The Robustness of In-context Learning to Word Shuffling

Zhouxiang Fang, Haonan Wang, Huiqi Zou

Introduction

1. Background

- LLMs demonstrates impressive performance with In-context learning (ICL), but the nature of ICL remains opaque.
- ICL is very sensitive (e.g. demonstration order, word framing …).

2. Goal: Investigate the robustness of ICL to word shuffling.

3. Hypothesis: If ICL is a real learning mechanism, LLMs should be able to recover the input in the original vocabulary since the mapping is reversible.

Methodology

- **Word Shuffling**
  - Randomly shuffle the original vocabulary, while maintaining bijection between shuffled and original vocabulary.
  - Perturb the demonstration input by mapping it to the shuffled vocabulary.

Shuffling Rate	Sentence
0.0	it 's a charming and often affecting journey .
0.2	it 's a heart polyonymal often affecting journey .
0.5	it 's cromfordite paternalistic and preadvance antimedical triantaphyllos .
0.8	it 's a valid and quadrilaterally billyboy palation .
1.0	excipulum 's punnic sultrily permeance clockwise protodonate myopathia .

Table: Sentence from SST-2 shuffled by different rates.

- **ICL Strategy**
  - 50-shots with Leave-one-out sampling
  - Prompt format:

    Input: {input_text}
    Output: {label}
    …
    Input: {question}
    Output: {prediction}

    } 50 examples.
- **Models**
  - Closed-source LLMs: GPT-3.5-turbo.
  - Open-source LLMs: Llama3-70B and Mistral-7B (a pre-retrained model without instruction-tuning or RLHF).

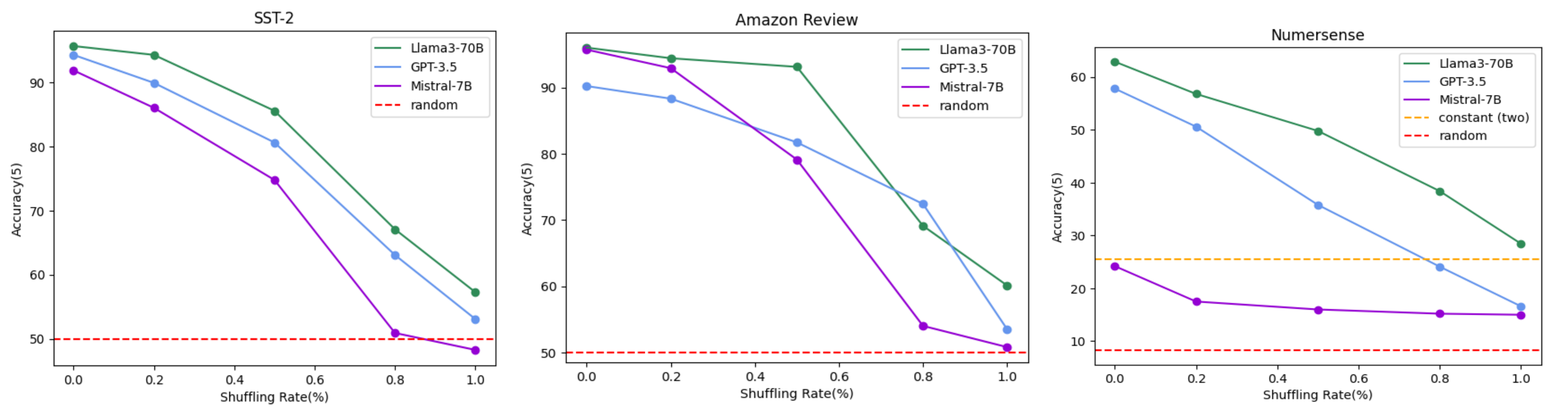
Datasets

- **Sentiment Classification:** `SST-2 and Amazon`
- **Masked LM:** `NumerSense`

Dataset	Task	Input	Ouput/Label
SST-2	sentiment classification	funny yet	1 (positive)
Amazon	sentiment classification	cheap! pulls hairs out	0 (negative)
Numersense	common sense	Dragonflies have <mask> wings.	nine

Table 1: Examples Input and Output/Label of three datasets. For SST-2 and Amazon, the label is either 1 (positive) or 0 (negative). For Numersense, the task is to predict the masked token.

Results



Finding 1: As shuffling rate increases, performance strictly decreases.

Finding 2: GPT-3.5 and Mistral-7B drop to random guess when vocab is completely shuffled.

Finding 3: Llama3-70B maintains robustness to complete word shuffling.

Dataset	Word			Average
	Good	Great	Bad	
SST-2:				
Frequency	3.21	1.37	2.29	-
Accuracy	75.0	75.0	65.0	58.14
Amazon:				
Frequency	8.3	10.7	2.1	-
Accuracy	61.44	62.62	61.90	60.7

Finding 4: Frequent words help prediction.

w/ original input	Few-shot	Accuracy
Yes	50	**65.2**
Yes	10	62.7
No	50	60.1
No	10	59.3

Finding 5: LLMs learn mapping via ICL.

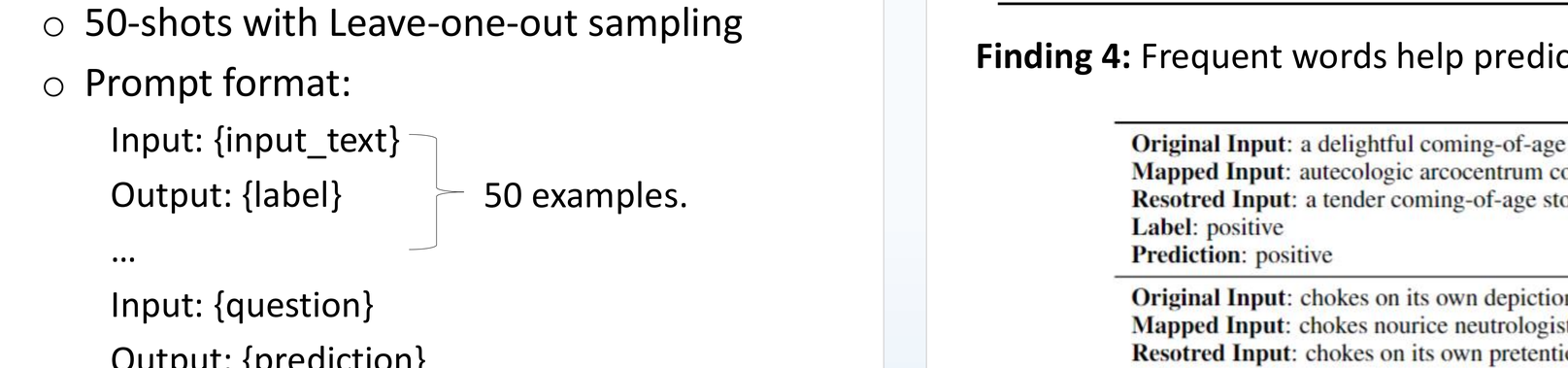
Original Input: a delightful coming-of-age story .
Mapped Input: autecologic arcocentrum coming-of-age pelargonic .
Resotred Input: a tender coming-of-age story .
Label: positive
Prediction: positive

Original Input: chokes on its own depiction of upper-crust decorum .
Mapped Input: chokes nourice neutrologistic circumagitation noup holostylic upper-crust pushful .
Resotred Input: chokes on its own pretentiousness and upper-crust aspirations .
Label: negative
Prediction: negative

Original Input: this movie seems to have been written using mad-libs .
Mapped Input: nondispersal nonloving seems silesia overpassionate sagai starlit using mad-libs .
Resotred Input: this movie seems like it was made using mad-libs .
Label: negative
Prediction: negative

Finding 6: LLMs can restore the original input.

Pipeline on Sentiment Classification Task



Future Works

- Include more tasks and datasets.
- Evaluate more models with different pretraining, instruction-tuning, and RLHF approaches.
- Investigate the impact of few-shot number.
- Develop strategies to handle out-of-vocabulary words.
- Conduct more in-depth analysis (e.g., attention-based method).