PseudoMC: Measuring How Language Model Understand Words by Pseudo-words

Yukai Yang
New York University
yy2949@nyu.edu

Yixin Zhu
New York University
yz5880@nyu.edu

Yuchen Zhu
New York University
yz4975@nyu.edu

Abstract

We propose a benchmark to measure and understand how large language models learn a new word. The benchmark contains 500 English words, each word with 3 context sentences, and three other words as distractors. We mask the word in the sentences by random sampling, adding or removing prefix and suffix to test how well models can learn this newly generated word. Our results show that the morphological structure of a word can significantly impact models' learning, even more than the contextual sentences. This gives a new idea of interpreting how LLMs work.

1 Introduction

It is known that large language models (LLMs) have a dominating performance on few-shot tasks and the score scales up with the model size (Brown et al., 2020). However, as we scale up the model size, it becomes harder for human intelligence to understand how and why it works.

To better understand what LLMs learn, numerous evaluations have been conducted to test the model in various aspects, such as linguistic knowledge (Ettinger, 2020), translation (Brown et al., 2020), natural language generation (Chen et al., 2019), etc. To find how machine learns new words in a human-like manner, we mainly test two abilities of LLM: (1) learning new words from its structure i.e. morpheme, (2) learning new words from the context.

While large-scale pretrained models are famous for their capability for such tasks (Brown et al., 2020), we tried on GPT-3 a fake word "conducing" replacing "confusing" with context to see if it can learn the meaning of target word from context or is it mislead by the fake word. The example query and model answer is:

[Query] *Question: Which of the following explains the word conducing in the sentence better, confusing, unhappy, clear, or angry? The instructions were not just conducing, they were positively misleading.*

[Answer from GPT-3] *The word conducing in this sentence means "clear."*

In this project, we introduce PseudoMC, a benchmark testing how models learn a new word from context, and test models, specifically for BERT and GPT family, on our benchmark in zero-shot and few-shot manner respectively. The results of such experiments, therefore, can serve as an insight of understanding how LLMs learn natural language.

2 Related Work

2.1 Machine Comprehension Tasks

MCTest (machine comprehension of text) is one of the earliest datasets to test machine comprehension of texts (Richardson et al., 2013). MCTest contains a passage and several reading questions to test if the model can extract information from the passage. Other popular reading comprehension datasets include SQuAD (Stanford Question Answering Dataset) with question-answer pairs sourced from Wikipedia pages and RACE (ReAding Comprehension Dataset From Examinations) that improves on the base of MCTest both in terms of both size and difficulty etc (Rajpurkar et al., 2016; Lai et al., 2017).

2.2 Word Prediction Tasks

Word prediction datasets are one form for testing machine understanding of language.

CNNDM (CNN/Daily Mail) dataset tests specifically on models' understanding of named entities by masking named entities in news articles and asking the model to predict the named entity in the summary of the article (Hermann et al., 2015). However its data is restricted to source from news articles and words of named entities. CBT

(Children's Book Test) dataset masks four types of words of different part of speech in the last sentence in a 21-sentence context and provides tested models with candidate choices of synonyms and random words (Hill et al., 2015).

The LAMBADA (LAnguage Model Broadened to Account for Discourse Aspects) dataset masks the last word in a sentence with a given context so that tested models need to understand the context to get the correct answer (Paperno et al., 2016). The ChineseWPLC (Chinese Word Prediction from Long Context) dataset is designed specifically for large language models that tests the model on predicting target word from a Chinese novel context (Ge et al., 2021). The two datasets leverage word prediction to evaluate models' capability to keep track of long-range context so that models would not score if they make predictions only based on the local sentence.

2.3 Word Understanding Tasks

Our word prediction dataset directly tests models on their understanding of words. A related benchmark is WDLMPro (Word Definition Language Model Probing) dataset that replaces the original words and tests the model on matching definition to word and matching word to definition (Senel and Schütze, 2021). Differences between WDLMPro and our dataset include that we ask the model to match the target word to context, which, comparing with WDLMPro matching words to definitions, contains less contextual information than definitions and is closer to real world tasks when the model needs to understand words from a given prompt.

Moreover, inspired by the TruthfulQA dataset that tests models to avoid producing false information (Lin et al., 2021), we mask the target word in misleading ways which, including inserting random typos and misleading prefix and suffix.

3 Methodology

3.1 Dataset

Our dataset is a benchmark designed to measure how well large language models can understand an unseen new word from several sentences containing rich context information for this target word.

The dataset contains 500 target words. For each ground truth word g, there are 3 independent sentences to form a context (denoted S) that fits different meanings of g, a mask word (denoted w), and 4 candidate answers (denoted C). We replace all appearances of the ground truth word g in context S with the mask word w to form the test context $\hat{S}$. The set of candidate answers C contains g and three misleading choices. The three misleading choices include one synonym and one antonym of g, and a purely random word. The model needs to identify the best answer (i.e., $\hat{g}$) that has the closest meaning to the mask word w among the candidate choices by inferring from the test context $\hat{S}$. Hence a question-answer pair in the dataset is a four-element tuple: $(\hat{S}, w, C, g)$. See further details in 3.2.

The source for g includes GRE vocabulary list for 150 words and TOEFL list for 350 words. The source for context S are a sentences pool built by scraping from the website `YourDictionary.com`. The authors invite friends as volunteers to manually select informative, well-written sentences from the pool to build a context S that well indicates the meaning of word g.

3.2 Experiment Setup

Model. In this project, we evaluate the performance of LLM on understanding new words from contexts through testing on the BERT family and GPT family. For the BERT family, we experiment with 1) BERT-base (Devlin et al., 2018) (referred to as BERT), 2) RoBERTa-base (Liu et al., 2019) (referred to as RoBERTa), and 3) distilbert-base-uncased (Sanh et al., 2019) (referred to as distilbert). For the GPT family, we experiment with 1) GPT-3 (175B) (Brown et al., 2020), 2) GPT-2-small (referred to as GPT-2) (Radford et al., 2019).

Prompt. Our task includes both zero-shot learning, i.e., there's no training or gradient updates happened during the evaluation phase and the tested model is not exposed to any examples with answers from our dataset a-priori, and few-shot learning. The zero-shot learning contains a natural language instruction (asking the model to choose the best answer from given choices), and the few-shot learning just adds 4 similar instructions with answers ahead as example.

The prompt we adopted for experiments is a standard multiple choice prompt widely used in the field, with a slight modification so that the output matches our desired format. The prompt includes 3 sentences containing the "unseen" word that provide a relatively rich context for the understanding of this word. The prompt also contains four choices and an instruction that guides the model to respond with A,B,C,D as answers, indicating the best choices it predicts. The context sentences and misleading choices are carefully chosen so that context sentences cover different meanings of the target word so that other three misleading candidates cannot fit in all three contexts. There is also carefully manipulated nuances between the target word and misleading candidates so that even misleading choices fit in the context, the target word makes the best candidate for the context. The following is a default prompt format for model evaluation:

`Sentence A , Sentence B , Sentence C.`
Which of the following explains the word {`Mask word`} in the sentences better?
A.{`Choice A`}, B.{`Choice B`}, C.{`Choice C`}, D.{`Choice D`}
Provide the answer in A, B, C, or D.

3.3 Masking strategy

Random typo. {Typo} A major source of "new word" in daily life are misspelled words. In this task, we simulate such type of "unseen word" by randomly corrupting the ground truth word g with a uniform noise. Examples of such corruption include "catalog" to "caataogg" and "facade" to "fccadee".

Randomly sampled letters. {Random} The target word is masked by a pseudo-word of the same number of letters. Each letter is uniformly sampled from the alphabet to sequentially form a mask word. Examples include "catalog" to "mynbiqp" and "facade" to "hjdmpe".

Prefix+Suffix. {Random_PS} We also try to adopt misleading information in random masking by prefix+suffix. The result mask word appears to be a real word, but has no meaning in the lexicon. We generate such mask words by randomly generating words with structure "prefix+suffix" or "prefix+embedding+suffix". Note that the generated pseduo-words in example does not need to have correlation to the ground truth word, but they may serve the same part-of-speech role. Examples for such words are, "catalog" to "motaarm", "facade" to "fusience".

Prefix/suffix replacement. {Replace_PS} The target word is masked by replacing its morpheme with prefix, suffix, and/or root. First we decompose the target word into a set of its morpheme parts (denoted M). If the set cardinality $|M| \leq 1$, the word will be replaced by a randomly sampled root; if $|M| = 2$, the first morpheme of the word is replaced by a randomly sampled prefix; and if $|M| \geq 3$, the first morpheme part will be replaced by a randomly sampled prefix and the last will be replaced by a randomly sampled suffix. Examples include "catalog" to "untaarm" and "facade" to "copulcade".

4 Results and Analysis

Of the models selected, BERT family models BERT, RoBERTa, and distilbert are fine-tuned before evaluation since BERT family cannot well adapt to the downstream task we provide. GPT-2 and GPT-3 are evaluated without fine-tuning. All the models except GPT-2 are experimented with zero-shot learning and few-shot learning to help them better perform on our specifically designed task. GPT-2 is only evaluated with few-shot learning since it cannot well understand the task in few-shot learning, generating random words instead of multiple choice answer. Moreover, we find GPT-2 is not good enough to understand the task even with examples and constantly produces unstable answers. Hence we let GPT-2 answer 10 times and take the most frequent choice as its answer for each target word question. However, GPT-2 still performs zero accuracy on all tasks. Hence we do not consider GPT-2 for model performance comparison.

The human baseline result is from several human annotator volunteers. Each annotator gets a randomly sampled set of 50 questions and the result is averaged. Annotators only work on questions with Random_PS masking, since Typo masking would be too evident for humans with only few letters replaced and the other three masking are not of big difference as humans can recognize the mask is a fake word. Thus we pick the Random_PS mask as it the most resembles the structure of a real word. We get an average human accuracy of 0.879 on the task.

The results of models on tasks of 4 masking strategies are presented in the table below. Our re-

Model	Typo	Random	Random_PS	Replace_PS
DistilBERT (ZL)	0.328	0.275	0.305	0.295
DistilBERT (FL)	0.356	0.279	0.297	0.293
RoBERTa (ZL)	0.422	0.330	0.305	0.301
RoBERTa (FL)	0.444	0.309	0.311	0.311
BERT (ZL)	0.504	0.485	0.479	0.471
BERT (FL)	0.461	0.442	0.434	0.444
GPT-2 small (FL)	0.000	0.000	0.000	0.000
GPT-3 (ZL)	**0.959**	**0.767**	0.708	0.772
GPT-3 (FL)	0.955	0.751	**0.804**	**0.793**

Table 1: The four columns on the right show experiment results of different versions each model (with zero-shot learning (ZL) and with few-shot learning (FL)) under respective masking strategy

sults reveal that models' ability to learn a word from context is loosely associated with model size. Model performance does not differ much when model does not have a large difference in the number of parameters. BERT-base after fine-tuning achieves higher accuracy on all tasks than RoBERTa after fine-tuning, while BERT-base is 10.6% smaller than RoBERTa-base with 123 million parameters. For distilbert-base with 66M parameters that is about one-third smaller than BERT and RoBERTa, its accuracy decreases from the previous two by roughly 0.1. On the other hand, when the model is large enough, like GPT-3 that is more than 1000 times larger than models in the BERT family, it presents an absolute advantage over all the smaller models on the given tasks. Zero-shot and few-shot learning also has different effects on model performance, as discussed below.

4.1 Typo

Of the four tasks, Typo task is the simplest with highest model performance. The most likely reason is Typo mask word highly resembles the original word with few letters replaced, and hence the model can well infer the original word from remaining letters. GPT-3 with zero-shot learning achieves the highest accuracy of 0.959, which is of the same magnitude of accuracy as humans. The smaller models BERT, RoBERTa, and distilbert achieves accuracy of below 0.5, no more better than random guess.

4.2 Random

Random task is the second simplest task. GPT-3 with zero shot learning achieves the highest accuracy of 0.767. However for the BERT family, even random masking with no misleading information well confounds the model, resulting in performance of BERT at 0.479, RoBERTa at 0.311, and distilbert at 0.279.

4.3 PS tasks: Random_PS & Replace_PS

For Random_PS task, GPT-3 with few-shot learning achieves the highest accuracy of 0.804 and for Replace_PS task, GPT-3 with few-shot learning achieves a similar accuracy of 0.793. Few-shot learning largely boosts model performance on PS tasks for GPT-3. Few-shot learning also boosts performance of RoBERTa at an accuracy of 0.479 for Random_PS and of distilbert at an accuracy of 0.311 for both PS tasks, but BERT performs better with zero-shot learning at an accuracy of 0.444 for Replace_PS.

5 Conclusion and Future Work

We build a benchmark of multiple choice questions to evaluate various masking strategies. We exclusively test LLMs with our benchmark. The results indicates that the performance gets significantly better as model size increases. Meanwhile, changes in morphological structure, in general, shows a stronger impact to model performance than changes of pure randomness. This serves as an evidence that models learn a new word not only based on the context, but also the structure of the words.

In the future, we can further apply our benchmark to different LLMs. Developing additional masking algorithms that focus on morphological structure is also a potential research direction.

6 Ethical Considerations

The ability to understand new and unseen words is crucial for both large language models and human beings. Also, understanding what LLMs have learnt from enormous texts and corpus is also becoming a concern. On the bright side, the PseudoMC dataset introduced in this paper is a new benchmark that enables us to gain insights into the internal logic behind LLMs' learning process. On the other side, it might also bright about additional risks if manipulated by malicious users. The masking strategies used in the dataset generation protocol might become new source of adversial attacks for LLMs. With simple and near-negligible corruptions of a word, LLMs will suffer a detrimental loss in their capability to understand correctly the meaning of this word, while human beings might not even notice the changes. Purposefully feeding LLMs with such texts could potentially create issues. Similarly, adversarially generated pseduo-words with prefix/suffix containing negative meaning might also trick LLMs into erroneous behaviors.

In a world where languages are evolving at an incredible space, each years there are hundreds of new words written into dictionaries, let alone the large number of prevailing expressions in the social media. Many of those new words don't possess a structure like most of the existing words, and look much more like a pseudo word generated in our protocol. For example, "BIPOC" stands for ”Black, Indigenous, (and) People of Color”, "BLM" stands for "Black Lives Matter". These acronyms are officially recognized as words, but their similarity to randomly generated words in our dataset might create unintended consequences if we fine-tune models on PseduoMC to enhance recognition ability of unseen words. While the newest words are often associated with the latest social affairs and international events, unexpected bias is likely to be created in this process.

7 Acknowledgements

We thank volunteers Yang Chen, Jinhan Luo, and Sichen Zhu et al. for their efforts in generating dataset and creating human benchmark results. We thank Yilun Kuang for his assistance in resolving technical issues. We also want to acknowledge kind support by the NYU High-Performance Computing Center.

8 Collaboration Statement

Yukai Yang

Designing collaboratively the dataset and experiments;

Generation of part of the dataset;

Implementation of the masking algorithms;

Implementation and Experiment for the GPT family;

Write-up of Abstract, Introduction, and Conclusion.

Yixin Zhu

Designing collaboratively the dataset and experiments;

Implementation of dataset generation and GPT-2 experiment;

Generation of part of the dataset;

Write-up of Literature Review, Methodology, and Results and analysis.

Yuchen Zhu

Designing collaboratively the dataset and experiments;

Generation of part of the dataset;

Implementation, fine-tuning and Experiment for BERT, RoBERTa, and DistilBERT;

Write-up of Methodology.

9 Github

All code and data files for the project can be found in the following link:
https://github.com/yk803/DS203_Project

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2019. Few-shot nlg with pre-trained language model. *arXiv preprint arXiv:1904.09521*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Huibin Ge, Chenxi Sun, Deyi Xiong, and Qun Liu. 2021. Chinese WPLC: A Chinese dataset for evaluating pretrained language models on word prediction given long-range context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3770–3778, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Lutfi Kerem Senel and Hinrich Schütze. 2021. Does he wink or does he nod? a challenging benchmark for evaluating word understanding of language models. *arXiv preprint arXiv:2102.03596*.