Evaluation of Large Language Models in Cooperative Language Games

Samuel Knoche
Independent

Abstract

This report investigates the potential of cooperative language games as an evaluation tool of language models. Specifically, the investigation focuses on LLM's ability to both act as the "*spymaster*" and the "*guesser*" in the game of Codenames, focusing on the *spymaster*'s capability to provide hints which will guide their teammate to correctly identify the "*target"* words, and the *guesser's* ability to correctly identify the *target* words using the given hint. We investigate both the capability of different LLMs at self-play, and their ability to play cooperatively with a human teammate. The report concludes with some promising results and suggestions for further investigation.

Keywords: Scale oversight, benchmarks, ML safety

1. Introduction

Recent advances in natural language processing (NLP) have enabled the development of language models (LMs) that can generate human-like language. These models have been used in a variety of applications, including natural language understanding and dialogue systems. Recent improvements of LLMs by fine-tuning them with human demonstration and feedback to better follow instructions have opened up the possibility to use LLMs on increasingly complicated tasks. Here, we investigate their ability at the game of Codenames.

Codenames is a word-based party game that requires two teams of players to work together in order to guess each other's secret agents. In the original game, 25 words are set in a grid on the table. Each team selects a *spymaster* who is shown which words correspond to secret agents of the other team. The *spymasters* take turns giving one-word clues that relate to multiple agents on the board, and a number corresponding to how many words the clue word refers to. The team must then guess the correct codenames before their opponents can.

2. Methods

For our purposes, we formulate the game into a one-team word-based game. The *spymaster* is given a selection of words and they must provide their teammate with a single word clue and a number (1 or 2) to indicate the number of *targets* the word relates

to. To provide an effective clue, the *spymaster* should consider the *target* words and think of synonyms, rhymes, or other creative uses of language.

We generate a prompt (see Appendix 1 for an example) with a random selection of 15 words out of a dataset of 400 words.[1] Out of the 15 words, 5 words are randomly selected to serve as "*targets*." We use the OpenAI API to feed a prompt to get different LMs to play the *spymaster* and provide the clue word. We instruct the LM to give a clue for only 1 or 2 words, and we change the hint number to "2" in case the LM gives a higher number.[1] The model then creates the best possible clue for the *guesser* to use to correctly guess the *target* words. This process can be used to generate effective clues for Codename games in a systematic and consistent manner.

We then use a different prompt (see Appendix 2 for an example) to show the 15 initial words and the clue to the *guesser* model. The *guesser* must then use logical deduction and the provided clue to figure out which of the words are the *target* words. The *guesser* is thus instructed to provide a list of its guesses. This allows for an efficient and automated way for the *guesser* to have their turn in the game.

In addition to fully automated self-play, we also allow a human player to take the role of either *spymaster* or *guesser* to play with different LLM available via the OpenAI API.

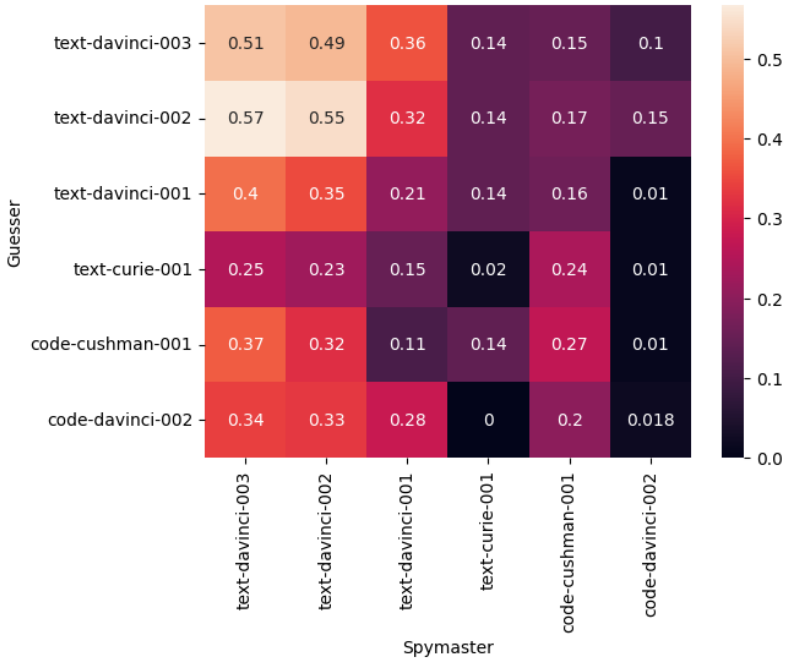
We evaluate different LLM-LLM and LLM-human team-ups according to the accuracy of the *guesser* to guess *target* words and according to the average number of words correctly guessed per turn.

We first evaluate every possible combination of the following models: text-davinci-003 text-davinci-002, text-davinci-001, text-curie-001, code-cushman-001, code-davinci-002, leading to 36 different possible team-ups. We have each team-up play 100 different games. We use a temperature setting of 0 in all experiments.

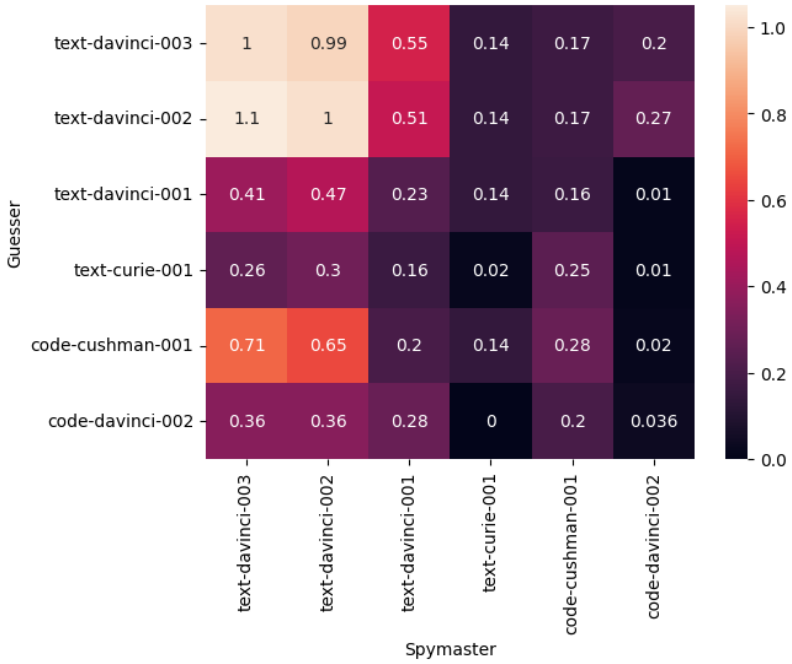
[1] We resort to this intervention after observing that the *spymaster* model tends to provide much higher numbers than warranted, even after extensive prompt engineering to mitigate this behavior.

3. Results

3.1. Self-play



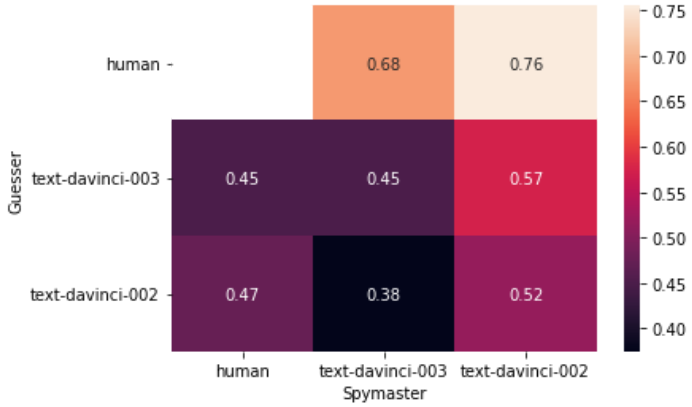
Accuracy on a run of 100 games each between human and LLMs (some of the code model numbers may only include 50 runs)



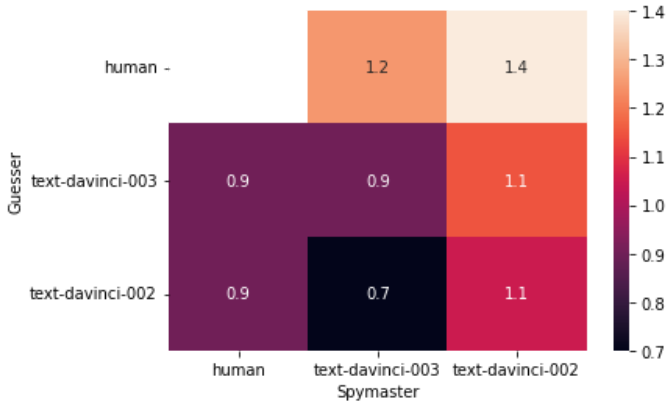
Average correct guesses on a run of 20 games each between human and LLMs (some of the code model numbers may only include 50 runs)

We find that text-davinci-002 and text-davinci-003 in particular perform very well when playing together.

3.2 Cooperative play with humans



Accuracy on a run of 20 games each between human and LLMs.



Average correct guesses on a run of 20 games each between human and LLMs.

We find that a human (average Codenames player, tired and under time pressure) is able to make good use of the hints given by the LLMs. We also find that the human in question does not outperform text-davinci-002 and text-davinci-003 in the task of *spymaster*.

4. Discussion and Conclusion

1. Quick summary of results

We find that

The need for the *spymaster* to be skillful at modeling how their clue word may be understood by the *guesser* also provides some additional evidence in the direction of

recent findings[2] that models like text-davinci-002 and text-davinci-003 are developing Theory of Mind-like abilities.

Furthermore we propose a new class of benchmarks for LLM, using cooperative, asymmetric-information language games. Such more dynamic benchmarks allow us to easily evaluate LLM on their ability to cooperate with other LLMs and humans, and on their ability to understand and propensity to use concepts and abstractions in a similar way to humans.

In addition to this, language games such as codenames allow us to evaluate models across a wider spectrum of ability when compared to traditional static benchmarks such as the ones present in BIG-bench[3]. The difficulty of the benchmark can be changed by simply changing a few parameters such as the number and difficulty of words used in the initial vocabulary, the number of words randomly selected, the initial prompts used… This reduces the risk of a benchmark quickly becoming obsolete.

We show a promising proof of concept of such a benchmark, and argue that the capability and safety of models may be better tested using such benchmarks rather than the traditional static Q/A style ones. Future research may look into collecting more human data for this benchmark, improving the prompts used, try to use a more challenging version of the game or try to come up with new cooperative language game benchmarks entirely.

Another potentially promising avenue for future research is the use of self-play and RL to improve LLMs performance on games such as codenames. Indeed, recent work has shown that such training may allow LLMs to self-improve.[4] In particular, it would be interesting to investigate if such a training setup on this language game would also lead to better play with human players, or if LLMs end up increasingly relying on alien conceptual relationships and references that a human player might not understand. This may provide some early evidence that sufficiently powerful AIs might begin to use internal representations that are difficult for humans to interpret.

Code for all experiments can be found here: https://colab.research.google.com/drive/1QF1GZM3PEDuPF5Hi4NdqEJFJmGCo9HHw?usp=sharing

5. References

[1] https://raw.githubusercontent.com/Gullesnuffs/Codenames/master/wordlist-eng.txt

[2] https://arxiv.org/abs/2302.02083

[3] https://github.com/google/BIG-bench

[4] https://arxiv.org/abs/2210.11610

Appendix

1. **Example Spymaster prompt:**

```
> You are a  spymaster in a game of codenames.
The game works as follows:
1. 15 random words are selected and are shown to everyone.
2. The spymaster is given which 5 of the words are the "targets".
3. The spymaster then has to give their teammates (the "guessers"), a single word clue and a number (either "1" or "2") to indicate the number of targets the word relates to. The clue should not be related to the other words.
4. The guessers must guess which of the words are targets.

To be a successful spymaster, it's important to think carefully about the clue word and the number of target words that the clue word is describing.
When coming up with a clue word, it can be helpful to think of synonyms, rhymes, associations and more creative uses of language.
For example, if the target words included "PLANE" and "KITE" the spymaster could use the clue word "FLY" and the number 2 ("FLY 2").
Finally, the spymaster should consider how well the clue word aligns with the target words. If the teammate guesses one of the target words incorrectly, it can lead to lost points. Therefore, giving a good clue for only 1 or 2 words is better than giving a more complex one for more words.

The spymaster's hint will be formatted like in the examples that follow:

```
Thus, the words to give a hint for are: HAWK HOTEL LAWYER POOL SPIDER

Now the Spymaster gives a clue word followed by either 1 or 2.

> Spymaster: SWIM 1
```

In this game the words are the following:
CAR ##
CLIFF
FALL
```

GIANT
HOTEL
JAM
MOUTH ##
NUT
OCTOPUS
PORT ##
PUMPKIN
RAY
STRING ##
TABLE ##
WHALE

Thus, the words to give a hint for are: CAR MOUTH PORT STRING TABLE

Now the Spymaster gives a clue word followed by either 1 or 2.

> Spymaster:
```

2. **Example Guesser prompt:**

```
> You are a highly proficient guesser in a game of Codenames.
The game works as follows:
1. 15 random words are selected and are shown to everyone.
2. The spymaster is given which 5 of the words are the "targets".
3. The spymaster then has to give their teammates (the "guessers"), a single word clue
    and a number to indicate the number of targets the word relates to. The clue
    should not be related to the other words.
4. The guessers must guess which of the words are targets.

To be a successful guesser in this game, one must listen closely to the spymaster's clues
    and use logical deduction to narrow down the potential target words.

The guesser should finish its reponse like this:
```
And the hint given to you by the spymaster is MASSIVE 2

> Guesser: ELEPHANT JUPITER
```

In this game the words are the following:

AZTEC
BUFFALO
CROSS
CROWN
DWARF
FORCE
HONEY
HORN
ICE CREAM
LINE
OCTOPUS
PORT
SHAKESPEARE
SPELL
TORCH

And the hint given to you by the spymaster is "FLY 2".

> Guesser:
```
```