Solving Cryptic Clues Using Large Language Models

Hsin-Chun Yin **Elchanan Schnaidman** **Tehila Mescheloff** **Peleg Oppenheimer**
Hebrew University of Jerusalem
{hsin-chun.Yin, schnaidman.elchanan, tehila.mescheloff, peleg.oppenheimer}@mail.huji.ac.il

Abstract

Cryptic crossword puzzles, with their intricate wordplay, present a significant challenge in natural language processing (NLP). This study aimed to improve the accuracy on Cryptonite dataset (7.64%) by exploring various techniques. Our most effective approach combined Auto Strategy Selection and Reflection, leading to a much improved results (29.5%). Despite ongoing challenges, these findings suggest that integrating strategy selection with inference scaling holds promise for future advancements (Our code is publicly available[1]).

1 Introduction

Cryptic crossword puzzles present a unique challenge in natural language processing (NLP), requiring models to decode clues that often rely on wordplay, misdirection, and lateral thinking. The goal of this project was to develop a model capable of solving cryptic clues by improving upon the existing T5 training baseline, which achieved only around 7.64% accuracy [Efrat et al. (2021)], while though top-tier human experts can solve them with almost 100% accuracy. We used the same dataset of cryptic clues and answers and explored a range of techniques to enhance the model's ability to solve these clues.

An improved performance on this cryptic crossword puzzle dataset can serve as a benchmark for evaluating whether the language model has developed logical deduction, and linguistic agility - a highly specific type of reasoning that is distinct from general problem-solving.

Previous studies have shown that Deits (2022) achieved an accuracy of 8.6% with a rule-based system incorporating probabilistic grammar. Improvements were noted by Rozner et al. (2021), who increased solving accuracy to 21.8% through a curriculum-based approach. Additionally, Wallace et al. (2022) demonstrated that strategic planning can lead to solving rates approaching 90% in non-cryptic crosswords.

2 Data

We used Cryptonite, a large-scale dataset based on cryptic crosswords, which is both linguistically complex and naturally sourced.

2.1 Categorization

Cryptonite features a cryptic crossword puzzle in every sample [Efrat et al. (2021)]. Correct categorization plays a key role in solving these puzzles. In cryptic crossword puzzles, there are many kinds of wordplays, and the wordplays are generally classified into several categories. After carefully considering various approaches of classifications [Wikipedia contributors (2024), Shuchi (2008), Moorey (2008)], we classified the wordplays into 10 categories. (See examples of wordplays in Appendix D)

Each clue can generally be classified according to the wordplay it employs, and the puzzle's difficulty often correlates with the variety of wordplay types it contains. A **Simple** clue are puzzle that involves a single wordplay element. A **Medium** clue are puzzles whose Solutions requiring a combination of different wordplay techniques. A **Hard** clue is puzzle that break from conventional patterns or include distinct, unconventional elements. (See details in Appendix D)

By incorporating this dimension into our analysis, we gain a deeper understanding of the complexity involved in solving the clues, which can help refine our approach to cryptic clue-solving strategies.

2.2 Identify Clue Types

To solve a puzzle, the general strategy involves first identifying the type of wordplay required. Indicators are often used to identify wordplay types,

[1] https://github.com/samyiin/CryptoniteAnalysis.git

although some wordplay types lack clear indicators. We compiled indicators from various sources (Shuchi (2008) and Deits (2022)) for identifiable types and analyzed their distribution by category [Figure 1]. The considerable overlap among indicators suggests that identifying the puzzle category thorough indicators, though useful, is imperfect. [Figure 2 , Figure 3].

3 Methods and Experiments

In this project, we employed several strategies to enhance the performance of solving cryptic clues from cryptic puzzles. Our approach evolved through two phases: We start with Fine tuning smaller models and then we shifted to Prompt Engineering on large language models. Below, we detail each phase and the methodologies employed.

3.1 Random Explorations: Fine Tuning LMs

The Cryptonite paper presents fine-tuning a T5-large model as a baseline, achieving roughly 7.64% accuracy. Based on this result, we chose to initiate with fine-tuning language models to address the challenge.

- **Full Fine-Tuning**: We performed full fine-tuning on Seq2Seq models to directly tailor them to our specific task. Despite this effort, the outcomes were less than satisfactory. We had also considered the possibility to customize the tokenizer to enhance the model's ability to process operations at the letter level.

- **LoRA Fine-Tuning**: We applied LoRA fine-tuning techniques to larger Seq2Seq models. This approach also yielded suboptimal results. Our future plan included employing transfer learning through adaptor fusion, aiming to decompose the solving process into simpler tasks and enable the model to learn each task by integrating adaptors.

- **Multiple Choices Approach**: We investigated a unique multiple-choice method inspired by "curriculum learning." This approach involved initially training a model to select the correct answer from other three synthetic wrong answers. We achieved an accuracy of 89% in this multiple-choice task. However, when fine tuning the pre-trained model to the Seq2Seq task, the resulting performance remained inadequate.

- **Mixture of Experts**: We investigated ensemble techniques by combining several models to potentially improve accuracy. This approach aimed to harness the strengths of different models to achieve better results collectively. We see that there is a slight improvement, but it's insignificant because the pretrained models are not good.

Initial attempts to address this problem through fine-tuning Seq2Seq models yielded low accuracy. Details of the models, hyperparameters, and results are provided in Table 1. Given these disappointing preliminary results, we proceeded to explore alternative methods instead of delving further in direction.

3.2 Prompt Engineering

Since solving the problem requires disambiguating semantic, syntactic, and phonetic wordplays, as well as real world knowledge. Recognizing the superior capabilities of larger language models, we opted to explore QA using these models. Early experiments with ChatGPT, where only the cryptic clue and enumeration were provided, produced encouraging results, with the model occasionally solving the puzzle correctly. The following section details how we refined this approach iteratively.

3.2.1 Chain-of-Thought (CoT)

We begin by implementing Chain-of-Thought (CoT) reasoning, guiding the model through a sequence of steps to solve each puzzle. Due to the unique nature of reasoning required for different types of wordplay, we crafted specific reasoning pathways tailored to each wordplay category (see Appendix C details).

The model experienced a **high mental load**, producing brief and occasionally inaccurate responses at each reasoning step. Additionally, it **struggled with concise operations** such as rearranging letters to form words or extracting initial letters from sentences.

- **Iterative Prompting (Iterative Reasoning)**

  To alleviate cognitive load, we employed an iterative prompting strategy, segmenting the reasoning process into smaller prompts and sequentially building on previous responses. We found that when the LLM was asked to format its output for use in the next step, accuracy diminished. To mitigate this, we devised a **two-agent system**: one agent produces the answers, while the second agent organizes the extracted information in a structured format.

- **Hybrid Approach**

  To address challenges with concise operations, we substitute certain reasoning steps with automated programs where feasible.

However, Clues often combine several types of wordplay, making the reasoning process far more complex than simply merging steps for each type. We soon recognized that manually constructing reasoning paths for every possible type of clue was unfeasible.

3.2.2 Auto (CoT) Strategy Selection

We aim to teach the model to **autonomously determine its reasoning paths** and intelligently follow its planned steps through **in-context learning**. To facilitate this, we have developed a structured five-step framework and curated 53 examples in this format from the book How to Master the Times Crossword: The Times Cryptic Crossword Demystified [Moorey (2008)], each showcasing a unique reasoning path.

- **In context learning**

  Given the token limit, only about 30 examples can be included as context at a time, so we randomly select 30 from the total of 53 for each trial. While **fine-tuning** is available to handle more examples, our analysis of the model's performance shows that with 30 examples, the model sufficiently understands and executes auto strategy selection. The real issue, as with handcrafted CoT, lies in the **high mental load** and **difficulty in concise operations** while executing the reasoning steps it planed, rather than in a lack of fine-tuning.

- **Auto Strategy Selection + Iterative Prompting? Inference Scaling!**

  To address this challenge, we propose using iterative prompting to alleviate the high cognitive load, employing a 'think and re-evaluate' mechanism for iterative answer refinement.

  Coincidentally, as our project concludes, OpenAI introduced the O1 model, designed to enhance reasoning through **inference scaling**, which also excels in cryptic crosswords(Clifford (2024)). So we also briefly tested the newly released o1 model to compare our results.

4 Results

We conducted our experiments across four LLMs. Due to budget constraints, we couldn't test on the entire dataset. However, within our available resources, we maximized the number of experiments to gain meaningful insights.

4.1 HandCrafted Chain of Thought

We tested each handcrafted reasoning chain on problems for which it was designed. The results for the CoT + Iterative Prompting + Hybrid approach are shown in Table 2.

4.2 Auto Strategy(CoT) Selection

For auto strategy selection though in context learning, we evaluated 200 samples, and the results are presented in Table 3. Nevertheless, we observe that the GPT-o1 models achieve higher accuracy in the number of letters, even when the answers are incorrect.

4.3 Comparison

Modern LLMs inherently employ a CoT approach when solving cryptic crossword puzzles. We have compared the accuracy of our method with that of zero-shot question answering; the comparative results are presented in Table 3.

5 Discussion

Despite various attempts to improve model performance beyond the 7.64% accuracy baseline, we found that the core issue may lie in the model's complexity, so we moved towards LLMs. Although handcrafted Chain-of-Thought prompting was effective, it is not scalable for all puzzle types. Auto Strategy Selection through In-Context Learning provided STOA performances but remains limited. Combining Auto Strategy selection with reflection offers more promising results.

6 Conclusion and Future Work

Our project demonstrated that solving cryptic crossword puzzles with large language models remains a difficult task. The most promising future direction would be to combine auto strategy selection and some form of inference scaling, to increase the model's reasoning ability and accuracy on complex reasoning tasks such as cryptic crossword puzzles.

References

Matthew Clifford. 2024. Tweet by matthew clifford. https://x.com/matthewclifford/status/1834485810113990786. Accessed: 2024-09-15.

Robin Deits. 2022. Crypticcrosswords.jl. https://github.com/rdeits/CrypticCrosswords.jl. GitHub repository.

Avia Efrat, Uri Shaham, Dan Kilman, and Omer Levy. 2021. Cryptonite: A cryptic crossword benchmark for extreme ambiguity in language.

T. Moorey. 2008. *How to Master the Times Crossword: the Times Cryptic Crossword Demystified (the Times Crosswords)*. The Times Crosswords Series. HarperCollins Publishers Limited.

Joshua Rozner, Christopher Potts, and Kyle Mahowald. 2021. Decrypting cryptic crosswords: Semantically complex wordplay puzzles as a target for nlp.

Shuchi. 2008. Crossword Unclued. [Online; accessed 12-September-2024].

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

Wikipedia contributors. 2024. Cryptic crossword — Wikipedia, The Free Encyclopedia. [Online; accessed 12-September-2024].

A Data Analysis

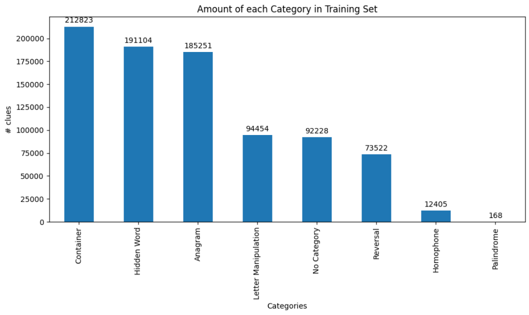

Figure 1: Amount of each category that was identified by a word that is indicative.

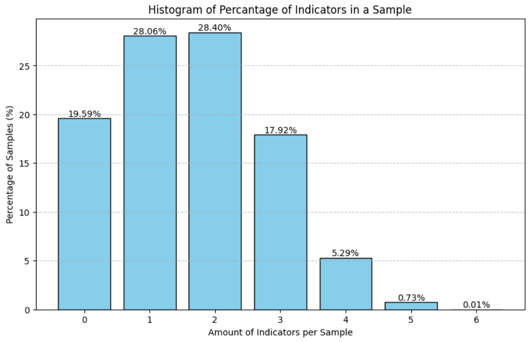

Figure 2: The percentage of amount of indicator that valid for a given clue. There are 71.94% of the clues without or more than one indicator.
Example for a 6 indicators clue: *recoil from report about dissolute romeo (9)*, where *out* in 'Anagram', *is* in 'Container', *from* in 'Hidden Word', *report* in 'Homophone', *recoil* in 'Reversal' and *about* in 'Letter Selection'.

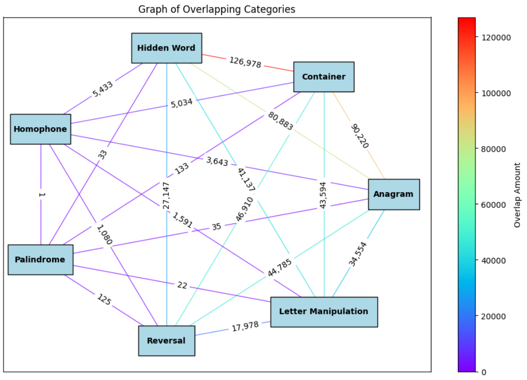

Figure 3: The division between overlapping in specific categories. we can see that Letter Manipulation and Hidden Words have the greater overlap among the categories, it's because words like *in* that indicative for both.

B Results

Model	Method	lr	batch size	Epochs #	Train sample size #	Validation sample size #	accuracy
T5-large	full	5e-03	7000 token	10	470804	26204	0.0764
Bart-base	full	5e-05	16	5	470804	26204	0.0137
Bart-large-cnn	LoRA	5e-04	16	3	470804	26204	0.0
T5-small	full	5e-05	16	5	470804	26204	0.0057
T5-large	LoRA	5e-04	16	3	48000	26204	0.0121
Roberta-base*	LoRA	5e-03	16	5	48000	26204	0.0

Table 1: Hyper Parameters and results For Each Model: We can see that all the model are not performing optimally.

Model	Anagram	Charade	Double Definition	Hidden	LM-Initial/Final	LM-Even/Odd	Palindrome
number of samples	20	20	13	20	20	20	20
gpt-4o-2024-08-06	55%	60%	38%	75%	100%	75%	75%
gpt-4o-mini	15%	20%	23%	55%	100%	90%	10%
gemini-1.5-pro	25%	65%	30%	25%	90%	70%	55%
gemini-1.5-flash	5%	40%	23%	30%	95%	90%	45%

Table 2: The table shows the performance of four models using tailored chain-of-thought prompts for different cryptic clue types. GPT-4o-2024-08-06 consistently performs well across most categories, while smaller models like GPT-4o-mini and Gemini-1.5 variants show varying success, particularly excelling in pattern-based clues such as LM-Initial/Final and LM-Even/Odd.

Model	number of samples	Acc. for ICL	ICL + ref	Acc. for Zero-Shot QA
gemini-1.5-pro	200	14.5%	17%	6%
gpt-4o-2024-08-06	200	6%	8.5%	7%
gemini-1.5-flash	200	21.5%	29.5%	23%
gpt-4o-mini	200	8%	10%	9.5%
gpt-o1-preview	10	N/A	N/A	40%
gpt-o1-mini	10	N/A	N/A	30%

Table 3: ICL stands for in-context learning. ICL+ ref stands for answers after reflection. The results show that ICL will casue a significant improvement in gemini-1.5-pro, while other models perform similar to zero shot QA. But the accuracy after reflection is higher than zero shot QA for all models.

C Solving Cryptic Puzzles

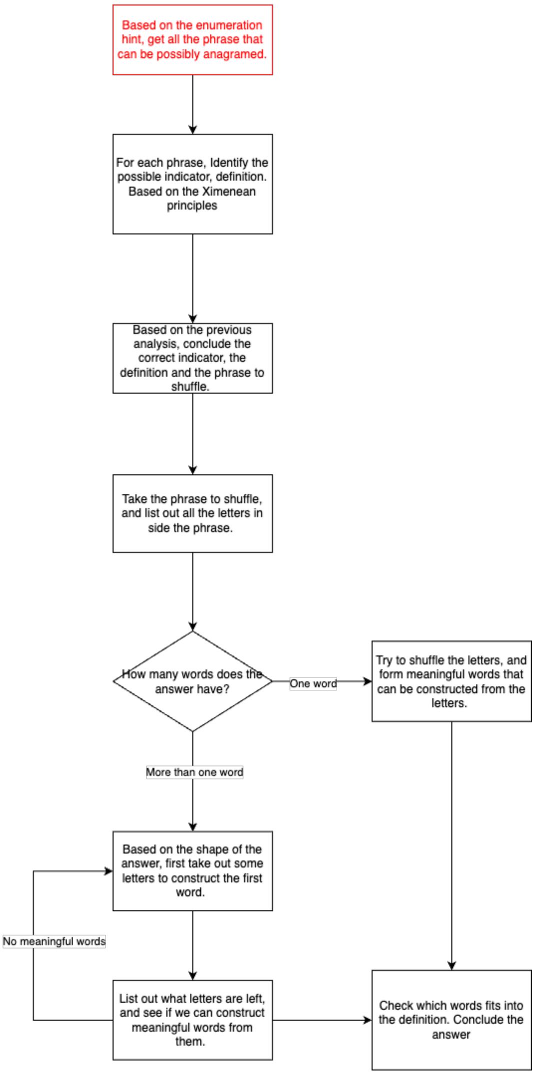
Based on the enumeration hint, get all the phrase that can be possibly anagramed.
For each phrase, Identify the possible indicator, definition. Based on the Ximenean principles
Based on the previous analysis, conclude the correct indicator, the definition and the phrase to shuffle.
Take the phrase to shuffle, and list out all the letters in side the phrase.
How many words does the answer have?
One word
Try to shuffle the letters, and form meaningful words that can be constructed from the letters.
More than one word
Based on the shape of the answer, first take out some letters to construct the first word.
No meaningful words
List out what letters are left, and see if we can construct meaningful words from them.
Check which words fits into the definition. Conclude the answer

Figure 4: Anagram CoT

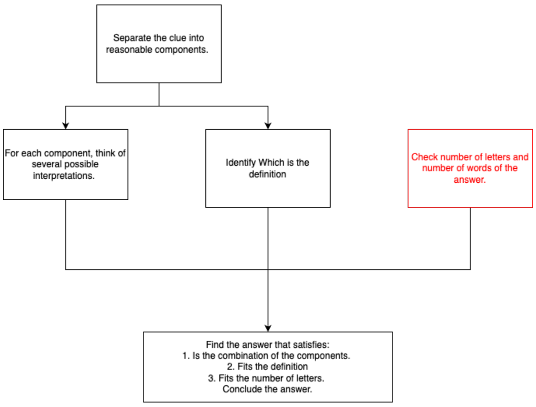
Separate the clue into reasonable components.
For each component, think of several possible interpretations.
Identify Which is the definition
Check number of letters and number of words of the answer.
Find the answer that satisfies:
1. Is the combination of the components.
2. Fits the definition
3. Fits the number of letters.
Conclude the answer.

Figure 5: Charade CoT

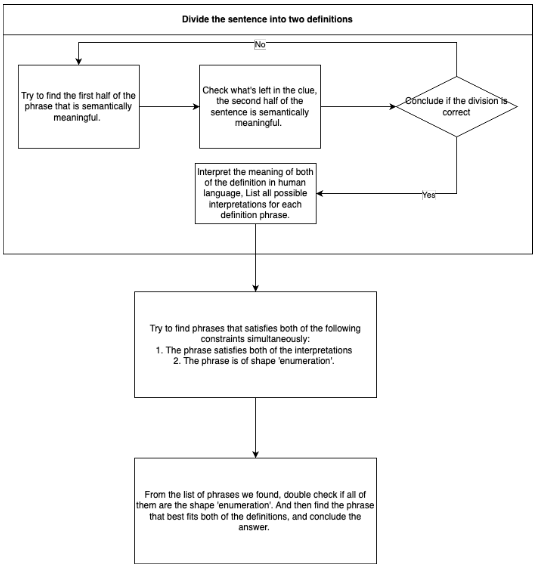
Divide the sentence into two definitions
No
Try to find the first half of the phrase that is semantically meaningful.
Check what's left in the clue, the second half of the sentence is semantically meaningful.
Conclude if the division is correct
Yes
Interpret the meaning of both of the definition in human language, List all possible interpretations for each definition phrase.
Try to find phrases that satisfies both of the following constraints simultaneously:
1. The phrase satisfies both of the interpretations
2. The phrase is of shape 'enumeration'.
From the list of phrases we found, double check if all of them are the shape 'enumeration'. And then find the phrase that best fits both of the definitions, and conclude the answer.

Figure 6: Double Definitions CoT

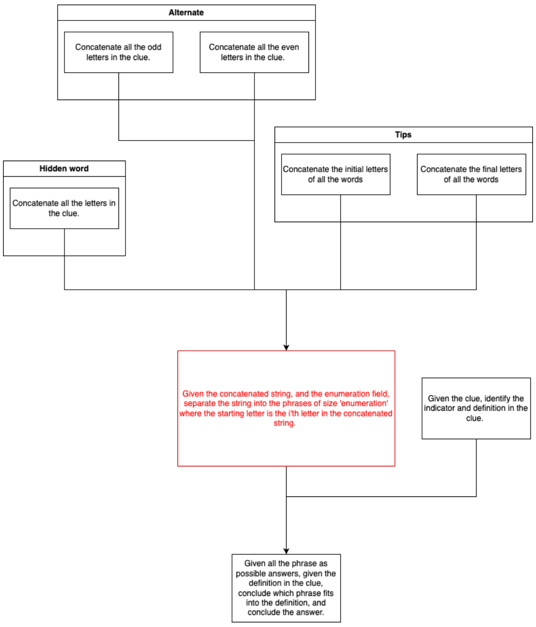
Alternate
Concatenate all the odd letters in the clue.
Concatenate all the even letters in the clue.
Tips
Concatenate the initial letters of all the words
Concatenate the final letters of all the words
Hidden word
Concatenate all the letters in the clue.
Given the concatenated string, and the enumeration field, separate the string into the phrases of size 'enumeration' where the starting letter is the i'th letter in the concatenated string.
Given the clue, identify the indicator and definition in the clue.
Given all the phrase as possible answers, given the definition in the clue, conclude which phrase fits into the definition, and conclude the answer.

Figure 7: Hidden Words Combined CoT

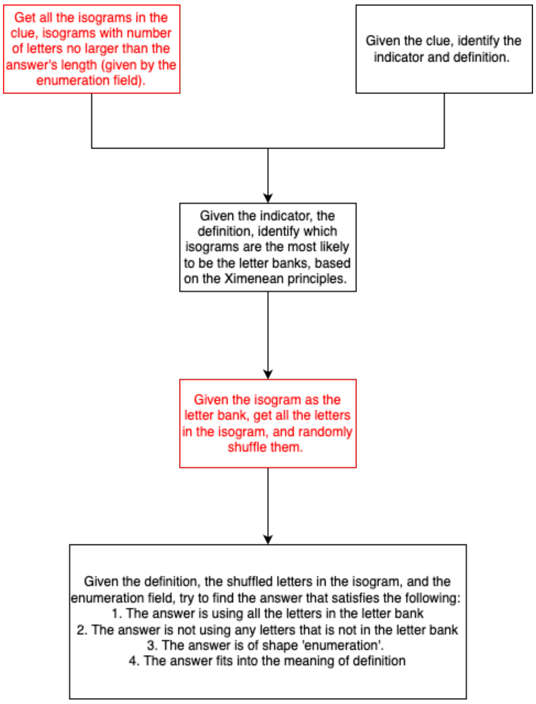
Get all the isograms in the clue, isograms with number of letters no larger than the answer's length (given by the enumeration field).
Given the clue, identify the indicator and definition.
Given the indicator, the definition, identify which isograms are the most likely to be the letter banks, based on the Ximenean principles.
Given the isogram as the letter bank, get all the letters in the isogram, and randomly shuffle them.
Given the definition, the shuffled letters in the isogram, and the enumeration field, try to find the answer that satisfies the following:
1. The answer is using all the letters in the letter bank
2. The answer is not using any letters that is not in the letter bank
3. The answer is of shape 'enumeration'.
4. The answer fits into the meaning of definition

Figure 8: Letter Banks CoT

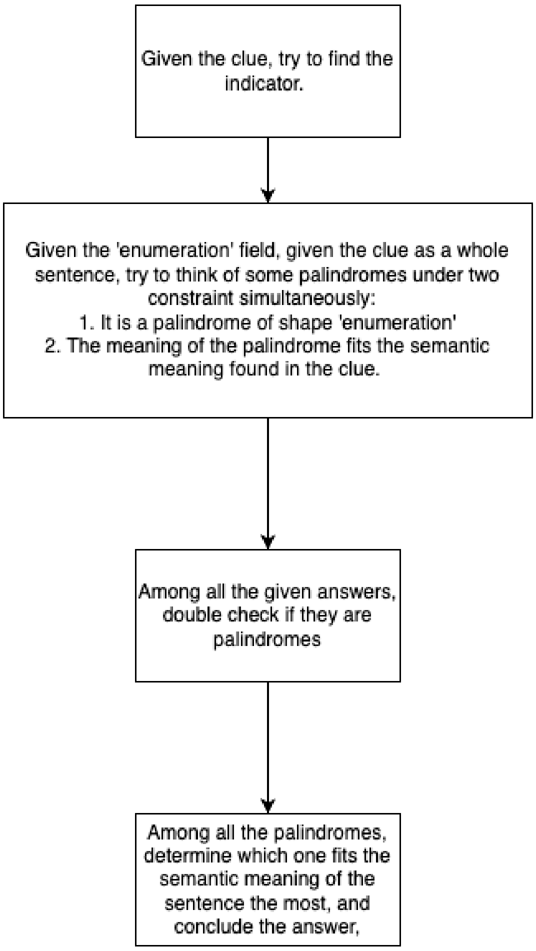
Given the clue, try to find the indicator.
Given the 'enumeration' field, given the clue as a whole sentence, try to think of some palindromes under two constraint simultaneously:
1. It is a palindrome of shape 'enumeration'
2. The meaning of the palindrome fits the semantic meaning found in the clue.
Among all the given answers, double check if they are palindromes
Among all the palindromes, determine which one fits the semantic meaning of the sentence the most, and conclude the answer,

Figure 9: Palindrome CoT

D Examples of Clues

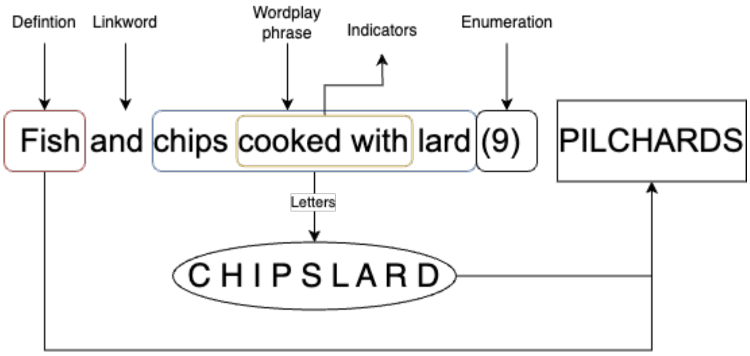

Figure 10: **Anagram.** This type usually have indicator words that suggests mixing of letters. Tha anagram puzzle requires mix the letters of wordplay words to form the answer

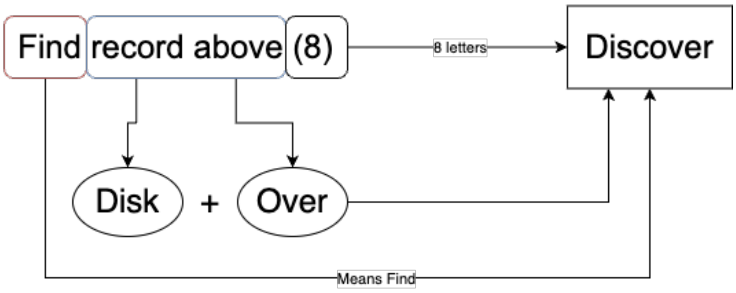

Figure 11: **Charade.** This type does not have indicators. The Charade puzzle requires associating words from wordplay words, and combine them to form the answer

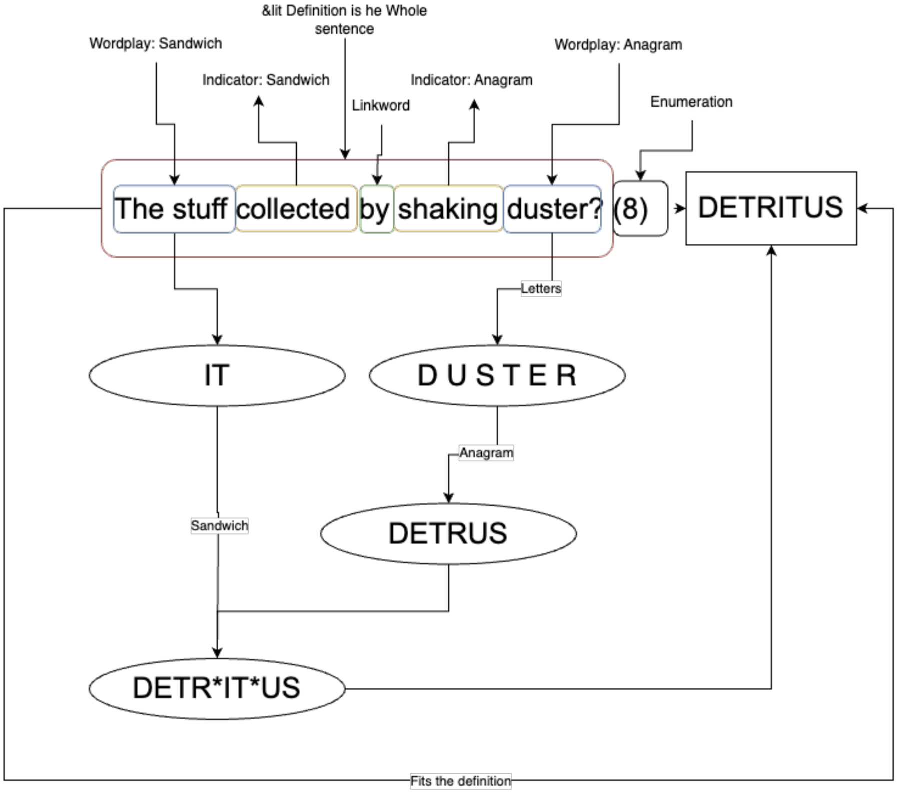

Figure 12: **Multiple Wordplays: Anagram + &lit + Sandwich.** In this example, we need to first perform anagram on "duster" to get "detrus", and then associate "this stuff" with "it", finally, we will stuff "IT" into "DETRUS" to get the word "DETRITUS". And the answer fits the definition, which is the entire clue.

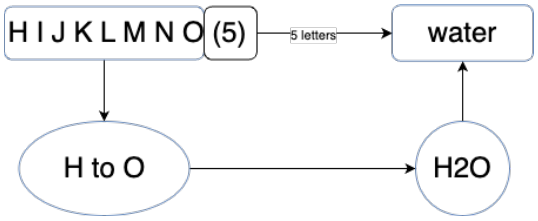

Figure 13: **Novelty.** The novel types of cryptic crossword puzzles doesn't have specific structure or pattern. There are hardly any way to define a CoT reasoning path for them.