# Sub-word Spotting: Uncovering Embedded Objects

Natural Language Processing Course – Final project

Daniel Bazar 314708181

Lior Krengel 315850594

## Abstract

Necessary?

while LLM’s are struggling with it, for humans it almost trivial.

## Introduction

Necessary?

## Linguistic task investigated

*A description and a definition of the linguistic phenomena or task –*

We demonstrate a type of word puzzle that involves identifying embedded words within other words. Specifically, this puzzle asks to find a subword that is also the name of an object from some category (for example, colors, animals, etc.).

question for example:

*which one of the following words contains a subword of an animal?*

1. *guitar*
2. ***million (contains "lion")***
3. *pasta*
4. *house*

We expect the LLM to succeed in distinguishing words as subwords. while standard LLM doesn’t trained on such tasks this task demands the ability of pattern recognition which we assumed they have. In contrast, we assume that this task is more challenging when there's no clear semantic or structural relationship between the subword and the surrounding characters. LLMs rely on patterns and associations in language to understand words and their meanings, If there's no meaningful connection, the model may struggle to detect the relevant subword (the pattern).

In the example above, "million" containing the subword "lion," the challenge lies in the fact that the remaining characters ("mil") don't carry any obvious semantic or structural significance in relation to "lion."

## Evaluation methodology

### Dataset

*explain the format of the dataset, as well as how you created it*

Our aim in generating the dataset was an automatic approach that construct pairs of word from a category (the subword) and a word containing the subword. our main guideline was to choose commonly used subwords so it would be easier to recognize them, especially for non-native English speakers.   
To achieve this, we first collected several word lists of different categories such as animals, body parts, colors, food, fruits, and vehicles. We ensured manually that the words are simple and known. Then we retrieved the “brown corpus” and used its vocabulary to fit subwords into words. We removed stopwords and lemmatized all the words to prevent unwanted and duplicates words. And after that, we remained with pairs of word and subword. This dataset contains thousands of records and for our task, we interested in a much smaller one and high quality one. To dilute it further we used a pretrained word2vec model in order to avoid similar pairs of word and subword. for example, burger and hamburger, boat and sailboat, brown and brownish, etc. This method also has another advantage that by choosing a pair with smaller similarity we force (hopefully) the LLM to rely less on the semantic meaning of the words and more on pattern recognition, in this way, we can evaluate the performance better.

Some of the words appeared more frequent as subwords than others (zipf’s law is everywhere!) *[FIGURE]*. so, to construct diverse dataset we decided to take each word once. Then we attached to each subword, the most frequent word according to the brown corpus that contains it. The last step in the process was a manual intervention, we passed through the dataset and validated the results.

The distillate dataset contains 60 records. Each record is of the following format: {subword, category, word, 3 random out-of-category words for multi-choice questions}

Then we would use this dataset to create different types of prompting as we will describe in *[SECTION]*. We left aside some more records for few-shot prompting.

### Metrics

Explain the metrics you used to quantify your work-

Accuracy.

Human benchmark.

How to extract answers.

### Experiments

Explain all the experiments you conducted.

Prompting technique, models, human benchmark

## Results

Discussion of the experiments and the results

Provide both quantitative (based on measurements) as well as qualitative (based on manual analysis of few examples) evaluations

## Summary\Conclusion

A summary of your findings and a conclusion.

## Mitigation

Propose ways in which the LLM limitation you uncovered can be remedied. Note that you do not need to implement it in this project, but you need to describe it clearly and assess its feasibility. Optionally you can show a proof of concept on a small scale.

Double check

string comprehension

retrieve data from web

## Appendices

all the prompts used in your reported experiments.