ClueGPT

Final Report

ECE1786H

Claire Phillips and Catherine Yeh

December 12th 2023

Word Count: 1955 (excluding references, tables)

Introduction (2 points)

The goal of ClueGPT is to supply people solving crosswords with an additional clue.

People get stuck when solving crossword puzzles and are left with few options on how to proceed. Often, puzzlers will be left to google the clue and be given the answer outright.

This situation can be frustrating to people who enjoy solving puzzles and don't want to feel like they are giving up or cheating. ClueGPT is an alternative to this and can be used to get help without being given the answer.

Illustration / Figure (2 points)

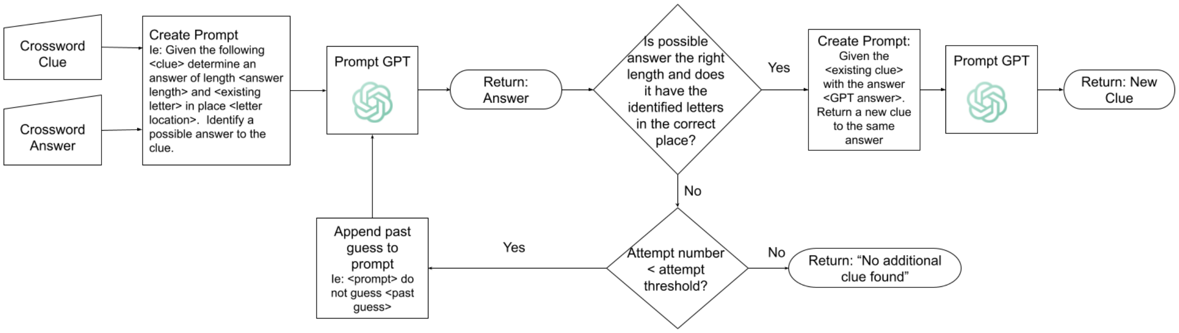

Crossword Clue
Crossword Answer
Create Prompt
Ie: Given the following <clue> determine an answer of length <answer length> and <existing letter> in place <letter location>. Identify a possible answer to the clue.
Prompt GPT
Return: Answer
Is possible answer the right length and does it have the identified letters in the correct place?
Yes
Create Prompt: Given the <existing clue> with the answer <GPT answer>. Return a new clue to the same answer
Prompt GPT
Return: New Clue
No
Attempt number < attempt threshold?
No
Return: "No additional clue found"
Yes
Append past guess to prompt Ie: <prompt> do not guess <past guess>

Figure 1 - ClueGPT.

Background & Related Work (2 points)

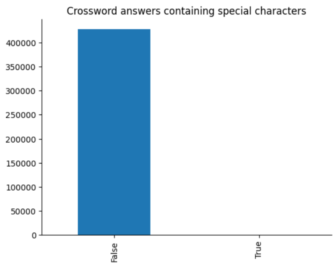
There is existing literature that looks at the problem of solving crossword puzzles and on crossword clue generation. For example, [1] proposes to solve a crossword puzzle by breaking down the problem into two subtasks. The first subtask is a question-answering task, where, given a clue, the system generates a list of possible answers for the given clue. For this subtask, the paper experimented with fine-tuned sequence-to-sequence models, BART [2] and T5 [3], as well as retrieval-augmented generation models (RAG). The second subtask is to select the answers that satisfy the given set of constraints. This paper suggests that using RAG and additional processing of the data (i.e. removing diacritics, punctuation and whitespace) leads to a better performance for both subtasks. However, there is room for improvement. [4] proposes a method to generate clues for cryptic crosswords. Cryptic crosswords mean that the clues consist of two elements: a definition of the answer (surface reading), and a wordplay element that suggests the same answer (a hidden meaning) [5]. It first generates a list of hidden meaning clues using strategies such as Charade and Anagram, then it generates the surface reading component and ensures it is semantically correct using additional resources such as a Collocational Semantic Lexicon, which is used to ensure the right words are in the right places in a sentence.

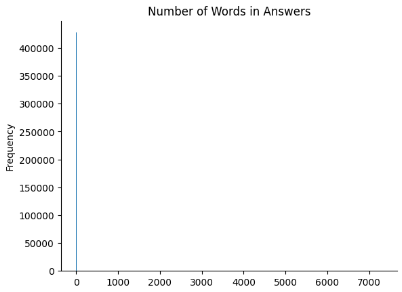
The task of generating clues with current AI technology has not yet been investigated to our knowledge. Hence in this project, we plan to explore the area of clue generation and the value of it to help people solve puzzles.

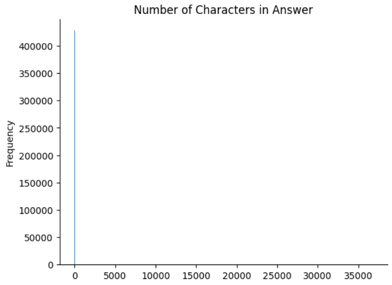
Data and Data Processing (4 points)

Our dataset is the New York Times crossword data [1]. The data is already split into train test and validation files (80/20 split).

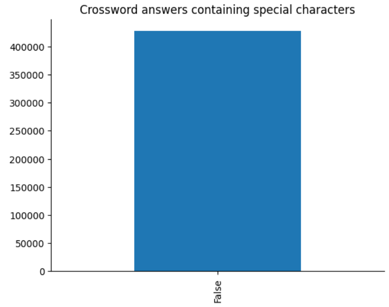
Evaluating the training answers dataset revealed a small number of outliers, likely data entry errors. These outliers have been removed by excluding answers containing special characters or which are over 50 characters long. Less obvious incorrect answers require manual identification. An example of this in the data is an answer “cr ation” which should be “creation”. About 1% of answers exhibit errors.

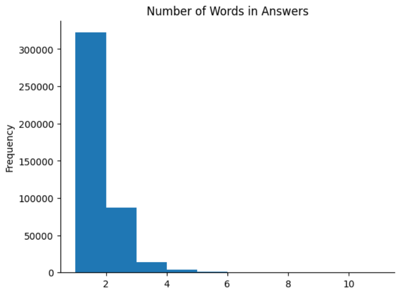


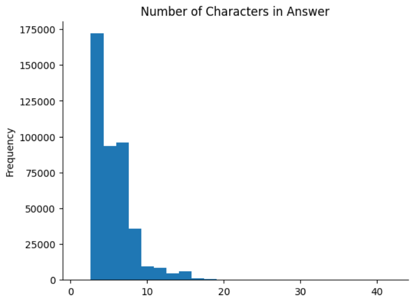




Removing outliers results in better training data. The plots above show the original distribution of features of the answer data, and the plots below show the distribution of features of the cleaned answer data.







To utilize the data, spaces are removed from the answer and the characters are made lowercase. Additionally, the length of the answer and some letters from the answer are included in the prompt.

Below is an example of cleaned data:

Clue	Original Answer	Processed Answer	Length of Answer	Supplied Letters
Kind of omlet	Egg White	eggwhite	8	{0: e}

Architecture and Software (4 points)

The diagram of ClueGPT is illustrated in the figure below. We first prompt GPT-4 to guess an answer based on the original clue, answer length and letter positions. If the guessed answer is correct, we ask GPT-4 to generate a new clue. If the guess is incorrect, we update the prompt to ask GPT-4 to guess an answer that is not any of the past incorrect guesses. We allow it to retry up to the maximum number of attempts. If it could not guess a correct answer, it returns “no additional clue found”.

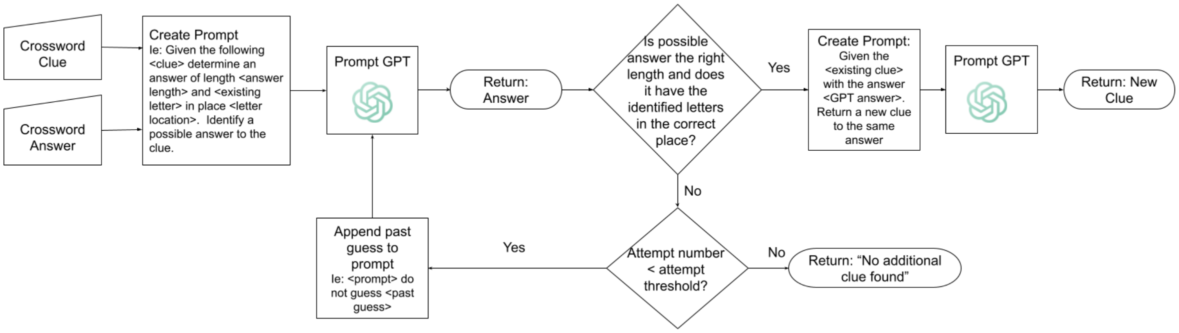
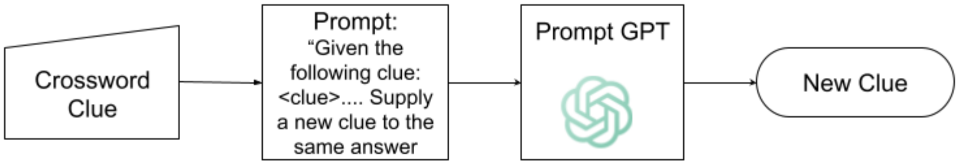

Crossword Clue
Crossword Answer
Create Prompt
Ie: Given the following <clue> determine an answer of length <answer length> and <existing letter> in place <letter location>. Identify a possible answer to the clue.
Prompt GPT
Return: Answer
Is possible answer the right length and does it have the identified letters in the correct place?
Yes
Create Prompt: Given the <existing clue> with the answer <GPT answer>. Return a new clue to the same answer
Prompt GPT
Return: New Clue
No
Attempt number < attempt threshold?
Yes
Append past guess to prompt
Ie: <prompt> do not guess <past guess>
No
Return: “No additional clue found”

Figure 4.1: ClueGPT architecture diagram.

Baseline Model or Comparison (2 points)

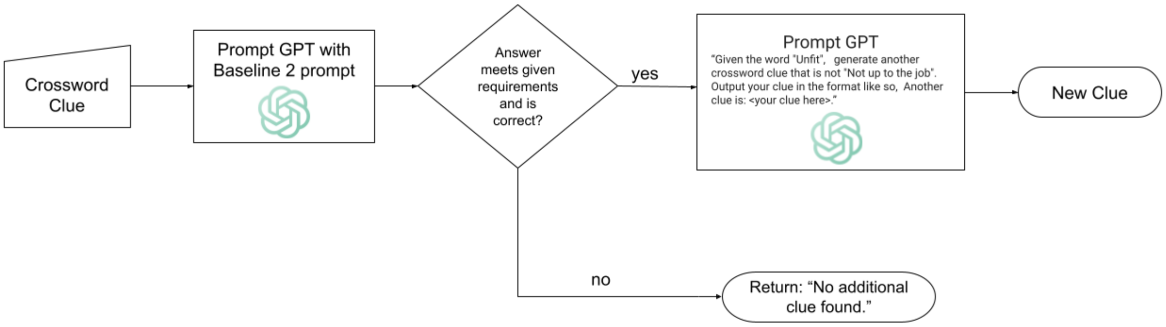
We compare ClueGPT with two different methods of prompting, Version 0 and Version 1. Both were provided with the existing clue, answer length, and the letter positions that were selected randomly at the data processing step. Version 0 uses zero-shot prompting on GPT-4 to generate a new clue. This method does not validate GPT-4's guesses.



"Given the following clue: <clue>.... Supply a new clue to the same answer

Figure 3.1: Version 0 architecture.

Version 1 uses chain-of-thought prompting, where it first prompts GPT-4 to generate an answer. If the answer is correct, it prompts GPT-4 to generate a clue, if the answer is incorrect, it outputs "no additional clue found".


Crossword Clue
Prompt GPT with Baseline 2 prompt
Answer meets given requirements and is correct?
yes
Prompt GPT
New Clue
no
Return: "No additional clue found."

"Given the word "Unfit", generate another crossword clue that is not "Not up to the job". Output your clue in the format like so, Another clue is: <your clue here>."

Figure 3.1: Version 1 architecture.

Quantitative Results (4 points)

The output for each clue was measured qualitatively via the following definition of success: A clue that leads to correct answers, but cannot give away the answer explicitly (eg. answer: 'ous'; clue: Common ending for words like "dangerous" or "hilarious".).

We compare the success rate of ClueGPT with the two other versions described earlier. The performance across the models was somewhat similar, with our best model (ClueGPT) only generating a correct clue for 70% of the answers. However, it never generated a clue we defined as being incorrect like Version 0, and has generated clues for 20% more clues than Version 1.

Table 1 - Comparison of three models.

	Version 0	Version 1	ClueGPT
Percent of generated clues which meet the definition of success	65.7%	88%	88%
Percent of clues the model did not generate a clue.	0%	38%	18%
Percent of clues generated which meet the definition of success and are not a “No clue generated” response	65.7%	50%	70%

[Please note that we made some miscalculation in our final presentation which have been correct in this report]

Table 2 - The number of unique guesses it took to find an answer (correct, wrong or no answer) given 12 guesses.

Guess(s)	Correct Answer	No Answer	Wrong Answer
1	58%		6%
2	8%		2%
3	2%		
4	2%		
9			18%
12 (maximum)		18%	

Qualitative Results (4 points)

Table 3 shows three cherry picked example outputs from the best ClueGPT model. Table 3 shows an incorrect clue in the last row, where the guessed clue is wrong and thus the created clue is wrong. The clues created for correct answers were correct 100% of the time.

Table 3 - Example outputs from ClueGPT

Original Clue	Letter Positions Given	ClueGPT Guessed Answer	ClueGPT Output	Answer
Au naturel	{1: 'a'}	bare	Another clue is: "Uncovered or exposed"	bare
___ said ...	{1: 'e'}	easier	Another clue is: Less difficult.	easier
Capital on the Mediterranean	{3: 'i'}	tunisia	Another clue is: "North African country known for its gold and olive oil production".	algiers

System Prompt for Answer Identification

Table 4 - Cherry picked examples of ClueGPT outputs including non answer text in response

ClueGPT answers	Correct Answer	System Prompt	Prompt
['batwoman', 'aquaman', 'superman', 'catwoman', 'The answer could be "Shazam', 'The answer could be "Batman".', 'The answer could be "Jagaman', 'The answer could be "Batlama".', 'The answer could be "Fatmama', 'The answer could be "Aquaman".']	starman	You are a helpful assistant who guess answers to crossword clues without any additional text.	Guess an answer to the following crossword clue which is 7 letters long and includes letters a as letter 3, m as letter 5, a as letter 6. and is not 'The answer could be "Fatmama', 'The answer could be "Shazam', 'aquaman', 'The answer could be "Batlama".', 'catwoman', 'The answer could be "Batman".', 'The answer could be "Jagaman', 'The answer could be "Aquaman".', 'superman' or 'batwoman': DC Comics superhero
['The answer to your crossword clue could be', 'immesh']	gobitween	You are a helpful assistant who guess answers to crossword clues without any additional text.	uess an answer to the following crossword clue which is 6 letters long and includes letters i as letter 1, m as letter 2, m as letter 3, e as letter 4, s as

			letter 5. and is not 'The answer to your crossword clue could be': Tangle in a net: Var.
['The answer is probably "Gobi Tween', 'The answer could be "Hobi Tween', 'The answer could be "Abbi Tween', 'The answer could be "Tobi Tween', 'The answer could be "Kibi Tween', 'The answer could be "Gobi Tween', 'The answer could be "Gobi Tween', 'The answer could be "Gobi Tween', 'The answer could be "Obbi Tween', 'The answer could be "Gobi Tween']		You are a helpful assistant who guess answers to crossword clues without any additional text.	Guess an answer to the following crossword clue which is 9 letters long and includes letters b as letter 3, i as letter 4, w as letter 6, e as letter 8, n as letter 9. and is not 'The answer could be "Obbi Tween', 'The answer could be "Abbi Tween', 'The answer could be "Tobi Tween', 'The answer could be "Gobi Tween', 'The answer is probably "Gobi Tween', 'The answer could be "Kibi Tween' or 'The answer could be "Hobi Tween': 11- or 12-year-old Mongolian desert dweller?

Table 5 - examples of outputs from varying system prompts with all other parameters kept consistent and the input clue being “Amerind shoe”

System Prompt	ClueGPT Answers	Correct Answer
You are a helpful assistant who guess answers to crossword clues without any additional text.	'moc'	moc
You are a helpful assistant.	'The answer to the crossword clue "A', 'The answer to the crossword clue could be', 'The answer to the crossword clue could be', 'The answer to this crossword clue is '"', 'The answer to the crossword clue "A', 'The answer to the crossword clue could be', 'The answer to the crossword clue could be', 'The answer to the crossword clue could be', 'The answer to the crossword clue "A', 'The answer to the crossword clue could be', 'The answer to the crossword clue could be', 'The answer to the crossword clue could be'	moc
You are a helpful assistant who guess answers to crossword clues.	'The answer to the crossword clue is '"', 'The answer to the crossword clue is '"', 'The answer to the crossword clue "A', 'The	moc

	answer to the crossword clue "A', 'The answer to the crossword clue "A', 'The answer to the crossword clue "A', 'The answer to the crossword clue "A', 'The answer to the crossword clue "A', 'The answer to the crossword clue "A', 'The answer to the crossword clue "A', 'The answer to the crossword clue "A', 'The answer to the crossword clue "A']	

Table 6 - Results from different system prompts on a sample of 10.

label	System prompt	Correct Answer Guessed Rate
Best system prompt	You are a helpful assistant who guess answers to crossword clues without any additional text.	60%
	You are a helpful assistant.	30%
	You are a helpful assistant who guess answers to crossword clues.	40%

The above tables show the impact of the system prompt on the first step of the ClueGPT model of selecting the correct crossword answer given the crossword clue. Table 6 shows that the best system prompt explicitly asked for a crossword answer and for no additional text. Table 6 gives examples of the response from GPT with varying system prompts. When the system prompt did not explicitly ask for an answer without any additional text, the responses often included auxiliary text (ie responses including “The answer is:”). Responses including auxiliary did occur with the best system prompt, as shown in table 4. Of the sample set the examples from table 5 were cherry picked from, all the samples which did not return an answer included additional text in the response. From these incorrect answers, only one response had a correct answer, shown in the last row of table 4. However, the ClueGPT memory function passes in the entire set of past guesses, so this auxiliary text leads to worse prompts (and likely responses) in iteration.

Top P and Temperature for Answer Identification

Table 7 - Identification of correct answer given varying top p and temperature values

Top P\ Temperature	Temeprature = 0.1	Temperature = 0.8	Temperature = 1.2	Temperature = 2.0
Top_p = 0.1	No Answer =40% Correct Answer =60%	No Answer = 30% Correct Answer = 60% Wrong Answer = 10%	No Answer = 30% Correct Answer = 60% Wrong Answer = 10%	No Answer =40% Correct Answer =60%
Top_p=1	No Answer = 30% Correct Answer = 70%	No Answer = 20% Correct Answer = 70% Wrong Answer = 10%	Correct Answer = 80% Wrong Answer = 20%	No Answer = 10% Correct Answer = 60% Wrong Answer = 30%

Table 8 - example answers given clue: "A place of rest" with varying top_p and temperature values.

Top P and Temperature	GPT Clue Guesses	Answer Selected	Correct Answer Identified

Top_p = 0.1
Temperature = 0.1
	['chair', 'chase', 'The answer could be "chair".', 'The answer could be "chair".', 'The answer could be "chair".', 'The answer could be "chair".', 'The answer could be "chair".', 'The answer could be "chair".', 'The answer could be "chair".', 'The answer could be "chair".', 'The answer could be "chair".', 'The answer could be "chair".']	None	No (no answer)

Top_p =1
Temperature = 0.1
	['chair', 'chase', 'shade', 'haven']	haven	Yes (correct answer)

Top_p = 1
Temperature = 2.0
	['chair', 'The answer could be "chase".', 'harem']	harem	No (wrong answer)

The above tables show the results of varying top_p and temperature for the ClueGPT step of guessing the correct crossword clue. The top_p value acts as a cut off, where a lower top_p value will limit the number of possible tokens in the set which can randomly be selected given their cumulative probability. Temperature is similar to statistical entropy, where as the temperature increases, the probability of choosing any token converges, which makes choosing a token which is not the most likely token increases. (Entropy can be thought as measuring the amount of surprise data in a variable [2]). Through performing a grid search over several values of temperature and top_p, over a sample size of 10, the best results came from a top_p of 1 and temperature of 0.1. The results of the grid search are displayed in table 7. A large temperature lead to an increase in wrong answers, while a decrease in top_p lead to an increase in no answer.

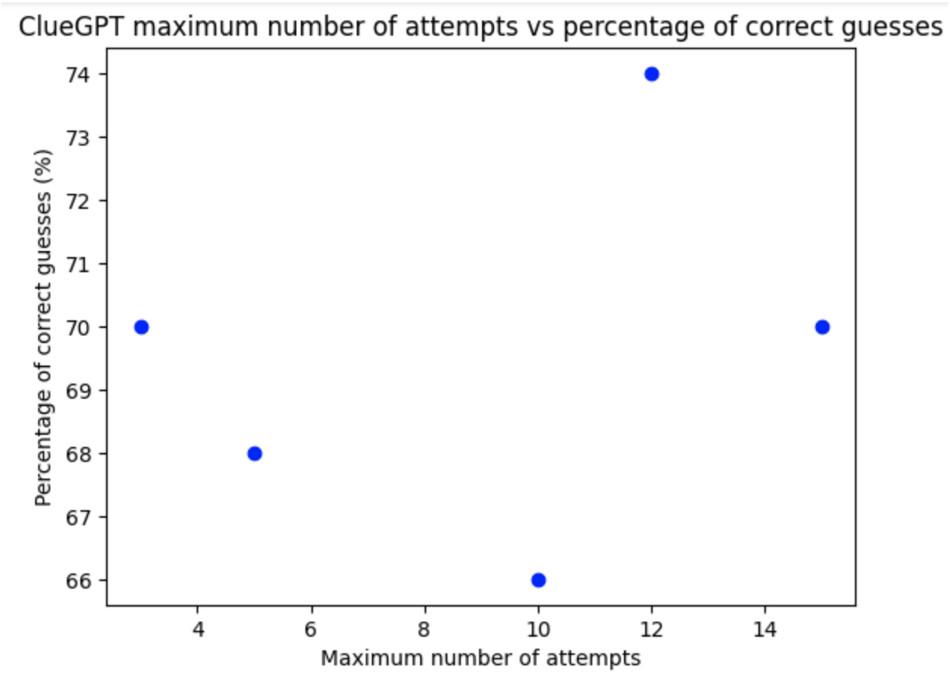
Table 8 shows the responses of a GPT with the same system and input prompts but varying top_p and temperature values. The model with a low top_p and temperature added additional text to the answers and chose the same answer several times (despite the prompted memory). The model with a high top_p and temperature selected a response which did not relate to the prompt.

Discussion and Learnings (4 points)

The results in general show that GPT performs quite well at creating a new crossword clue based on an existing clue and correct answer (100% of the new clues were correct if GPT knew the correct crossword answer) but is not as skilled at guessing the correct answer given an existing clue and information regarding the answer (length of clue and existing letters). This makes sense, as GPTs are designed to perform well at generating new content, rather than creating an exact correct sequence. Guessing the answer to a crossword requires returning the exactly correct text. Further, ClueGPT will not identify the answer at this step if there is any auxiliary text. Despite this, the final model still performed well at both tasks.

To allow for the GPT API model to perform the task it was not built to perform we needed to optimize several parameters - top_p, temperature, the system prompt and the content prompt and build architecture around the model to allow for memory. The GPT API we used (gpt-4 from the OpenAI library using python) did not contain a memory function, so each prompt was its own gpt session. This impacted how the ClueGPT was built, as the past attempts to guess a clue needed to be included in the input prompt. This lack of memory also impacted the input prompt as we needed to explicitly ask the GPT to exclude additional text everytime in the prompt. Even when asked, the GPT would regularly return auxiliary words. To improve ClueGPT, additional architecture could be built to ensure correct answers and formatting. This could include further prompt engineering, adding regex parsing to identify what text is the crossword answer or adding another LLM after the first GPT call to pull out the explicit answer and validate the guessed answers.

Through creating the version 0 model, we also gained a deeper understanding about GPT’s inclination to posture, especially with the default system prompt of “you are a helpful assistant”. Despite being very incorrect several times, the GPTs never returned a message like “I do not know the answer”.



Plot 1 - The maximum number of attempts to predict the answer vs the number of correct guesses.

Plot 1 shows the maximum number of attempts ClueGPT was allowed to take to predict the correct crossword answer against the percentage of correct guesses. There is no clear trend in this data. Table 2 in the qualitative results section, shows that for one model, most correct answers are identified by the first guess (58%) and by the 5th guess, no more correct answers were guessed but the number of incorrect guesses increased. If this model were to be built again, lowering the maximum attempts threshold should likely be lowered to decrease the number of incorrect guesses.

Individual Contributions (10 points)

Claire preprocessed the crossword dataset, came up with the ClueGPT logic, implemented Version 0, and analyzed the outputs from ClueGPT, labeled if each clue is a success or not.
Catherine implemented Version 1, analyzed the outputs from Version 0 and 1, and labeled clues.
Both investigated how to work with the GPT API, coded ClueGPT, and ran prompt and parameter experiments.

Permissions

- Permission to post video
  - yes/no or wait till see video
  - Catherine: yes
  - Claire: yes
- Permission to post final report: yes no
  - Catherine: yes
  - Claire: yes
- Permission to post source code: yes no
  - Claire: yes
  - Catherine:yes

Bibliography

[1] GPT-3: All you need to know about ai language model. Sigmoid. (2023, December 6). https://www.sigmoid.com/blogs/gpt-3-all-you-need-to-know-about-the-ai-language-model/

[2] Seth, N. (2023, November 3). *Entropy in machine learning: Definition, examples and uses*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/11/entropy-a-key-concept-for-all-data-science-beginners/#:~:text=Entropy%20measures%20the%20amount%20of,higher%20uncertainty%20have%20higher%20entropy.