What Makes Cryptic Crosswords Challenging for LLMs?

Anonymous ACL submission

Abstract

Cryptic crosswords are puzzles that rely on general knowledge and the solver's ability to manipulate language on different levels, dealing with various types of wordplay. Previous research suggests that solving such puzzles is a challenge even for modern NLP models. However, the abilities of large language models (LLMs) have not yet been tested on this task. In this paper, we establish the benchmark results for two popular LLMs: `LLaMA3` and `ChatGPT`, showing that their performance on this task is still far from that of humans. We also investigate why the models struggle to achieve superior performance. [1]

1 Introduction

A cryptic crossword is a type of crossword puzzle that is known for its enigmatic clues (Friedlander and Fine, 2016). Unlike standard crossword puzzles, where clues are straightforward definitions or synonyms of the answers, cryptic crosswords involve wordplay, riddles, and cleverly disguised hints that make solving them more challenging (Moorey, 2018). Figure 1 demonstrates an example of a cryptic crossword clue.

To solve a cryptic clue, one should be able to not only apply generic rules in the specific context of the clue but also use general and domain-specific knowledge to arrive at a reasonable answer. Tackling cryptic crosswords with modern NLP methods, therefore, provides an interesting challenge. It has been shown that current NLP models are far from human performance: Rozner et al. (2021), and Efrat et al. (2021) report accuracy of 7.3%, and 8.6% for rule- and transformer-based models against 99% achievable by expert human solvers (and 74% by self-proclaimed amateurs) (Friedlander and Fine, 2009, 2020), and there are still no official statistics for average human performance.

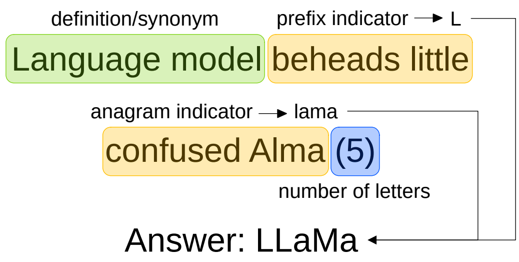

Figure 1: An example of a cryptic clue: number 5 at the end of the clue denotes the number of characters in the answer and is called **enumeration**. The **definition** part here is *language model*, with the rest being the **wordplay** part. *Beheads* or similar words point to the first letters of the next word, while *confused* (as well as *mixed up*, etc.) is likely to indicate an anagram. As we should look for a language model's name that starts with the letter *l* plus an anagram of *Alma* and consists of 5 letters, the answer here is *LLaMA*.

This identifies a challenging area for current NLP research, while also opening up possibilities for improvement and innovation.

Prior work suggests that LLMs can show emergent capabilities (Wei et al., 2022), so it can be assumed that they should be able to solve cryptic puzzles if not on a par with human solvers, then at least somewhat successfully. However, to the best of our knowledge, this assumption has not been tested before. In this work, we address this research gap as we believe that trying to solve cryptic clues with LLMs might reveal their limitations as well as important aspects of natural language understanding and interpretation captured by LLMs.

Typically, a cryptic clue can be split into two parts: the **definition** and the **wordplay**. The definition consists of one or more words in the clue that can be used interchangeably with the answer. Definition usually appears either at the beginning or at the end of the clue. The wordplay can take many forms: the most popular ones include anagrams, hidden words, and double definitions, among oth-

[1] The code will be available online upon acceptance

ers (see Table 3 for the most popular wordplay types and examples for each of them). Previous work has explored explicitly splitting the solution into these two parts (Deits, 2015; Rozner et al., 2021).

Past approaches applied to solving cryptic clues range from rule-based models,[2] to traditional machine learning models like KNN (Rozner et al., 2021), to Transformer models like T5 (Rozner et al., 2021; Efrat et al., 2021). However, all these models achieve only modest accuracy on the task (§2).

Our preliminary investigation suggests that a zero-shot, naive approach to LLMs evaluation yields very low accuracy. In this work, we try to understand the shortcomings of LLMs and figure out the aspects of the task that cause models to struggle. We focus on three main areas to analyze the model reasoning. First, we explore if the models can extract the definition part from the clue. Next, we test the models' ability to extract the wordplay type in prompts with different levels of information. Finally, we test the models' internal reasoning by prompting them to explain how to reach the clue's answer.

Our main contributions are as follows: **(1)** We explore the general abilities of LLMs on the challenging task of solving cryptic crosswords using simple prompting strategies; **(2)** We investigate models' understanding of the task by addressing 3 auxiliary tasks; **(3)** To facilitate reproducibility of our results and follow-up experiments, we release our data and code.

2 Related Work

LLMs' emergent capabilities LLMs have been shown to follow the scaling law (Kaplan et al., 2020), which has motivated researchers to explore the performance limit by increasing the size of both model and data. This has led to the discovery of the emergent abilities of LLMs (Wei et al., 2022), which occur when training models with similar architectures and on the same tasks at different scales. As a result, models may exhibit unexpected abilities in solving a series of novel tasks: for instance, a relatively small LLM like GPT-3 (Brown et al., 2020) does well on arithmetic tasks, question answering or passage summarization just through in-context learning (Yousefi et al., 2024).

Solving puzzles with NLP models Although there is prior work on wordplay (Luo et al., 2019; He et al., 2019; Ermakova et al., 2023) and traditional crosswords (Littman et al., 2002; Zugarini et al., 2023), much less attention has been paid to cryptic crosswords specifically. Deits (2015) achieved 8.6% accuracy on the task with a rule-based solver, which applied hand-crafted probabilistic context-free grammar to find the best split.

Efrat et al. (2021) introduced `Cryptonite`, a dataset of 523,114 cryptic clues collected from *The Times* and *The Telegraph*. They fine-tuned a T5 (Raffel et al., 2023) model, which helped set the benchmark accuracy for Transformer methods at 7.6%. Rozner et al. (2021) introduced a dataset extracted from *The Guardian*, and introduced a curriculum approach, which involved training a model on simpler tasks before progressing to more complex compositional clues. This increased the performance to 21.8%.

3 Data

3.1 The Guardian dataset

In our experiments, we primarily use the dataset introduced by Rozner et al. (2021) and extracted from *The Guardian*. Most models were tested on this dataset, so we chose it as well for comparison purposes. The dataset contains 142,380 clues in total. We evaluate our models on the test subset of 28,476 examples, referred to as "naive random split". Rozner et al. (2021) introduced different splits for the dataset because of the tendency of T5 to remember certain clues/answers during fine-tuning.

3.2 Times for the Times dataset

To test models' performance across datasets, we used the dataset collected by George Ho,[3] where every clue has a marked definition. The original dataset contains around 600k clues from many sources, which would result in extremely expensive experimentation with LLMs. For that reason, for our experiments, we sampled 1000 representative examples collected from Times for the Times blog. We made sure that the distribution of these examples with respect to the number of words in the definition and their position in the clue is similar to the full dataset. We rely on the available definitions to estimate how well our models understand what the definition is and where it is located in the clue. In addition, this information is helpful for the

[2] https://github.com/rdeits/cryptics

[3] https://cryptics.georgeho.org/

investigation of whether including the definition explicitly helps the models solve the clues.

3.3 Small explanatory dataset

Unfortunately, there is no large-scale dataset that contains information about the wordplay types of the clues. To investigate whether our models can detect wordplay types, we manually select 25 examples from the additional dataset (see §3.2), including 5 examples per each major wordplay type (anagram, assemblage, container, hidden word, and double definition – see Table 3).

4 Methodology

4.1 Zero-shot solving

Base prompt We begin by defining a simple prompt (see Figure 2) that only includes the minimal information required to solve the task. We include the line "you are a cryptic crosswords expert", as it has been shown that it can help the model to act like an expert on a specific task (Xu et al., 2023).

All-inclusive prompt We then try to combine general information about cryptic crossword solving without expanding it with a few shots or Chain of Thoughts (CoT) (see Figure 3). We include information about the parts of a clue and their meaning. We also add information about the usual position of the definition in the clue. Finally, our preliminary experiments suggest that LLMs tend to suffer from understanding the constraints of the answer length mentioned in the clue, so we explicitly tell the model that the number of letters in the answer is mentioned inside the parentheses at the end of the clue.

4.2 Dividing solution process into sub-tasks

Next, we investigate why the models struggle to solve the task. To do that, we design a set of experiments that test the models' ability to (1) extract the definition word(s) in the clue, (2) detect the wordplay type in the clue, using different levels of information about the wordplay types, (3) explain the solution process, given the clue and the answer. In addition, we experiment with giving the models definition part of the clue.

5 Experiments and Discussion

We choose one open-source LLM (`LLaMA3`) and one closed-source model (`ChatGPT`). For the latter, we use `gpt3.5-turbo` version. All our results are summarized in Table 1.

5.1 Zero-shot solving

The first 4 rows of Table 1 show the models' accuracy in solving cryptic clues on two different datasets and using two different prompts. From the results, we can see that `ChatGPT` has far better results than `LLaMA3`. Also, we can conclude that providing the model with the definition significantly improves `ChatGPT` performance. To put these results into perspective, in Table 2, we compare our results with the results obtained by Rozner et al. (2021). We can see that `ChatGPT` achieves the same results as the naive fine-tuning, despite that they fine-tuned the model vs. zero-shot prompting in our case. On the other hand, Rozner et al. (2021) mentioned that their approach has data memorization problems.

5.2 Understanding different aspects of the task

5.2.1 Definition extraction

We ask the model to extract the definition part of the clue with the prompt illustrated in Figure 5. We say that the definition should be a synonym for the answer but do not mention that the definition usually comes at the beginning or end of the clue. We see that both models do better with the definition extraction. One reason for that might be that the definition is explicitly included in the clue itself, so the task is to repeat part of the clue, which is arguably easier than inventing new words as an answer.

5.2.2 Wordplay

Determining the wordplay type We identify 5 major types of wordplay listed in Table 3. Then we investigate if our models could identify the wordplay type by the clues. Usually, professional solvers note so-called indicator words that relate the clue to one type or another: for example, *confused, mixed up, mad* usually indicate anagrams. To test the models' ability to identify the wordplay type, we design three experiments that gradually add information for the models. In the first one, we just give the model the names of the 5 different wordplay types and ask it to predict which wordplay type the given clue has (see Figure 6). We notice that `LLaMA3` fails to understand the task and just produces one type for all examples, which suggests that the model does not pay much attention to

Task	No Examples	Info / Prompt	Accuracy LLaMA3	Accuracy ChatGPT
Cryptic Clue Solution	28476	base prompt	2.2*	10.9
Cryptic Clue Solution	28476	all inclusive prompt	2.1*	11.4
Cryptic Clue Solution	1000	all inclusive prompt	3.3*	13.4
Cryptic Clue Solution	1000	all inclusive prompt + definition	3.8*	16.2
Definition Extraction	1000	synonym of the answer	19.3	41.2
Wordplay Type Detection	25	wordplay types	20	36
Wordplay Type Detection	25	+ explanations and examples	20	40
Wordplay Type Detection	25	+ clue answer	32	40

Table 1: Summarized results for our experiments. * means that changes in accuracy from one prompt to another (withing the same dataset) are not statistically significant according to the sign test for the difference between model answers.

Model	Accuracy
LLaMA3 Best	2.2
ChatGPT Best	11.4
Rule-based	7.3
T5 fine-tuned	16.3
T5 fine-tuned + curriculum	21.8

Table 2: Comparison with previous results on naive random test set.

the given clue. Next, we experiment by also giving the model the explanation of each wordplay type and one example for each (Figure 7). Finally, we also add the answer for each clue to test whether the model can infer information about the wordplay from the answer (Figure 8).

From the results, we can see that adding the definition for the wordplay and providing the model with the answer does not help improve the model's ability to extract the wordplay type much. We are aware that the small size of the dataset might not fully support such a conclusion, but one important observation is that the models more frequently identify some "easy" types like **container**, while "harder" types like **assemblage** cause the models more trouble to extract.

5.2.3 Explanation

We ask a model to explain the solution, given the clue and the answer. Our analysis of the models' answers show that: (1) both models follow some kind of structure in their explanations, breaking the clue into parts of 1-3 words. (2) LLaMA3 does not mention any wordplay operations and works only on a synonym level, which is not enough for solving. (3) ChatGPT says some word operations should be applied and sometimes even gets them right. However, it does not properly "understand" the procedure: e.g., *rearranging the letters of "pan" and adding "to cook cheese" results in "parmesan"*.

6 Conclusions and Future Work

We focus on researching the inner workings of LLMs rather than trying to improve the performance on this task. We began by evaluating the models under zero-shot settings, and then we tried to gain insight into the models' understanding of the main task of solving the cryptic clue by using auxiliary tasks. The results suggest that, although the ChatGPT model overall outperforms open-source LLMs, in general, cryptic crosswords still present a very challenging task for LLMs, with a large room for improvement.

We believe performance can be improved in future work with several possible research directions. Firstly, a promising avenue for research in this area is chain-of-thought (Wei et al., 2023) and tree-of-thought (Yao et al., 2023) prompting techniques, which can potentially teach models how to arrive at the solution step by step. Secondly, given a considerable increase in performance achieved by using curriculum learning with T5 (Rozner et al., 2021), we consider this direction is worth exploring with LLMs as well. Finally, such approaches as a mixture of experts (Jacobs et al., 1991; Gale et al., 2022) used to train open-source models like Mixtral (Jiang et al., 2024) can be applied to the task, as they may end up developing expert layers specializing in separate wordplay types.

Limitations

Limited set of LLMs experimented with Experiments with an extensive set of state-of-the-art LLMs can get quite expensive. Due to limitations of time and budget, we have been selective in terms of the LLMs that we use in this study. Specifically, we chose only a few of the most popular open-source and closed-source LLMs. We believe that the obtained results shed light on the current LLMs' capabilities on this task, however, we acknowledge that the set of LLMs we tested here is limited, and our results cannot be extrapolated to other LLMs. In addition, in many experiments, we have observed that certain changes in settings do not bring substantial improvement to the results – this motivated us to perform only a limited set of experiments with some of the models in some of the settings as is elaborated in the paper.

Limitations of the dataset size Some datasets that we used don't have a bigger size in terms of the number of examples. The main reason for this is the lack of datasets with rich annotation and the limitations of the fund, so we had to create the dataset ourselves. We are aware that these datasets can not give a numerical benchmark, but we used it as a theoretical indication of the models' abilities. The main

Closeness to real-world scenario We focus on solving one clue at a time due to simplicity of this task. However, in the real-word scenario human solvers encounter 20-30 clues in one grid. Solving one clue usually reveals letters of the other answers, which can be quite helpful in the solution process.

Dangers of data contamination Finally, we observe in our experiments that `ChatGPT` outperforms the open-source model. We admit that we lack the information about its training setup, since `ChatGPT` is a proprietary model, and therefore, we cannot guarantee that this model's training data is free from contamination.

Ethics Statement

We foresee no serious ethical implications from this study.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

Robin Deits. 2015. Cryptics.

Avia Efrat, Uri Shaham, Dan Kilman, and Omer Levy. 2021. Cryptonite: A cryptic crossword benchmark for extreme ambiguity in language. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4186–4192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liana Ermakova, Anne-Gwenn Bosser, Adam Jatowt, and Tristan Miller. 2023. The JOKER Corpus: English-French Parallel Data for Multilingual Wordplay Recognition. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2796–2806, New York, NY, USA. Association for Computing Machinery.

Kathryn J. Friedlander and Philip A. Fine. 2016. The grounded expertise components approach in the novel area of cryptic crossword solving. *Frontiers*.

Kathryn J. Friedlander and Philip A. Fine. 2020. Fluid Intelligence is Key to Successful Cryptic Crossword Solving. *Journal of Expertise*, 3(2):101–132.

KJ Friedlander and PA Fine. 2009. Expertise in cryptic crossword performance: an exploratory survey. In *Proceedings of the International Symposium on Performance Science, Auckland, eds A. Williamon, S. Pretty, and R. Buck (Utrecht: European Association of Conservatoires (AEC))*, pages 279–284.

Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. 2022. MegaBlocks: Efficient Sparse Training with Mixture-of-Experts.

He He, Nanyun Peng, and Percy Liang. 2019. Pun Generation with Surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1734–1744, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert Jacobs, Michael Jordan, Steven Nowlan, and Geoffrey Hinton. 1991. Adaptive Mixture of Local Experts. *Neural Computation*, 3:78–88.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models.

Michael L. Littman, Greg A. Keim, and Noam Shazeer. 2002. A probabilistic approach to solving crossword puzzles. *Artificial Intelligence*, 134(1):23–55.

Fuli Luo, Shunyao Li, Pengcheng Yang, Lei Li, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Pun-GAN: Generative adversarial network for pun generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3388–3393, Hong Kong, China. Association for Computational Linguistics.

Tim Moorey. 2018. *How to Crack Cryptic Crosswords*. Collins Puzzles.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

Josh Rozner, Christopher Potts, and Kyle Mahowald. 2021. Decrypting cryptic crosswords: Semantically complex wordplay puzzles as a target for NLP. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11409–11421.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models.

Safoora Yousefi, Leo Betthauser, Hosein Hasanbeig, Raphaël Millière, and Ida Momennejad. 2024. Decoding In-Context Learning: Neuroscience-inspired Analysis of Representations in Large Language Models.

Andrea Zugarini, Thomas Röthenbacher, Kai Klede, Marco Ernandes, Bjoern M Eskofier, and Dario Zanca. 2023. Die Rätselrevolution: Automated German Crossword Solving.

```
You are a cryptic crossword expert.
You are given a clue for a cryptic
crossword. Output only the answer.
clue:
{clue}
output:
{output}
```

Figure 2: Base prompt.

A Wordplay types

Common wordplay types are listed in the table 3 with examples [4] and explanations. We identify 5 main types: anagram, assemblage, container, hidden word and double definition.

B Journals Links

In the text of the paper we mention several sources of cryptic crosswords:

1. *The Times*[5]
2. *Telegraph* [6]
3. *The Guardian* [7]
4. *Times for the Times* blog[8]

We do not parse their data specifically and completely but rather use already prepared for us datasets or sample from them.

C Prompts

We present all the prompts we used in this section.

[4] Examples are taken from https://crypticshewrote.wordpress.com/explanations/

[5] https://www.thetimes.co.uk/puzzleclub/crosswordclub/home/crossword-cryptic

[6] https://puzzles.telegraph.co.uk/crossword-puzzles/cryptic-crossword

[7] https://www.theguardian.com/crosswords/series/cryptic

[8] https://times-xwd-times.livejournal.com/

Type	Example Clue	Answer
Anagram: certain words or letters must be jumbled to form an entirely new term.	Never upset **a Sci Fi writer** (5)	Verne
Assemblage: the answer is broken into its component parts and the hint makes references to these in a sequence.	**Bitter** initially, but extremely enjoyable refreshment (4)	Beer
Container: the answer is broken down into different parts, with one part embedded within another.	**The family member** put us in the money (6)	Cousin
Hidden word: the answer will be hidden within one or multiple words within the provided phrase.	**Confront them** in the tobacco store (6)	Accost
Double definition: contains two meanings of the same word.	**In which you'd place the photo** of the **NZ author** (5)	Frame

Table 3: Examples of common wordplay types. The definition part is bolded.

You are a cryptic crossword expert.
The cryptic clue consists of a definition and a wordplay.
The definition is a synonym of the answer and usually comes at the beginning or the end of the clue.
The wordplay gives some instructions on how to get to the answer in another (less literal) way.
The number/s in the parentheses at the end of the clue indicates the number of letters in the answer.
Extract the definiton and the wordplay in the clue, and use them to solve the clue. Finally, output the answer on this format:
Answer: <answer>,
Clue:
{clue}

Figure 3: All inclusive prompt.

You are a cryptic crossword expert.
The cryptic clue consists of a definition and a wordplay.
The definition is a synonym of the answer and usually comes at the beginning or the end of the clue.
The wordplay gives some instructions on how to get to the answer in another (less literal) way.
The number/s in the parentheses at the end of the clue indicates the number of letters in the answer.
Use the given definition, and extract the wordplay in the clue, and use them to solve the clue. Finally, output the answer on this format:
Answer: <answer>,
Clue:
{clue}
Definition:
{definition}

Figure 4: All inclusive prompt with included definition.

You are a cryptic crossword expert. I will give you a cryptic clue. Every clue has two parts: a definition and a wordplay. The definition is a synonym of the clue's answer. Extract the definition word/s from this clue. Only output the definition.
Clue: {clue}
Definition:

Figure 5: Prompt for the definition extraction.

You are a cryptic crosswords expert. I will give you a clue. Every clue has two parts: a definition and wordplay. Definition is a synonym of the answer. Wordplay is the rest of the clue.
Please extract the wordplay type for this clue.
Here is a list of all possible wordplay types: anagram, hidden word, double definition, container, assemblage.
Only output the wordplay type.
Clue: {clue}
Output:

Figure 6: Prompt for the wordplay type classification.

You are a cryptic crosswords expert. I will give you a clue. As you know, every clue has two parts: a definition and wordplay. Please extract the wordplay type from this clue.
Here is a list of all possible wordplay types, and their descriptions:
- anagram: An anagram is a word (or words) that, when rearranged, forms a different word or phrase.
  Example: Ms Reagan is upset by the executives (8)
  The answer: Managers

- hidden word: The answer is found in the clue itself, amongst other words.
  Example: Confront them in the tobacco store (6)
  The answer: Accost

- double definition: Clues contain two meanings of the same word. The words may be pronounced differently, but must be spelt the same.
  Example: Footwear for pack animals (5)
  The answer: Mules

- container: One word is placed inside another (or outside another) to get the answer.
  Example: Curse about the Maori jumper (7)
  The answer: Sweater

- assemblage: The answer is broken up into smaller parts and each syllable or part is given a separate clue. These separate clues are then put together into one clue.
  Example: Brash gets a Prime Minister employment, but it's drudgery (6,4)
  The answer: Donkey work
Only output the wordplay type.
Clue: {clue}
Output:

Figure 7: Prompt for the wordplay type classification with examples for each wordplay type.

You are a cryptic crosswords expert. I will give you a clue. As you know, every clue has two parts: a definition and wordplay. Please extract the wordplay type from this clue.
Here is a list of all possible wordplay types, and their descriptions:
- anagram: An anagram is a word (or words) that, when rearranged, forms a different word or phrase.
    Example: Ms Reagan is upset by the executives (8)
    The answer: Managers

- hidden word: The answer is found in the clue itself, amongst other words.
    Example: Confront them in the tobacco store (6)
    The answer: Accost

- double definition: Clues contain two meanings of the same word. The words may be pronounced differently, but must be spelt the same.
    Example: Footwear for pack animals (5)
    The answer: Mules

- container: One word is placed inside another (or outside another) to get the answer.
    Example: Curse about the Maori jumper (7)
    The answer: Sweater

- assemblage: The answer is broken up into smaller parts and each syllable or part is given a separate clue. These separate clues are then put together into one clue.
    Example: Brash gets a Prime Minister employment, but it's drudgery (6,4)
    The answer: Donkey work
Only output the wordplay type.
Clue: {clue}
The answer: {ans}
Output:

Figure 8: Prompt for the wordplay type classification with examples for each wordplay type. Here we also add the answer for the clue.