ECWCA - Educational CrossWord Clues Answering A CALAMITA Challenge

Andrea Zugarini[1,*], Kamyar Zeinalipour[2], Achille Fusco[3] and Asya Zanollo[3]

[1]expert.ai, Siena, Italy
[2]University of Siena, DIISM, Via Roma 56, 53100 Siena, Italy
[3]USS Pavia, Piazza della Vittoria 15, 27100 Pavia (PV)

Abstract

This paper presents ECWCA (Educational CrossWord Clues Answering), a novel challenge designed to evaluate knowledge and reasoning capabilities of large language models through crossword clue-answering. The challenge consists of two tasks: a standard question-answering format where the LLM has to solve crossword clues, and a variation of it, where the model is receives hints about the word lengths of the answers, which is expected to help models with reasoning abilities. To construct the ECWCA dataset, synthetic clues were generated based on entities and facts extracted from Italian Wikipedia. Generated clues were then selected manually in order to ensure high-quality examples with factually correct and unambiguous clues.

Keywords

Educational Crosswords Dataset, Large Language Models, CALAMITA

1. Challenge: Introduction and Motivation

Crossword puzzles are well-known linguistic games that are usually used for entertainment, but they are also applied in education as a tool to assess knowledge, reasoning skills and linguistic abilities of students [1, 2, 3]. Large Language Models (LLMs) [4, 5, 6] have shown impressive abilities and strong knowledge about the world. Recently, Language Models have been extensively used to both solve [7, 8, 9, 10, 11] and create crossword clues [12, 13] for educational purposes.

In this challenge instead, we make use of educational crossword clues to build a benchmark to assess the LLM clue-answering skills on popular entities and facts about the world. We refer to it as ECWCA, standing for Educational CrossWord Clues Answering. ECWCA is an Italian benchmark presented at [14], designed to include Entities and Facts that are popular in the Italian culture.

2. Challenge: Description

In this challenge, we evaluate the knowledge abilities of LLMs by testing them on crossword clue-answering tasks. We propose two slightly different tasks in the challenge. The first one, is essentially a Question Answering problem, where the question is a clue and we expect the LLM to reply with the correct answer. In the second case, the goal is analogous, but we assist the model with hints related to the length of the words in the answer. Suggestions reduce the number of possible answers, therefore models with reasoning skills are supposed to take advantage of that.

To build ECWCA, we created a dataset of synthetic clues grounded on entities and facts extracted from Italian Wikipedia pages. Clue-answer pairs were generated following the same methodology of clue-instruct [13]. In a nutshell, we create multiple clues for a given answer. The generation is grounded to a content that is about the given answer, and a topic. A sketch of the method is outlined in Figure 1. Since the approach produces multiple definitions for a single answer, and the quality may not be good enough for all of them, we perform a manual selection step to preserve only high-quality clues.

3. Data description

3.1. Origin of data

The dataset was constructed following the clue-instruct [13] approach. In clue-instruct it was faced a clues generation problem. Indeed, the task was to generate multiple clues given a certain answer, its context and its category. Here instead, we exploit the approach to build a QA dataset of clue-answer pairs. This happens in two steps, first we generate a set of examples constituted by an answer and the generated clues (as in clue-instruct), then we manually select the most suited clue-answer pairs (see Section 3.2 for further details).

In order to construct the examples with clue-instruct,

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
azugarini@expert.ai (A. Zugarini); kamyar.zeinalipour2@unisi.it (K. Zeinalipour); achille.fusco@iusspavia.it (A. Fusco); zanolloasya@gmail.com (A. Zanollo)

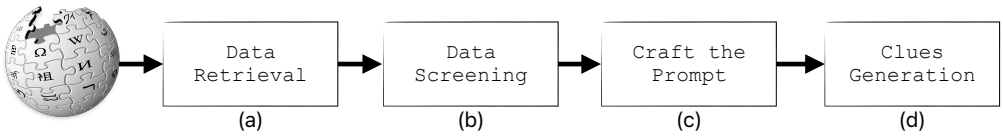

Figure 1: Sketch of clue-instruct method. Picture taken from [13].

we identified the most visited Italian Wikipedia[1] pages. To count visits, we considered a period between September 10, 2023 and May 31, 2024 and gathered stats from Wikimedia APIs[2]. We considered the page title as the answer. Titles with non-alphabetic characters, with less than two characters or more than 20 were excluded. On the remaining pages, we extracted their content. Differently from clue-instruct, we did not dispose of the category information, therefore we generated it by querying GPT-4o [6], asking to choose the category of the answer given its page content within a set of 20 predefined categories. We then randomly sampled the pages and we interrogated GPT-4o to create three clues for the answer. Finally, those examples underwent through the manual selection process, to keep only one clue amongst the three. The dataset is publicly available[3].

3.2. Annotation details

The clue-instruct method produces three different clues for each given answer and its context. To select only one clue we add a human selection step. Doing so, we avoid the presence of multiple occurrences for the same answer. Moreover, we guarantee high quality definitions and answers.

The example selection process was carried out by three native Italian speaking annotators. Examples were split in 18 chunks of 100 examples each, equally distributed among the annotators.

Each example was presented with the answer, the three generated clues and the Wikipedia page paragraph that was used to create the clues. Annotators were tasked with selecting the best one, if any, based on the following criteria:

Truthfulness and Accuracy. It was imperative that the content of the selected clue was factually correct. Annotators cross-verified the accuracy of the clue from the provided Wikipedia page content to ensure that it did not contain misleading or false information, thereby ensuring the integrity of the dataset.

Answerability. Annotators were instructed to choose a clue that could be answered without a high degree of ambiguity. The focus was on clues that provided enough information to infer the correct answer with confidence. Clues that left room for multiple interpretations or guesses were rejected. For example, generic definitions, such as 'a large mammal', does not fit this criteria, since there are many possible species fitting for this answer.

No clue-answer overlap. Clues including the answer or a significant portion of it should be discarded.

In cases where more than one clue satisfied all the criteria, annotators were directed to select the clue that provided the most relevant information with most clarity and simplicity. When no clue matched the criteria, the whole example was discarded.

3.3. Data format

Each example includes the clue-answer pair, the word length hint, some additional metadata (such as the category and the page views) and the reference to the wikipedia page url, whose content was exploited to generate the clue. More precisely, there are the following columns: `clue`, `answer`, `answer_len`, `url`, `content`, `views`, `category`, `length_hint`, `raw_entity`. A few examples are showcased in Table 1, where for the sake of simplicity, we only report the clue-answer pair, the hint and the category of the example.

3.4. Example of prompts used for zero or/and few shots

We defined two different prompts, one with and the other without indications about the words length of the answer. The two prompts are presented in Figure 4 and Figure 3, respectively.

[1] https://it.wikipedia.org/
[2] wikimedia.org
[3] https://huggingface.co/datasets/azugarini/crossword-clues-QA

Table 1
Some examples of generated clues in the dataset, their answers, the hint suggesting the character length of each word in the answer and the category representing the topic of the clue.

Clue	Length Hint	Category	Answer
Sovrana che instaurò rapporti con Giulio Cesare e Marco Antonio	(9)	History	Cleopatra
Autore de I Malavoglia e Mastro-don Gesualdo	(8,5)	Literature	Giovanni Verga
Pilota austriaco tre volte campione del mondo di Formula 1	(4,5)	Sports	Niki Lauda
Attore canadese protagonista di Blade Runner 2049	(4,7)	Entertainment	Ryan Gosling
Opera divisa in tre cantiche: Inferno, Purgatorio e Paradiso	(6,8)	Literature	Divina Commedia
Stato dell'Oceania con capitale Canberra	(9)	Geography	Australia

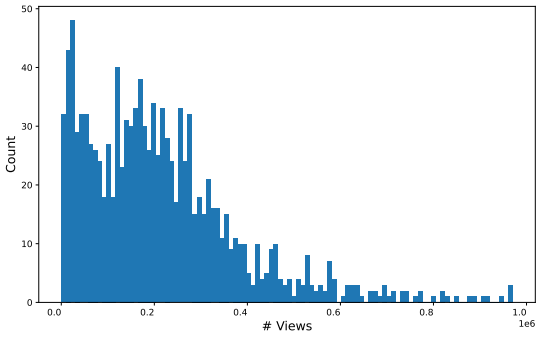

Figure 2: Page views distribution (the very few examples above one million visits were excluded).

```
Sei un esperto di enigmistica. Devi risolvere
definizioni di cruciverba.
Trova la risposta alla definizione. Ritorna solo la
risposta, nient'altro.

Esempi:

DEFINIZIONE: Protagonista di Titanic al fianco di
Kate Winslet
RISPOSTA: leonardo dicaprio

DEFINIZIONE: capitale dell'Impero romano d'Occidente
nel 313 d.C.
RISPOSTA: milano

Ora tocca a te:

DEFINIZIONE: {clue}
RISPOSTA:
```

Figure 3: Prompt task without hints.

Task without hints. We construct a 2-shot prompt (Figure 3) for the task. First, we instruct the model to act as an expert in solving crossword clues without any additional hints related to the structure of the answer (such as words length). The format is clear and concise, focusing on the core task: resolving the crossword definition and providing only the solution. Then, the two static demonstration examples are showcased to illustrate to the model how to approach the task. Finally, following the same layout, we present a new clue and expect the model to complete it with the answer.

Task with word length hints. This prompt (see Figure 4) is very similar to the first one, but introduces an hint indicating the words length of the expected answer. The hint is a constraint that reduces the number of valid answers, giving indications on both how many words there are and their lengths, therefore, ideally, it should aid the language model.

3.5. Detailed data statistics

Overall we collected 1,171 clue-answer pairs belonging to 16 different categories. The distribution of answers among categories is outlined in Figure 5. Most of the examples belong to Entertainment topic, indeed the dataset includes many actors, tv shows, movies and fictional characters. Sports, Geography, History and Society are also well represented, whereas the remaining categories are less frequent, which some, like Applied Science, Philosophy and Education being rare.

The pages from which clue-answer pairs were built have about 234 thousand views each on average, with a minimum of 1,108 up to almost five million views. However, only a few examples outreach the million and the vast majority of them is within the half million visits, as we can observe from Figure 2.

4. Metrics

To evaluate the performance on the tasks we rely on the following metrics: Edit Distance (ED), Exact Match (EM), and average F1 score on words (F1).

Edit Distance. Edit Distance (also known as Levenshtein Distance) measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one sequence into another. In this context, ED measures how close the generated

```
Sei un esperto di enigmistica. Devi risolvere
definizioni di cruciverba.
Ti verrà data una definizione corredata da un
suggerimento, una sequenza di numeri indicante di
quanti caratteri è composta ciascuna parola della
risposta.
Trova la risposta alla definizione.
Ritorna solo la risposta, nient'altro.

Esempi:

DEFINIZIONE: Protagonista di Titanic al fianco
di Kate Winslet
SUGGERIMENTO: (8,8)
RISPOSTA: leonardo dicaprio

DEFINIZIONE: capitale dell'Impero romano
d'Occidente nel 313 d.C.
SUGGERIMENTO: (6)
RISPOSTA: milano

Ora tocca a te:

DEFINIZIONE: {clue}
SUGGERIMENTO: {length_hint}
RISPOSTA:
```

Figure 4: Prompt task with word length hints.

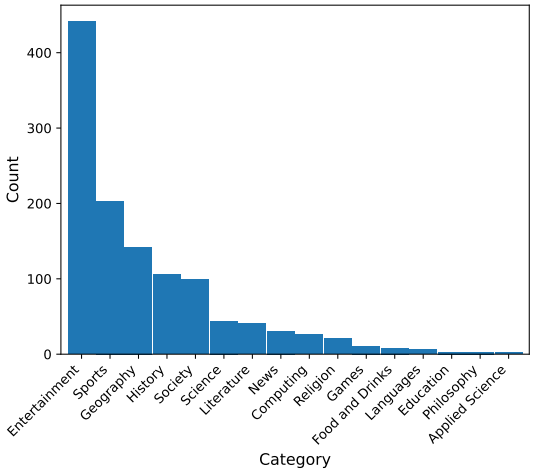

Figure 5: Distribution of the examples across the categories.

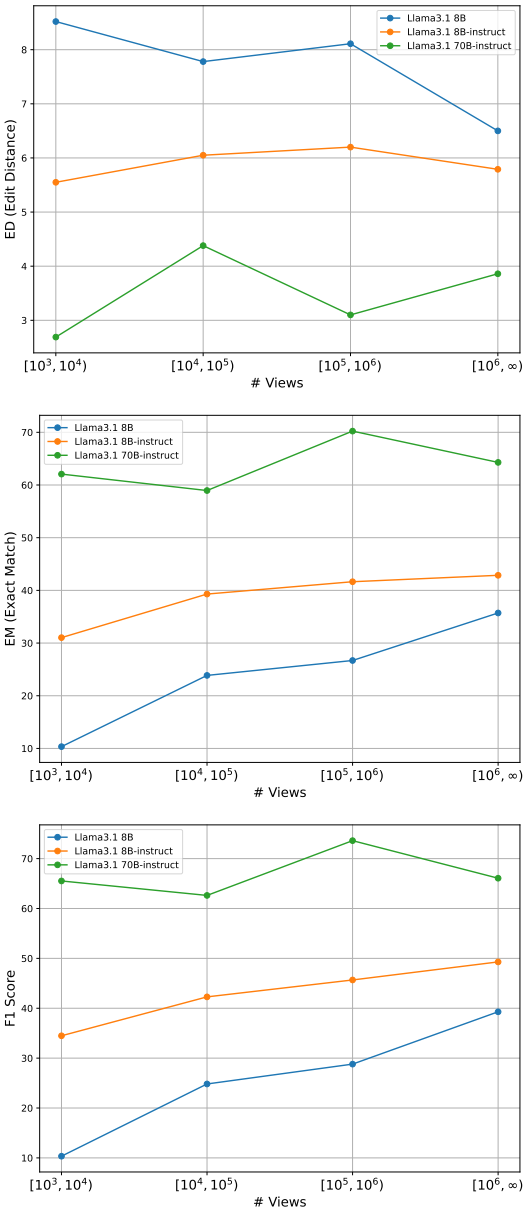

Figure 6: ED, EM and F1 score performance varying with respect to the number of page views for 3.1 llama models.

response is to the ground truth answer. A lower ED indicates better performance, as it signifies that the predicted text is more similar to the target text.

Exact Match. Exact Match (EM) is a binary metric that evaluates whether the generated answer exactly matches the ground truth. We report in percentage the EM score obtained in each example, which corresponds to the percentage of correctly predicted answers.

F1 score. The F1 score evaluates how well the predicted words overlap with the ground truth answer. For example, if the ground truth is ”leonardo dicaprio” and the model predicts ”dicaprio”, the model would have perfect precision, but imperfect recall (50%), resulting in a 66.67% F1 score.

Table 2
Performance on the task with and without word length hints.

Model	Hint	ED ↓	EM	F1
Llama3 8B	No	11.43	14.82	16.37
Llama 8B	Yes	11.52	10.82	11.91
Llama3 8B-instruct	No	11.43	14.82	16.37
Llama3 8B-instruct	Yes	12.07	14.48	16.07
Llama3.1 8B	No	6.99	34.16	37.35
Llama3.1 8B	Yes	8.01	25.72	27.51
Llama3.1 8B-instruct	No	7.31	39.69	44.47
Llama3.1 8B-instruct	Yes	6.14	40.80	44.58
Llama3.1 70B-instruct	No	3.32	66.61	70.16
Llama3.1 70B-instruct	Yes	**3.27**	**67.89**	**71.24**

Preliminary Results. We establish baseline results on ECWCA, testing some of the models in the Llama family. In particular, we consider Llama3 8B and Llama3.1 8B in both instructed and non-instructed versions, and the Llama3.1 70B-instruct, to observe how model size affects the results. Table 2 illustrates the performance of the LLMs on the two tasks (with and without word-length hints), both evaluated on the defined scores. We can observe that Llama3.1 8B consistently outperforms its predecessor across all the metrics, both with and without hints. The gap between smaller LLMs and Llama3.1 70B-instruct is remarkable, proving once again that larger LLMs preserve much more knowledge.

Word-length hints instead are generally not helping the models, actually harming the performance in non-instructed models. For example, the F1 score of Llama3.1 8B drops significantly, from 37.35 without hints to 27.51 with hints, and similarly, EM decreases from 34.16 to 25.72 as well. Instructed models instead are not affected by this, but the suggestions lead to a small increase in all the metrics. Only in Llama3.1 70B-instruct, we can observe some statistically significant improvement. This may suggest that constraints are beneficial only on models with stronger understanding capabilities.

In Figure 6, we show how the performance of Llama3.1 family models vary with respect to the number of page views. We group examples in intervals, then we compute the metrics on each of them. Edit distance shows no significant trends, whereas EM and F1 exhibit an increasing trend on more visited pages for 8B sized models, whereas the 70B model has a behaviour that seems uncorrelated with the number of views. This suggests that the larger number of weights in 70B model, stored a broader and deeper knowledge about world facts and entities, covering also less popular ones, whereas smaller LLMs did embody only the most popular factual knowledge seen during training.

5. Limitations

Large Language Models have all been exposed to vast amount of data. The clues proposed in this dataset were created from Wikipedia pages that were definitely seen by the LLMs during training. Clues are also generally very adherent to the pages content, since they were created from it. Indeed, one of the goals of the benchmark is to assess their memorization capabilities on facts that were likely to be well known by them. However, the proposed dataset is new, hence it could not have been part of the training set of such LLMs.

6. Data license and copyright issues

Data is released under apache-2.0 license.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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).

[6] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[7] A. Zugarini, M. Ernandes, A multi-strategy approach to crossword clue answer retrieval and ranking (2021).

[8] E. Wallace, N. Tomlin, A. Xu, K. Yang, E. Pathak, M. Ginsberg, D. Klein, Automated crossword solving, arXiv preprint arXiv:2205.09665 (2022).

[9] A. Zugarini, T. Rothenbacher, K. Klede, M. Ernandes, B. M. Eskofier, D. Zanca, Die rätselrevolution:

Automated german crossword solving., in: CLiC-it, 2023.

[10] G. Angelini, M. Ernandes, T. Iaquinta, C. Stehlé, F. Simões, K. Zeinalipour, A. Zugarini, M. Gori, The webcrow french crossword solver, in: International Conference on Intelligent Technologies for Interactive Entertainment, Springer, 2023, pp. 193–209.

[11] S. Saha, S. Chakraborty, S. Saha, U. Garain, Language models are crossword solvers, arXiv preprint arXiv:2406.09043 (2024).

[12] K. Zeinalipour, T. Iaquinta, A. Zanollo, G. Angelini, L. Rigutini, M. Maggini, M. Gori, Italian crossword generator: Enhancing education through interactive word puzzles (2023).

[13] A. Zugarini, K. Zeinalipour, S. S. Kadali, M. Maggini, M. Gori, L. Rigutini, Clue-instruct: Text-based clue generation for educational crossword puzzles, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 3347–3356. URL: https://aclanthology.org/2024.lrec-main.297.

[14] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.