Clue-Instruct: Text-Based Clue Generation for Educational Crossword Puzzles

Andrea Zugarini[1], Kamyar Zeinalipour[2], Surya Sai Kadali[2],
Marco Maggini[2], Marco Gori[2], Leonardo Rigutini[1]
[1]expert.ai, Siena, Italy
[2]University of Siena, Italy
{azugarini, lrigutini}@expert.ai, suryasai.kadali@student.unisi.it,
{kamyar.zeinalipour2, marco.maggini, marco.gori}@unisi.it

Abstract

Crossword puzzles are popular linguistic games often used as tools to engage students in learning. Educational crosswords are characterized by less cryptic and more factual clues that distinguish them from traditional crossword puzzles. Despite there exist several publicly available clue-answer pair databases for traditional crosswords, educational clue-answer pairs datasets are missing. In this article, we propose a methodology to build educational clue generation datasets that can be used to instruct Large Language Models (LLMs). By gathering from Wikipedia pages informative content associated with relevant keywords, we use Large Language Models to automatically generate pedagogical clues related to the given input keyword and its context. With such an approach, we created `clue-instruct`, a dataset containing 44,075 unique examples with text-keyword pairs associated with three distinct crossword clues. We used `clue-instruct` to instruct different LLMs to generate educational clues from a given input content and keyword. Both human and automatic evaluations confirmed the quality of the generated clues, thus validating the effectiveness of our approach.

Keywords: Educational Crossword Clues, LLMs, Instruction Tuning, Natural Language Generation

1. Introduction

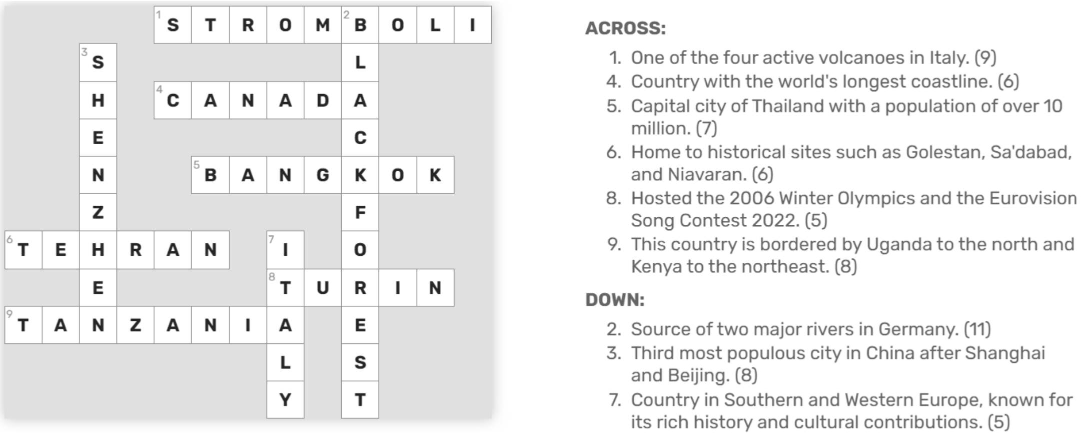
The conventional structure of crossword puzzles merged with scholastic elements results in an engaging learning tool: educational crosswords. They encompass a variety of subjects such as science, vocabulary, and history (Nickerson, 1977; Sandiuc and Balagiu, 2020; Yuriev et al., 2016). Educational crosswords differ from traditional puzzles because they are designed for teaching rather than entertainment. Consequently, their are less cryptic, and usually, they present a less constrained puzzle scheme, as in the example shown in Figure 1. They are particularly beneficial in language acquisition or when mastering technical jargon for specific topics (Orawiwatnakul, 2013; Dzulfikri, 2016; Bella and Rahayu, 2023). Additionally, the requirement of correlating appropriate hints with correct words fosters learners' problem-solving skills (Kaynak et al., 2023; Dol, 2017). Memory enhancement is another merit of educational crosswords, as learners need to summon previously learned material to solve the puzzle (Mueller and Veinott, 2018; Dzulfikri, 2016). Moreover, the interactive nature of crosswords makes the learning experience captivating, inducing learners to persist in honing their abilities (Zirawaga et al., 2017; Bella and Rahayu, 2023). Summarily, educational crosswords serve as an entertaining resource for strengthening educational skills (Zamani et al., 2021; Yuriev et al., 2016).

Harnessing the power of Large Language Models (LLMs) presents an opportunity in the field of educational crossword production, traditionally known for requiring specialized skills and labor. Through an extensive training process on huge language corpora comprising internet resources, academic papers, and books, LLMs acquire the ability to generate high-quality text to accomplish many different tasks. This proficiency can be exploited to automatically generate clues, so to ease the process of educational crossword crafting.

In this work, we propose a methodology to construct datasets for educational crossword clue generation. In particular, we present `clue-instruct`, a corpus made of 44,075 clue generation instructions. Each example is constituted by a source text, serving as context, a category of interest, and a keyword, all paired with three target clues to generate. The dataset is built by gathering content from Wikipedia pages about relevant keywords, whereas clues were automatically generated by an LLM. Upon `clue-instruct`, we carried out a detailed experimentation with different open-source LLMs varying in size and family, and we fine-tune them on the dataset. Results, assessed with both automatic and human evaluations, indicate that fine-tuning remarkably improves the generation quality of those models. The dataset[1] and all the models are publicly available.

The paper is organized as follows. Section 2 reports the related works on crosswords in NLP. We describe the proposed methodology in Section 3 and analyse in detail the properties of the gener-

[1] `https://huggingface.co/datasets/azugarini/clue-instruct`



ACROSS:

1. One of the four active volcanoes in Italy. (9)
4. Country with the world's longest coastline. (6)
5. Capital city of Thailand with a population of over 10 million. (7)
6. Home to historical sites such as Golestan, Sa'dabad, and Niavaran. (6)
8. Hosted the 2006 Winter Olympics and the Eurovision Song Contest 2022. (5)
9. This country is bordered by Uganda to the north and Kenya to the northeast. (8)

DOWN:

2. Source of two major rivers in Germany. (11)
3. Third most populous city in China after Shanghai and Beijing. (8)
7. Country in Southern and Western Europe, known for its rich history and cultural contributions. (5)

Figure 1: Example of an educational crossword puzzle on Geography-related keywords.

ated dataset in Section 4. In Section 5, we discuss the experimental outcomes in-depth. Finally, we draw our conclusions in Section 6.

2. Related Works

Crossword puzzles are a fascinating linguistic game that has been a subject of study in the Natural Language Processing field in the past few years. Literature can be divided into two main research branches: crossword solving and crossword generation (Rigutini, 2010). We briefly review both of them, then we finally discuss about existing crossword datasets.

Crossword solving. Crossword resolution can be tackled as a constrained satisfaction task where the objective is to maximize the probability of filling the grid with answers coherent with the given clues. The main challenge in the problem is retrieving correct candidate answers. Existing solutions heavily rely on clue-answer databases. Proverb (Littman et al., 1999), one of the earliest crossword-solving systems, used a probabilistic version of the A* with candidate answers retrieved from databases of American crosswords. Similarly, Dr. Fill (Ginsberg, 2011) converted them into weighted CSPs and used advanced heuristics. WebCrow (Ernandes et al., 2005; Angelini et al., 2005b,a) was a crossword-solving Italian project based on human-machine competitions. Webcrow distinguished from other solutions for exploiting the information present in the web. It was developed for the Italian language and English. Recently, it was extended to other languages (Angelini et al., 2023; Zugarini et al., 2023) making use of neural representations of clue-answer pairs (Zugarini and Ernandes, 2021). Based on WebCrow, SACRY leveraged syntactic structures for re-ranking and answer extraction to enhance answer quality by incorporating syntactic analysis (Barlacchi et al., 2015). Lately, the Berkeley Crossword Solver (Wallace et al., 2022) was presented. It was based on neural question-answering models for candidate answer retrieval, and belief propagation with local searches to fill the grid, achieving state-of-the-art performance in English crossword solving.

```
You are a crosswords expert.

Generate short and clever definitions for
crosswords, based on a given keyword, a
category and a keyword-related context,
following the instructions provided below.

        KEYWORD: {keyword}

        CATEGORY: {category}

        CONTEXT: {text}

Follow these steps:

1. Find parts of the given context related
to the {keyword} and {category}.

2. Select three key pieces of information
related to {keyword} and {category} that
are present in the context.

3. Create short clues from these key
facts, making sure not to include the
keyword in the clues.

4. Put these clues into a JSON file under
the key: 'clues'.
```

Figure 2: `clue-instruct` prompt used to generate the clues.

Crossword Generation. Building a crossword puzzle automatically encompasses different linguistic problems, such as identifying the answers, composing the grid, and above all, creating the clues. Early approaches Rigutini et al. (2008, 2012) leveraged NLP techniques to generate lists of clue-answer pairs by analyzing online documents

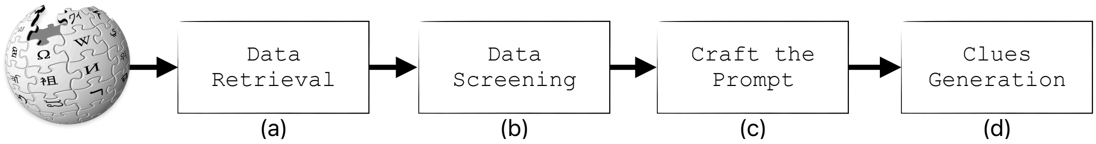


Figure 3: This figure illustrates the pipeline employed in constructing `clue-instruct`: (a) Information extraction from Wikipedia pages of text, keywords, and categories. (b) Data refinement and filtering to enhance data quality, by selecting the most crucial and highly-viewed pages, eliminating excessively short or overly detailed text, and more. (c) Design of the prompt for generating crossword clues based on input text and specified keywords within specific categories. (d) Exploit of GPT-3.5-Turbo to generate clues from the collected data and defined prompts.

(Wikipedia pages). Clues were identified and extracted with NLP techniques (POS tagging, dependency analysis, and WordNet). Analogously, the methods proposed in Ranaivo-Malançon et al. (2013) and Esteche et al. (2017) followed a step-based approach to construct crosswords via NLP tools. The steps involved preliminary data extraction of sentences from a text that was then used to produce the clue-answer pairs. A software tool utilizing NLP techniques to extract crucial keywords for crossword creation in Indian languages was proposed by Arora and Kumar (2019). The resulting SEEKH framework combines statistical and linguistic methods to identify vital keywords. More recently, Zeinalipour et al. (2023a,b,c) moved from handcrafted design of generated crosswords to generative solutions utilizing pre-trained LLMs. Crosswords were generated in English, Arabic, and Italian, thus demonstrating the effectiveness of computational linguistics in creating culturally diverse and engaging puzzles. Analogously, our method makes use of LLMs to generate clues for a given answer, but we ground the generation to a source context with the purpose of producing clues that are adherent to a given input text.

Clue-Answer datasets. Despite many works have been published in both crossword solving and generation, few datasets have been created and publicly released. Most of them consist of clue-answer pair corpora, generally collected from crosswords or clue databases (Ernandes et al., 2008; Ginsberg, 2011) sometimes enriched by metadata such as publication date, publisher, and difficulty. Unfortunately, for copyright reasons, they are not always publicly available. In (Barlacchi et al., 2015), to test their proposed system, the authors created a corpus by downloading crossword puzzles from some web sources. Wallace et al. (2022) collected a validation and test set of complete 2020 and 2021 puzzle grids from several US news (The New York Times, The LA Times, Newsday, The New Yorker, and The Atlantic) and they publicly released code, models, and dataset. However, all these clue-answer pairs corpora are constructed from traditional crossword puzzles. In these types of puzzles, the clues usually have extremely enigmatic linguistic structures that are quite different from those typically adopted for educational purposes. Furthermore, by design they lack of any reference to textual passages in which the clue can be found. This information is very important in the educational use-case where the clue must be related to a subject of study. Moreover, a grounding context allow to steer the generation of a Language Model, thus dramatically reducing the occurrence of hallucinated or unrelated clues.

In this work instead, we propose a method to create a clue generation corpus where clues are tied with an answer and a source context. The obtained dataset is, to the best of our knowledge, the first corpus associating such information together.

3. Method

Differently from traditional clue-answer crossword databases, we necessitate aligning the clue-answer pair with a grounding text, where the answer to the clue can be inferred from it. The grounding text is crucial in education both from the perspective of a teacher and from the point of view of the student. In order to construct such a context-keyword-clue triplet, we follow a pipeline, starting from collecting and gathering data from Wikipedia. The entire pipeline is sketched in Figure 3. Here we describe it step by step.

Data Retrieval. We initiate the information extraction process by mining Wikipedia pages. This involves accessing the initial section of each page, which typically contains the most pertinent information. From this portion, we emphasize keywords presented in bold, which often correspond to the page's title but can include additional terms. These selected keywords become the focal points of the Wikipedia page, shaping the content to provide in-depth definitions and explanations. In addition to the content, we gather various metadata about the page, including the number of page views, an overall importance rating, text within paragraphs, its title, associated keywords, relevant categories, and individual URLs. Leveraging the standardized layout of Wikipedia pages, we extract keyword-rich

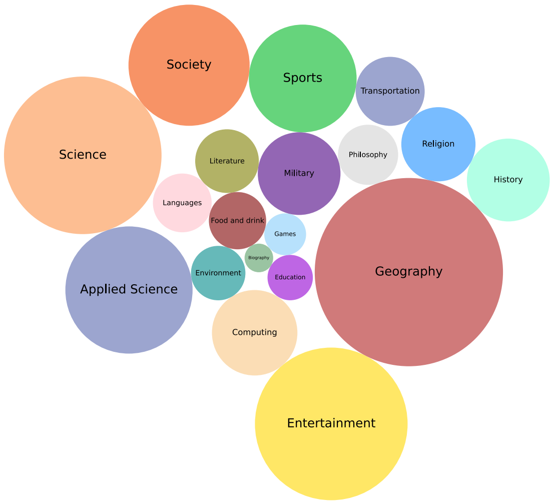


Figure 4: Distribution of the examples among the twenty categories.

opening paragraphs that encapsulate the core content, offering succinct explanations or definitions, and contributing to the construction of a valuable dataset.

Data Screening. With the goal of discarding low-quality data, we adopt several filters: (1) We select pages based on the number of views and importance rating; (2) We remove pages with too long or too short contents; (3) All the data with keywords made of more than three words were removed; (4) all the keywords outside typical English crossword boundaries – $[3, 20]$ character length – or containing non-alphabetical symbols were excluded.

Craft the prompt. The creation of an effective prompt was a crucial aspect of our methodology. We carefully designed prompts for crossword clue generation by incorporating the relevant keywords extracted from the Wikipedia pages. These prompts were structured to provide contextual guidance for generating clues that were both informative and engaging. By using the extracted keywords along with the context of the Wikipedia page, the prompts acted as input signals to guide the generation of crossword clues. Our goal was to create prompts that were well-suited to each specific topic or subject area, taking into account the unique characteristics of the information we had gathered. Crafting the prompt effectively played a key role in the success of our approach, enabling our system to produce high-quality crossword clues tailored to educational needs. In Figure 2, the prompt employed in the study is depicted.

Clues Generation. After assembling content, keywords and categories into the prompt, in the last pipeline step, we generate educational clues for such data. Inspired by SELF-INSTRUCT (Wang et al., 2022), we make use of Large Language Models for automatically generating clues. Differently from SELF-INSTRUCT, generation is strongly conditioned by the information in the input context of the LLM. Therefore, we expect it to produce more faithful clues, thus significantly mitigating the risks of hallucinations.

clue-instruct	
# contexts	44,075
# keywords	44,075
# categories	20
# clues	132,225

Table 1: General statistics on `clue-instruct` dataset.

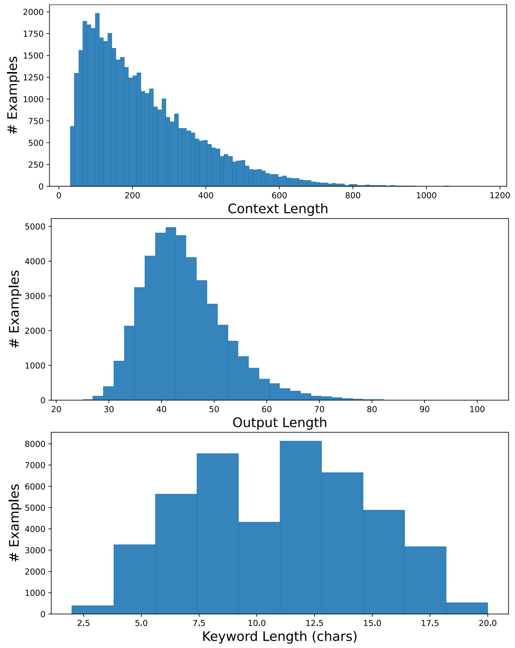


Figure 5: Word length distribution of contexts and outputs. Char length distribution over keywords.

4. Clue-Instruct Dataset

With the method described in Section 3, we construct a dataset for English educational clues, start-

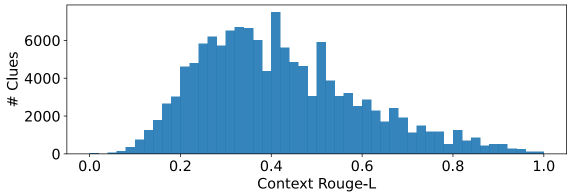


Figure 6: ROUGE-L score distribution of clues vs sentences in wiki page contents computed over the entire `clue-instruct` dataset.

Answer	Category	Clue	Rating
Robocall	Society	May be blocked by phone companies to prevent scams	A
Ministry Of Magic	Literature	Corrupt and incompetent government in J.K. Rowling's Wizarding World	A
Lovesick	Literature	Renewed for a third season, released exclusively on Netflix	C
South American tapir	Science	One of the four recognized species in the tapir family	E

Table 2: Some examples of generated clues alongside the human rating assigned to each of them.

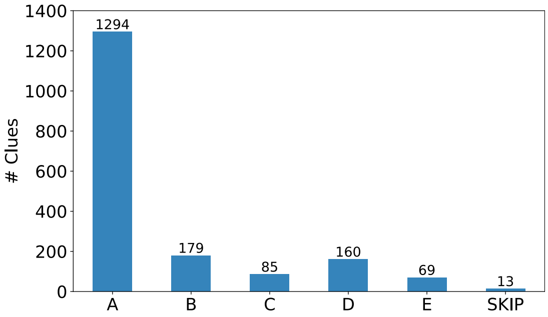


Figure 7: Ratings assigned by humans on the test set.

ing from the most popular pages of 20 distinct categories, that initially contained 258,325 Wikipedia pages. We kept all the pages with more than 10,000 views or with an importance rating equal to 'Top'. Contexts below 30 or above 1000 words were deleted. After data screening, we obtained a corpus of 44,075 examples in total. We used GPT-3.5 Turbo (Brown et al., 2020) as clue generator LLM, with the prompt depicted in Figure 2.

4.1. Statistics

In this section, we delve into the statistical properties of `clue-instruct`. Table 1 presents an overview of the dataset. It comprises overall 44,075 textual content-keyword pairs across 20 different categories. Three distinct clues were generated from each content-keyword-category triplet, resulting in a total of 132,225 clues.

As previously highlighted, examples are divided into 20 distinct categories. In Figure 4, we visually represent the frequency of each category within our dataset using a bubble plot. The size of each bubble corresponds to the frequency of the respective category in the dataset. Upon analyzing this plot, it becomes evident that 'Geography', 'Science', and 'Applied Science' are the most prevalent categories, in that order. Conversely, 'Biography', 'Games', and 'Education' are the least frequent ones.

In Figure 5, we outline the distribution of context and output lengths in relation to the number of words. Such a figure also presents the keyword length distribution in terms of characters. We can observe how the context length falls in a wide range going from 30 to 1000 words, with the vast majority of examples having context lengths between 50 to 400. Conversely, most outputs have word lengths between 35 to 50. Additionally, the keyword character length spans from 3 to 20 as imposed during the corpus creation.

4.2. Measuring Data Quality

To evaluate the dataset quality, we resorted to both automatic metrics and human evaluations.

Automatic Metrics. Due to the absence of a reference corpus for educational crosswords, there is no reference set to compare the generated clues with. Therefore, we cannot produce standard automatic metrics such as ROGUE scores. Nonetheless, in the specific educational clue generation task, good clues should tightly adhere to the reference context, being simple reformulations of some information stated in the text. Hence, the problem is highly extractive.

From such considerations, we exploited as automatic evaluation, the ROUGE-L score between the sentences in the input context against the generated clue. Intuitively, scores should be high enough to indicate strong adherence to the context, thus reducing the chances of hallucinations, but not too close to perfect matches, which would be an indication of poor clue styling and high chances of injecting the target keyword within the clue itself. On average, we obtained about 42 ROUGE-L, which indicates a significant entailment between the generated clue and the most similar sentence in the context. The distribution over the dataset is outlined in Figure 6.

Human Evaluation. To assess the quality of generated data, we cannot solely rely on automatic metrics. Thus, we sample a portion of clue-instruct for human evaluation. Similarly to Wang et al. (2022), we consider a five-level rating, under the following guidelines:

- `RATING-A`: The clue is valid and coherent to the given context, answer, and category.

model type	model name	# params	ROUGE-1	ROUGE-2	ROUGE-L
Off-the-shelf LLMs	LLAMA2-CHAT	7B	–	–	–
	MPT-INSTRUCT	7B	23.98	11.79	19.69
	LLAMA2-CHAT	13B	**31.80**	**15.32**	**25.27**
	MPT-INSTRUCT	30B	29.92	14.47	24.30
Finetuned LLMs	LLAMA2-CHAT	7B	59.92	40.98	52.28
	MPT-INSTRUCT	7B	59.26	40.37	51.68
	LLAMA2-CHAT	13B	**62.97**	**44.97**	**55.40**
	MPT-INSTRUCT	30B	61.42	42.63	53.77

Table 3: Performance of off-the-shelf LLMs with and without fine-tuning. Without clue generation instruction tuning, smaller models struggle to follow the request. Fine-tuning greatly improves the performances of all the LLMs.

- `RATING-B`: Acceptable clue with minor imperfections - loose correlation with category.
- `RATING-C`: The clue is relevant to the answer but loosely correlates with the context, or it is too generic.
- `RATING-D`: The clue is irrelevant and/or incorrect with respect to the answer or the context.
- `RATING-E`: Not acceptable clue because it contains the answer (or a variant of it).

We also allow annotators to skip examples (marked with `SKIP`), in case there are issues not strictly related to the clue itself, such as odd keywords or documents.
Overall, 600 examples were annotated, for a total of 1,800 clues evaluated, since there are three clues proposed by the model for each given context, keyword, and category triplet. We report rating distributions in Figure 7. More than two out of three (about 72%) clues were marked with `RATING-A`, the highest score, which grows to 81% if we consider as acceptable also the clues rated with `B`.

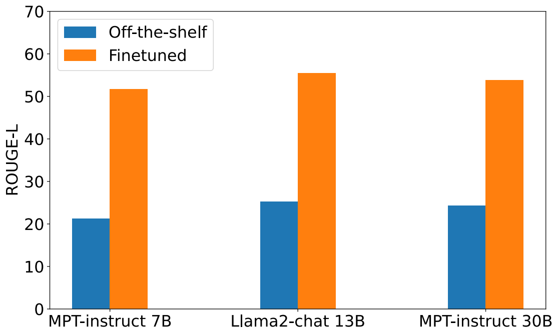


Figure 8: ROUGE-L of different LLMs with and without finetuning.

5. Experiments

We investigate the usage of `clue-instruct` to fine-tune different families of LLMs with various dimensions.

5.1. Experimental Setup

Data. LLMs were trained for instruction tuning on `clue-instruct`. We kept the 600 annotated examples as a test and used them to evaluate all of our models using GPT-3.5 Turbo as an oracle. The remaining 43,475 examples were used for training. LLMs were instructed with the same prompt used for GPT-3.5 Turbo depicted in Figure 2.

Baselines. We focus on four instruction-tuned LLMs: LLAMA2-CHAT Touvron et al. (2023) in 7B and 13B sizes, and MPT-INSTRUCT Team (2023) in both 7B and 30B releases.

Training details. All the models were fine-tuned with LORA (Hu et al., 2021), $r = 16$, and $\alpha = 32$ over the course of two training epochs and batch size set to 32. Learning rate was initialized to $3 \cdot 10^{-4}$ with a linear warm-up of 200 steps. At inference time, clues were generated by sampling from the model distribution. The temperature was set to 0.1, while top-p and top-k (Holtzman et al., 2019) were set to 0.75 and 50, respectively. All the experimentation was carried out on a server equipped with four NVIDIA A6000 GPUs.

5.2. Results

Off-the-shelf LLMs. First of all, we evaluate the four baseline models in zero-shot, i.e. without any fine-tuning on `clue-instruct`. Comparison is shown in Table 3. Despite being previously trained to follow generic instructions, all the models struggle to produce a valid set of clues. Results are in general not satisfactory. In particular, LLAMA27B-chat always fails to produce an acceptable output with the given prompt. Probably, different prompt designs would have led to better results for such a model, however, this inquiry goes beyond the goals of our paper. Also MPT-INSTRUCT-7B often poorly fails to produce the correct JSON output and often generates a single clue, instead of the three requested. With the increase of models' parameters, also the quality grows. Both LLAMA2-13B-CHAT and MPT-INSTRUCT-30B have higher ROUGE-L scores, with the former slightly better than the latter. This

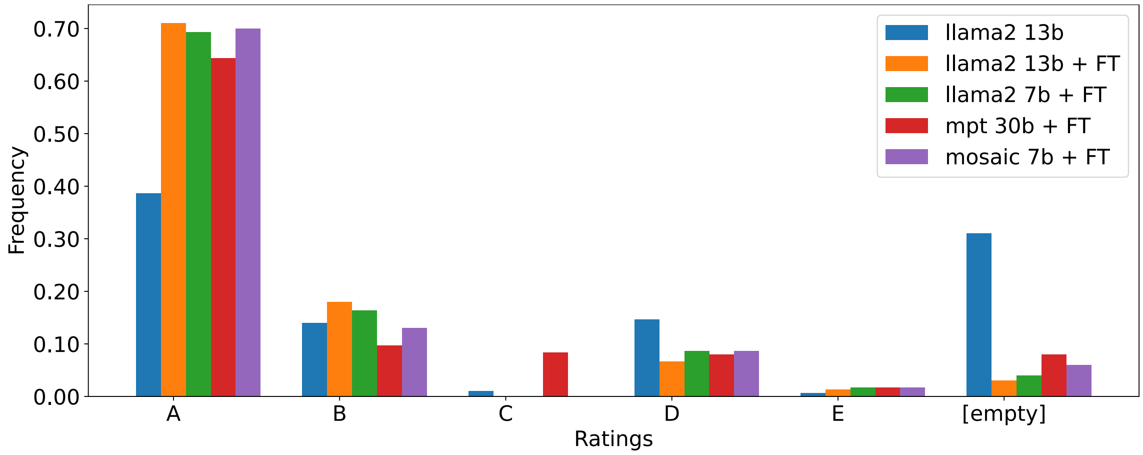


Figure 9: Human evaluation of clues generated by different LLMs. Models finetuned on `clue-instruct` are indicated with "+ FT".

is mainly due to the fact that Llama2-13B-chat always produced the exact JSON schema, whereas mpt-instruct-30B failed almost once every four times.

Finetuned LLMs. When finetuning the baseline LLMs on `clue-instruct`, all the models exhibit a remarkable improvement. Such a comparison is clearly shown in Figure 8. ROUGE-L results are outlined in Table 3. The outputs always align with the expected format. Finetuned LLMs surpass off-the-shelf models by a large margin, with an increase above 20 points in ROUGE-L. It is worth noticing that, Llama2-chat 13B is confirmed to be the best model, and that Llama2-chat 7B can recover from catastrophic results. All the finetuned LLMs are publicly available[2,3,4,5].

Impact of model size and LLM family. Analyzing the results from Table 3, we can notice that larger models tend to outperform smaller ones. In particular, larger LLMs are more robust to unseen instructions, thus showing wider gaps when not finetuned on the downstream task. Moreover, we can observe that Llama2-chat 13B model is particularly well-performing, surpassing mpt-instruct 30B, which is more than twice its size, as already observed in the literature.

Impact of dataset size. We also measure how the performance changes when using different amounts of training examples. Training size was cut at 1%, 10%, and 100%, to see the trend at different orders of magnitude. To slightly cope with the reduced amount of training steps, we increase the number of epochs to 3 for 1% and 10% pieces of training, and we reduce the number of warm-up gradient steps to 20. In this experiment, we only focus for simplicity on the Llama2 family (7B, 13B). From the results, outlined in Figure 10, we can observe that a small number of examples are enough to align the LLMs to the task, even for Llama2-chat-7B that failed to produce valid clues when applied as zero-shot. Thus, the biggest leap in performance is given by just a small amount of instructions, coherently with findings in literature (Zhou et al., 2023).

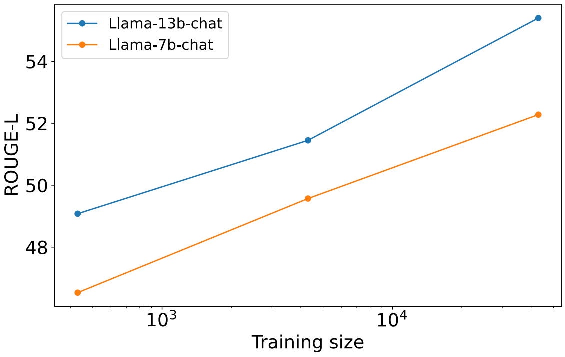


Figure 10: Impact of training size on ROUGE-L. `clue-instruct` is truncated at sizes corresponding to 1%,10% and 100% of the training corpus.

Human Evaluation. In addition to automatic evaluation, human annotators are asked to evaluate the output of the fine-tuned LLMs. We compare the models on a portion of 100 documents of the test set. Due to the poor performance of off-the-shelf models, we only consider Llama2-chat-13B in this evaluation with the purpose of highlighting once again the differences between base models and

[2] https://huggingface.co/azugarini/clue-instruct-llama-7b

[3] https://huggingface.co/azugarini/clue-instruct-llama-13b

[4] https://huggingface.co/azugarini/clue-instruct-mpt-7b

[5] https://huggingface.co/azugarini/clue-instruct-mpt-30b

their fine-tuned versions on clue-instruct. In addition to rating scores, we marked as [EMPTY] all examples where a clue was not produced. We report the results in Figure 9. All the tuned models exhibit a major reduction of malformed outputs ([EMPTY]). In contrast, the number of A-rated examples suddenly increased. Also, D-rated examples diminish, whereas B, C, and E rates have a slight increase, with some exceptions. These results suggest that finetuning is extremely effective in aligning the generated output to the expected format, but there is also a positive contribution to the quality of the generated clues. To help understanding what kind of clues were generated and the ratings assigned, we showcase some examples in Table 2.

6. Conclusions

In this paper, we presented a methodology to generate clues for educational crosswords, from which we constructed clue-instruct, an instruction-tuning dataset with keyword-clue pairs grounded on an input context, specifically designed for educational crosswords. To the best of our knowledge, the corpus is the first resource that combines such information, which is necessary to build systems that can generate educational crosswords from a given document. We then leveraged clue-instruct to fine-tune different open-source Large Language Models, showing that aligning LLMs to this kind of instructions greatly improves the output quality in terms of both automatic and human evaluation. Both the dataset and the models have been publicly released.

In the future, we plan to further extend our methodology to non-English languages in order to facilitate the diffusion of educational crosswords also in less represented languages.

7. Acknowledgements

This work was supported by the IBRIDAI project financed by the Regional Operational Program “FESR 2014-2020” of Emilia Romagna (Italy), resolution of the Regional Council n. 863/2021.

8. Bibliographical References

Giovanni Angelini, Marco Ernandes, and Marco Gori. 2005a. Solving italian crosswords using the web. In *AI* IA 2005: Advances in Artificial Intelligence: 9th Congress of the Italian Association for Artificial Intelligence, Milan, Italy, September 21-32, 2005. Proceedings 9*, pages 393–405. Springer.

Giovanni Angelini, Marco Ernandes, and Marco Gori. 2005b. Webcrow: A web-based crosswords solver. In *Intelligent Technologies for Interactive Entertainment: First International Conference, INTETAIN 2005, Madonna di Campiglio, Italy, November 30–December 2, 2005. Proceedings 1*, pages 295–298. Springer.

Giovanni Angelini, Marco Ernandes, Caroline Stehlé, Fanny Simões, Kamyar Zeinalipour, Andrea Zugarini, Marco Gori, et al. 2023. The webcrow french crossword solver. *arXiv preprint arXiv:2311.15626*.

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.

Gianni Barlacchi, Massimo Nicosia, and Alessandro Moschitti. 2015. Sacry: Syntax-based automatic crossword puzzle resolution system. In *Proceedings of 53nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Beijing, China, July. Association for Computational Linguistics*.

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.

Marco Ernandes, Giovanni Angelini, and Marco Gori. 2005. Webcrow: A web-based system for crossword solving. In *AAAI*, pages 1412–1417.

Marco Ernandes, Giovanni Angelini, and Marco Gori. 2008. A web-based agent challenges human experts on crosswords. *AI Magazine*, 29:77–90.

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).

Matthew L Ginsberg. 2011. Dr. fill: Crosswords and an implemented solver for singly weighted csps. *Journal of Artificial Intelligence Research*, 42:851–886.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

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

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

Michael L Littman, Greg A Keim, and Noam M Shazeer. 1999. Solving crosswords with proverb. In *AAAI/IAAI*, pages 914–915.

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

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

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

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

Leonardo Rigutini. 2010. *Automatic Text Processing: Machine Learning Techniques*. LAP Lambert Academic Publishing.

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

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

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

MosaicML NLP Team. 2023. Introducing mpt-30b: Raising the bar for open-source foundation models. Accessed: 2023-06-22.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Eric Wallace, Nicholas Tomlin, Albert Xu, Kevin Yang, Eshaan Pathak, Matthew Ginsberg, and Dan Klein. 2022. Automated crossword solving. *arXiv preprint arXiv:2205.09665*.

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

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

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

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

Kamyar Zeinalipour, Mohamed Saad, Marco Maggini, and Marco Gori. 2023b. ArabIcros: AI-powered Arabic crossword puzzle generation for educational applications. In *Proceedings of ArabicNLP 2023*, pages 288–301, Singapore (Hybrid). Association for Computational Linguistics.

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment.

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

Andrea Zugarini and Marco Ernandes. 2021. A multi-strategy approach to crossword clue answer retrieval and ranking. In *Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it Milan, Italy.*

Andrea Zugarini, Thomas Röthenbacher, Kai Klede, Marco Ernandes, Bjoern M Eskofier, and Dario Zanca. 2023. Die rätselrevolution: Automated german crossword solving. In *Proceedings of the 9th Italian Conference on Computational Linguistics, CLiC-it, Venice, Italy.*