoint three concise sentences that contain the Keyword and best characterize the keyword. Try to select sentences from
different parts of the Text.
4. Generate short and clever crossword clues in Italian from the selected sentences. Make sure that  the keyword remains absent
from the clues. If the Keyword is not the subject of the sentence, make sure that it is substituted with an appropriate clitic,
possessive or demonstrative pronoun. Generate clues from all the parts of the text and use all of the information provided to
generate the clues.
5. Ensure that each clue functions as a description or definition of the keyword rather than a query, focusing on details about
the keyword.
6. Make sure that each clue's information can be traced back to the text. Make sure that the clues are relevant and that they are
sufficient to identify the keyword. Make sure that the keyword does not appear in the clues. Make sure that any part of the
keyword is not present in the clues.
7. Select only the three best clues for educational purposes.
8. Compile these clues into a list formatted as follows: [clue1, clue2, clue3]  into a JSON file under the key: 'clues'. Make sure
the output is in the requested format and do not include the whole process in the output, but only the clues.
```

Figure 6: Illustration of the prompt used for unrestricted format clues in the research.

```
You are a crossword expert.
Generate concise and clever clues in Italian for educational crossword puzzles based on a specified Keyword and its relation to an
assigned Text. To execute this task properly, replicate the guidelines below:
KEYWORD: {keyword}
TEXT: {text}

Observe the following steps:
1. Substitute every pronoun in the text with full phrases expressing their referents.
2. Split the text into small independent sentences that could be understood out of context.
3. Pinpoint three concise sentences that contain the Keyword and best characterize the keyword. Try to select sentences from
different parts of the Text.
4. Generate short and clever crossword clues in Italian from the selected sentences. Make sure that the keyword remains absent
from the clues. Each clue must have the syntax of a bare noun phrase (zero determiner): the root node of each clue must be a
common or proper noun and it can be followed by a relative clause or other complements or adjuncts.  Generate clues from all the
parts of the text and use all of the information provided to generate the clues.
5. Ensure that each clue functions as a description or definition of the keyword rather than a query, focusing on details about
the keyword.
6. Make sure that each clue's information can be traced back to the text. Make sure that the clues are relevant and that they are
sufficient to identify the keyword. Make sure that the keyword does not appear in the clues. Make sure that any part of the
keyword is not present in the clues.
7. Select only the three best clues for educational purposes.
8. Compile these clues into a list formatted as follows: [clue1, clue2, clue3]  into a JSON file under the key: 'clues'. Make sure
the output is in the requested format and do not include the whole process in the output, but only the clues.
```

Figure 7: Illustration of the prompt used for noun phrases format clues in the research.

```
You are a crossword expert.
Generate concise and clever clues in Italian for educational crossword puzzles based on a specified Keyword and its relation to an
assigned Text. To execute this task properly, replicate the guidelines below:
KEYWORD: {keyword}
TEXT: {text}

Observe the following steps:
1. Substitute every pronoun in the text with full phrases expressing their referents.
2. Split the text into small independent sentences that could be understood out of context.
3. Pinpoint three concise sentences that contain the Keyword and best characterize the keyword. Try to select sentences from
different parts of the Text.
4. Generate short and clever crossword clues in Italian from the selected sentences. Make sure that  the keyword remains absent
from the clues. Each clue must have the syntax of a determiner phrase with the definite article (followed by a noun and possibly
adjectives). It can be followed by a relative clause or other complements or adjuncts. Generate clues from all the parts of the
text and use all of the information provided to generate the clues.
5. Ensure that each clue functions as a description or definition of the keyword rather than a query, focusing on details about
the keyword.
6. Make sure that each clue's information can be traced back to the text. Make sure that the clues are relevant and that they are
sufficient to identify the keyword. Make sure that the keyword does not appear in the clues. Make sure that any part of the
keyword is not present in the clues.
7. Select only the three best clues for educational purposes.
8. Compile these clues into a list formatted as follows: [clue1, clue2, clue3]  into a JSON file under the key: 'clues'. Make sure
the output is in the requested format and do not include the whole process in the output, but only the clues.
```

Figure 8: Illustration of the prompt used for determiner phrases format clues in the research.

```
Generate concise and clever clues in Italian for educational crossword puzzles based on a specified Keyword and its relation to an
assigned Text. To execute this task properly, replicate the guidelines below:
KEYWORD: {keyword}
TEXT: {text}

Observe the following steps:
1. Substitute every pronoun in the text with full phrases expressing their referents.
2. Split the text into small independent sentences that could be understood out of context.
3. Pinpoint three concise sentences that contain the Keyword and best characterize the keyword. Try to select sentences from
different parts of the Text.
4. Generate short and clever crossword clues in Italian from the selected sentences. Make sure that  the keyword remains absent
from the clues. Each clue must be a copular sentence, in which the keyword constitutes the subject. The syntax of each clue then
must corresponds to a copular sentence without the subject. For example: "è <clue>". Generate clues from all the parts of the text
and use all of the information provided to generate the clues.
5. Ensure that each clue functions as a description or definition of the keyword rather than a query, focusing on details about
the keyword.
6. Make sure that each clue's information can be traced back to the text. Make sure that the clues are relevant and that they are
sufficient to identify the keyword. Make sure that the keyword does not appear in the clues. Make sure that any part of the
keyword is not present in the clues.
7. Select only the three best clues for educational purposes.
8. Compile these clues into a list formatted as follows: [clue1, clue2, clue3]  into a JSON file under the key: 'clues'. Make sure
the output is in the requested format and do not include the whole process in the output, but only the clues.
```

Figure 9: Illustration of the copular sentences prompt used for copular sentences format clues in the research.

Clue	Answer	Rating	Explanation
È il sesto album in studio del gruppo rock inglese The Who 'It's the sixth studio album by English rock band The Who'	Quadrophenia	A	
Il distretto con status di borough del Lancashire 'The district with the status of borough of Lancashire'	South Ribble	B	Definite determiner is not appropriate: there are other boroughs in Lancashire.
Duo composto da Hayley Williams e Taylor York fino al 2017 'Duo composed by Hayley Williams and Taylor York until 2017'	Paramore	C	The clue provides accurate but incomplete information: the band was a duo for a limited period.
Gruppo musicale statunitense 'American music band'	Pixies	D	The clue is too generic.
Terrier di proporzioni minuscole, cacciatore eccezionale 'Terrier of minuscule proportions, excellent hunter'	Patterdale Terrier	E	The clue contains part of the answer.

Table 2
Examples of evaluation ratings

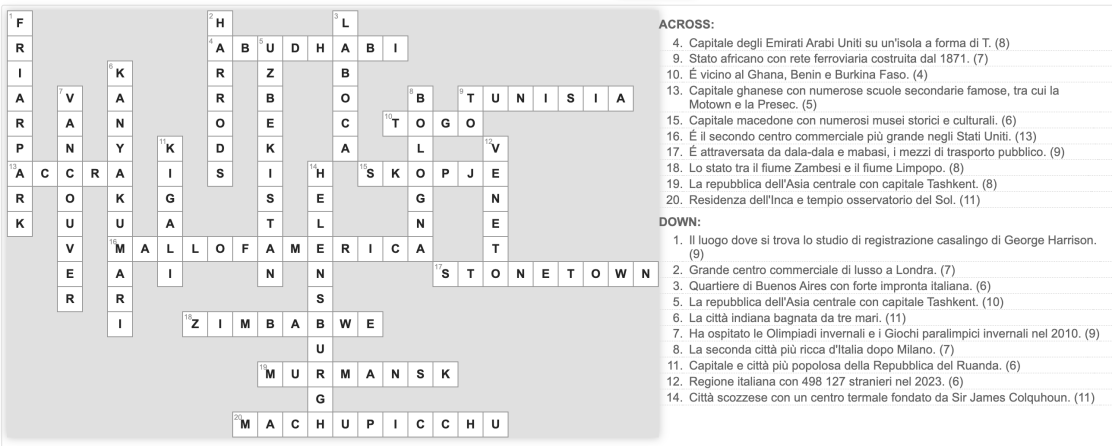

Figure 10: Crossword crafted using the proposed system.