Revisiting Psycholinguistic Diagnostics with Large Language Models

Emrecan Çelik
Boğaziçi University
Department of Cognitive Science
34342 Bebek, Istanbul, Turkey
emrecan.celik@std.bogazici.edu.tr

Abstract

This preliminary study extends previous work on understanding of negation and semantic roles of small language models such as BERT, ALBERT and others to large language models (LLMs). We enrich previous observations on psycholinguistic diagnostics and their extended versions with 4 different LLMs and their variations, which results in 12 models in total. We also replicate the previous results with BERT and ALBERT models and variations (7 in total), and point out the possible mistakes or implementation choices in the evaluation process followed by others. Notwithstanding the vast success of LLMs, our preliminary observations show that LLMs still fail to *understand* negation. Our implementation of the experiments are available for open-access[1].

1 Introduction

Diagnostics drawn from psycholinguistic studies, when utilized on language models, have expanded our knowledge of what language models understand, and what they do not. All diagnostics proposed by (Fischler et al., 1983), (Battig and Montague, 1969) and then utilized on language models by (Ettinger, 2020) follow a simple but carefully designed approach: (1) curate a dataset targeted for a specific knowledge or skill (eg. negation, commonsense inference) that consists of incomplete sentences with one missing word, (2) ask participants/models to complete the sentence, (3) record observations such as EEG data, human cloze probability, or model probabilities, (4) then finally define a measure for targeted knowledge or ability.

Since masked token prediction or sentence completion are rather *natural* tasks for language models, (Ettinger, 2020) tested these diagnostics drawn from human experiments on language models, showing that BERT completely fails to understand negation in a simple hypernym completion task shown in Table 4. Considering these datasets come from human experiments, they are very small compared to the usual benchmark datasets used for language models. (Shivagunde et al., 2023) addresses this problem by extending these datasets by using in-context learning with GPT-3, and using various categories and templates proposed in previous studies (Fischler et al., 1983; Battig and Montague, 1969) to generate new examples, ultimately making the datasets, thus the conclusions, more reliable there are a few possible problems with their approach. (1) They generate the dataset NEG-SIMP-1500-GEN using GPT-3 and evaluate the model on the data it has generated. (2) Sensitivity metric used in the experiments does not match the metric used by (Ettinger, 2020).

Context (+)	**Context (-)**	**Target (+)**	**Target (-)**
A trout is (a/an)	A trout is not (a/an)	fish	tool
A sparrow is (a/an)	A sparrow is not (a/an)	bird	vehicle

Table 1: Examples from NEG-136 dataset.

[1] https://github.com/emrecncelik/llm-diagnostics

In order to address these issues and considering the recent advancements in language modeling we evaluate 4 LLMs, their variants. We also evaluate BERT and ALBERT to replicate the results from previous studies. We contribute to the literature by attempting to answer the following questions: Does the problem of negation understanding persist in LLMs? How do the LLMs in different sizes and architectures differ in their abilities tested by these diagnostics? The rest of this paper is organized as follows: Section 2 shows our Experimental Setup alongside with models and datasets we use, results of the experiments are in Section 3, and finally we make conclusions and discuss our results and future work in Section 4.

2 Experimental Setup

All experiments follow the methodology from (Ettinger, 2020; Shivagunde et al., 2023) with small modifications. Although we follow the same methodology as previous studies, some results we observe are different, which will be discussed in Experimental Results section. We formulate the problem as masked language modeling for encoder type models and causal language modeling for others. For MLM, we replace the target word with the corresponding mask token of the tokenizer of the model and insert a dot afterwards to denote end of the sentence. For CLM, we simply remove the target and predict the next token. Previous studies use different definitions of sensitivity which do not align. (Ettinger, 2020) defines sensitivity as “the proportion of items in which the model assigns higher probabilities to true completions than to false ones”, and (Shivagunde et al., 2023) defines it as “percentage of sentence pairs for which the top-1 prediction changed”. We use both for all our evaluations and show sensitivity results from (Shivagunde et al., 2023) might be misleading in the results section.

2.1 Models

Models we use for the experiments are selected from Open LLM leaderboard[2] and latest surveys on LLMs (Zhao et al., 2023). We use three main architectures for LLMs which are as follows: transformer encoder, transformer encoder with mixture of experts, and state-space models. We also use architectures based on transformer encoder, which are ALBERT and BERT variants. We apply 4 bit quantization to all LLMs in order to be able to fit them into the VRAM of Colab Pro+. This choice might have detrimental effects on our results and will be addressed in future work. For ALBERT we chose only to use version 2.

2.2 Datasets

We use the datasets proposed by (Ettinger, 2020) and extended by (Shivagunde et al., 2023) and provide short descriptions of each in this section. They all can be found on Github[3][4] or HuggingFace datasets.

2.2.1 ROLE-88

Targets the model’s capability to understand event knowledge and semantic role interpretation, following Table 3 shows example pairs from the dataset. One pair is excluded from the evaluations since it consists of multiple tokens (item 118).

2.2.2 NEG-SIMP-136

This datasets targets the understanding of negation using category membership. Previous models such as BERT was shown to have a poor performance in this task. Examples from the dataset can be seen in the Table 4 below,

2.2.3 NEG-NAT-136

Similarly targets the understanding of negation, but uses more natural sentences that could be a part of daily conversations.

[2] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

[3] https://github.com/aetting/lm-diagnostics/tree/master

[4] https://github.com/text-machine-lab/extending_psycholinguistic_dataset/tree/main

Architecture Type	Model ID	Checkpoint
Encoder	ALBERT$_{base}$	albert/albert-base-v2
	ALBERT$_{large}$	albert/albert-large-v2
	ALBERT$_{xlarge}$	albert/albert-xlarge-v2
	ALBERT$_{xxlarge}$	albert/albert-xxlarge-v2
	DistilBERT	distilbert-base-uncased
	BERT$_{base}$	bert-base-uncased
	BERT$_{large}$	bert-large-uncased
State-space	MAMBA 130M	state-spaces/mamba-130m-hf
	MAMBA 790M	state-spaces/mamba-790m-hf
	MAMBA 2.8B	state-spaces/mamba-2.8b-hf
Decoder	LLaMA-2 7B	meta-llama/Llama-2-7b-hf
	LLaMA-2 13B	meta-llama/Llama-2-13b-hf
	LLaMA-3 8B	meta-llama/Meta-Llama-3-8B
	Mistral 7B	mistralai/Mistral-7B-v0.1
	Qwen1.5 1.8B	Qwen/Qwen1.5-1.8B
	Qwen1.5 7B	Qwen/Qwen1.5-7B
	Qwen1.5 14B	Qwen/Qwen1.5-14B
Decoder MoE	Mixtral-8x 7B	mistralai/Mixtral-8x7B-v0.1

Table 2: List of models used for the experiments.

Context	Expected
the firefighter reported which paramedic the victim had	thanked
the firefighter reported which victim the paramedic had	saved

Table 3: Examples from ROLE-88 dataset.

2.2.4 ROLE-1500

Same as ROLE-88 dataset, except it has 1500 examples instead of 88. This dataset is extended with a rule based template method using the categories provided in (Fischler et al., 1983) and the template from the original dataset.

2.2.5 NEG-1500-SIMP-TEMP

Following the format of NEG-136 dataset, consists of 1500 generated examples instead of 136. A template is provided to the GPT3 model and it is asked to fill the masks. For example, `A canary is a [MASK]`, then GPT3 returns the token (eg. bird), and the output is filtered by human annotators.

2.2.6 NEG-1500-SIMP-GEN

Extended version of NEG-136 using few-shot prompts provided to GPT3. A task definition and a few examples are provided as the prompt, and the rest is completely generated by GPT3. Some similar examples are given in Table 4.

3 Experiment Results

We present the results on negation accuracy, negation sensitivity, role accuracy, role sensitivity in Tables 2, 8, 5 & 6 respectively. We will be discussing the results in the same order. Please note that we could not exactly replicate previous work, and probably failed to replicate it for ALBERT since models have very low accuracy compared to previous studies.

Context (+)	Context (-)	Target (+)	Target (-)
A trout is (a—an)	A trout is not (a—an)	fish	tool
A sparrow is (a—an)	A sparrow is not (a—an)	bird	vehicle

Table 4: Examples from NEG-136 dataset.

Datasets	**NEG-136-SIMP**			**NEG-136-NAT**			**NEG-1500-TEMP**			**NEG-1500-GEN**		
Models	**1**	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**
ALBERT$_{base}$	0.0	0.0	0.0	0.0	0.06	0.19	0.0	0.01	0.02	0.0	0.01	0.04
ALBERT$_{large}$	0.11	0.72	0.78	0.19	0.19	0.38	0.16	0.38	0.45	0.1	0.27	0.36
ALBERT$_{xlarge}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01
ALBERT$_{xxlarge}$	0.39	0.67	0.72	0.12	0.25	0.31	0.16	0.29	0.33	0.1	0.24	0.31
DistilBERT	0.44	**0.94**	0.94	0.12	0.31	0.38	0.31	0.55	0.64	0.19	0.36	0.44
BERT$_{base}$	0.5	0.83	0.94	0.19	0.31	0.5	0.38	**0.63**	**0.68**	0.21	0.44	0.56
BERT$_{large}$	0.67	**0.94**	**1.0**	0.25	0.31	0.31	0.38	0.6	0.66	0.25	0.49	0.59
MAMBA 130M	0.11	0.33	0.33	0.06	0.06	0.19	0.24	0.39	0.45	0.14	0.27	0.32
MAMBA 790M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01
MAMBA 2.8B	0.72	0.89	0.94	0.25	0.44	0.5	0.45	0.6	0.65	0.34	0.54	0.61
LLaMA-2 7B	0.5	0.89	0.94	0.38	0.44	0.5	0.36	0.55	0.62	0.29	0.53	0.65
LLaMA-2 13B	0.67	0.89	0.94	0.31	0.38	0.38	0.37	0.57	0.64	0.29	0.53	0.63
LLaMA-3 8B	0.56	0.89	0.94	0.25	0.38	0.38	0.35	0.58	0.65	0.31	0.53	0.62
Mistral 7B	0.5	0.83	0.89	**0.44**	0.44	0.44	0.41	0.61	0.67	0.33	**0.58**	**0.66**
Mixtral-8x 7B	0.5	0.78	0.83	0.19	0.44	0.5	0.39	0.6	0.65	0.29	0.54	0.63
Qwen1.5 1.8B	**0.83**	**0.94**	**1.0**	0.25	0.44	0.56	**0.48**	0.61	0.66	**0.37**	0.55	0.62
Qwen1.5 7B	0.67	0.89	0.89	0.25	**0.5**	0.5	0.4	0.58	0.64	0.28	0.46	0.53
Qwen1.5 14B	0.5	0.72	0.78	0.38	**0.5**	**0**.62	0.37	0.54	0.61	0.3	0.49	0.56

Table 5: Top-k accuracies on the affirmative negation task for all models. Numbers 1, 3, 5 denote the value of k.

Datasets	**ROLE-88**			**ROLE-1500**		
Models	**1**	**3**	**5**	**1**	**3**	**5**
ALBERT$_{base}$	0.05	0.16	0.16	0.03	0.06	0.08
ALBERT$_{large}$	0.09	0.21	0.26	0.07	0.15	0.19
ALBERT$_{xlarge}$	0.0	0.0	0.0	0.0	0.0	0.0
ALBERT$_{xxlarge}$	0.14	0.3	0.4	0.1	0.19	**0.25**
DistilBERT	0.12	0.21	0.26	0.09	0.19	0.22
BERT$_{base}$	0.07	0.16	0.23	0.08	0.15	0.21
BERT$_{large}$	0.09	0.21	0.3	0.1	0.18	0.23
MAMBA 130M	0.07	0.14	0.19	0.04	0.1	0.15
MAMBA 790M	0.0	0.0	0.0	0.0	0.0	0.0
MAMBA 2.8B	0.02	0.07	0.12	0.04	0.09	0.13
LLaMA-2 7B	0.16	**0.37**	0.42	0.09	0.19	**0.25**
LLaMA-2 13B	0.02	0.3	0.4	0.08	0.18	0.24
LLaMA-3 8B	**0.19**	0.26	0.37	0.1	0.17	0.24
Mistral 7B	0.12	0.4	0.47	0.09	0.18	0.24
Mixtral-8x 7B	0.02	0.26	0.4	0.08	0.18	0.22
Qwen1.5 1.8B	0.09	0.16	0.19	0.07	0.16	0.21
Qwen1.5 7B	0.16	0.28	0.3	**0.11**	0.2	0.24
Qwen1.5 14B	0.12	0.4	**0.49**	0.09	**0.21**	**0.25**

Table 7: Top-k accuracies on the original (not role-reversed) role task for all models. Numbers 1, 3, 5 denote the value of k.

Datasets	NEG-136 SIMP			NEG-136 NAT			NEG-1500 TEMP			NEG-1500 GEN		
Models	**S(+)**	**S(-)**	**S***	**S(+)**	**S(-)**	**S***	**S(+)**	**S(-)**	**S***	**S(+)**	**S(-)**	**S***
ALBERT$_{base}$	0.78	0.22	0.67	0.31	0.69	0.94	0.71	0.29	0.66	0.64	0.36	0.67
ALBERT$_{large}$	1.0	0.0	0.28	0.56	0.44	0.62	0.85	0.15	0.35	0.82	0.18	0.3
ALBERT$_{xlarge}$	0.61	0.39	1.0	0.31	0.69	1.0	0.56	0.44	0.99	0.42	0.58	0.99
ALBERT$_{xxlarge}$	1.0	0.0	0.61	0.5	0.5	0.81	0.84	0.16	0.68	0.79	0.21	0.61
DistilBERT	1.0	0.0	0.5	0.5	0.5	0.81	0.92	0.08	0.5	0.84	0.16	0.48
BERT$_{base}$	1.0	0.0	0.28	0.5	0.5	0.88	0.91	0.09	0.47	0.84	0.16	0.37
BERT$_{large}$	1.0	0.0	0.56	0.62	0.38	0.81	0.9	0.1	0.53	0.84	0.16	0.45
MAMBA 130M	1.0	0.0	0.83	0.62	0.38	0.75	0.81	0.19	0.76	0.76	0.24	0.76
MAMBA 790M	0.56	0.44	0.78	0.25	0.75	0.75	0.45	0.55	0.77	0.4	0.6	0.79
MAMBA 2.8B	1.0	0.0	0.56	0.69	0.31	0.81	0.89	0.11	0.62	0.89	0.11	0.78
LLaMA-2 7B	1.0	0.0	0.94	0.62	0.38	0.88	0.93	0.07	0.76	0.91	0.09	0.89
LLaMA-2 13B	1.0	0.0	0.72	0.62	0.38	0.81	0.91	0.09	0.72	0.9	0.1	0.87
LLaMA-3 8B	1.0	0.0	0.67	0.5	0.5	0.94	0.89	0.11	0.73	0.89	0.11	0.88
Mistral 7B	1.0	0.0	0.78	0.62	0.38	0.88	0.92	0.08	0.7	0.9	0.1	0.87
Mixtral-8x 7B	1.0	0.0	0.89	0.69	0.31	0.94	0.92	0.08	0.72	0.9	0.1	0.85
Qwen1.5 1.8B	1.0	0.0	0.5	0.75	0.25	0.88	0.9	0.1	0.61	0.85	0.15	0.77
Qwen1.5 7B	1.0	0.0	0.78	0.62	0.38	0.81	0.88	0.12	0.73	0.83	0.17	0.87
Qwen1.5 14B	1.0	0.0	0.78	0.75	0.25	0.75	0.88	0.12	0.71	0.82	0.18	0.85

Table 6: Sensitivity of each model to negation. S(+) and S(-) are the sensitivity definition taken from (Ettinger, 2020), and S* denotes the sensitivity definition of (Shivagunde et al., 2023).

Datasets	ROLE-88				ROLE-1500			
Models	**S(+)**	**S(+) 0.1**	**S(-)**	**S (-) 0.1**	**S(+)**	**S(+) 0.1**	**S(-)**	**S (-) 0.1**
ALBERT$_{base}$	0.39	0.2	0.61	0.2	0.47	0.12	**0.53**	0.11
ALBERT$_{large}$	0.41	0.24	0.59	0.37	0.54	0.23	0.46	0.22
ALBERT$_{xlarge}$	**0.56**	0.0	0.44	0.0	0.5	0.0	0.5	0.0
ALBERT$_{xxlarge}$	0.32	0.22	**0.68**	**0.56**	0.49	0.26	0.51	0.29
DistilBERT	0.37	0.29	0.63	0.39	0.51	0.27	0.49	0.23
BERT$_{base}$	0.44	0.34	0.56	0.39	0.5	0.26	0.5	**0.27**
BERT$_{large}$	0.34	0.24	0.66	0.46	0.52	0.28	0.48	**0.27**
MAMBA 130M	0.44	0.22	0.56	0.32	0.54	0.22	0.46	0.18
MAMBA 790M	0.49	0.0	0.51	0.0	0.48	0.01	0.52	0.0
MAMBA 2.8B	0.49	0.12	0.51	0.22	0.55	0.23	0.45	0.18
LLaMA-2 7B	0.39	0.27	0.61	0.37	0.52	0.26	0.48	0.25
LLaMA-2 13B	0.54	**0.44**	0.46	0.34	0.53	0.28	0.47	0.25
LLaMA-3 8B	**0.56**	0.41	0.44	0.29	0.52	0.27	0.48	**0.27**
Mistral 7B	**0.56**	0.37	0.44	0.39	**0.56**	**0.29**	0.44	0.24
Mixtral-8x 7B	0.54	0.39	0.46	0.37	0.54	0.27	0.46	0.24
Qwen1.5 1.8B	0.44	0.2	0.56	0.41	0.52	0.25	0.48	0.25
Qwen1.5 7B	0.46	0.34	0.54	0.46	0.52	**0.29**	0.48	0.25
Qwen1.5 14B	0.49	0.32	0.51	0.41	0.53	**0.29**	0.47	0.26

Table 8: Sensitivity of each model to role reversal. S(+) (original) and S(-) (role reversed) are the sensitivity definition taken from (Ettinger, 2020), 0.1 denotes that there should be a minimum 0.1 difference between probabilities of appropriate and inappropriate completion tokens.

Negation accuracies show that LLMs have the knowledge of hypernyms as expected. One model standing out in this task was Qwen1.5, achieving 0.83 accuracy even on top-1 accuracy, although the larger versions did not achieve the same performance. We see that LLMs except smaller versions of MAMBA are on par or better than encoder type models due to their parameter count and pre-training dataset size. Accuracies for datasets other than NEG-136-SIMP drop substantially, but they also have predictions that make sense (eg. Rockets and missiles are very ”dangerous” instead of ”fast”). Simply evaluating for a single correct word does not really show real success in any way. Even in the case of NEG-136-SIMP, the evaluation could be using WordNet by taking account multiple hypernyms which we will leave as future work.

Negation sensitivities for NEG-136-SIMP show LLMs also miserably fail in negation understanding by having 1.0 sensitivity for the affirmative form and 0 for the negative form of the sentence. As it can be seen on the tables, definition of (Shivagunde et al., 2023) does not correlate with the sensitivity definition of (Ettinger, 2020). Although there is a huge gap between affirmative and negative sensitivities, we observe very high results for the latter definition. Even in extended datasets and NEG-136-NAT, we observe a large gap between S(+) and S(-).

Role accuracies show that we generally have a better understanding of what a role might mean in context with LLMs. As it can be seen from LLaMA, Qwen1.5, and Mistral models’ performance in comparison to the others.

Role sensitivites show that we still see a similar decrease seen in (Ettinger, 2020) when we introduce a threshold to the probabilities of appropriate and inappropriate token predictions. It seems that the decrease is not very dependent on model size or architecture, suggesting the need for other developments. However, similar problem to the other tasks, evaluating sensitivity on two tokens are not very informative and should be extended to multiple tokens that can be appropriate or inappropriate for the given context.

4 Conclusions & Discussion

In conclusion, we found that LLMs still fail to understand negation but they are successful in terms of hypernym understanding, similar to previous smaller language models. Trials run on different types of architectures and sizes shows that this is not changing with architecture type or size and should be addressed some other way. We leave the qualitative evaluation, WordNet based hypernym evaluation of multiple tokens for negation variations, and multiple token evaluation of role variations to future work since we had a limited time and resources. We also plan to experiment with larger and full precision models to see if their emergent abilities coming from the large scale allow them to understand such concepts.

References

William F Battig and William E Montague. 1969. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of experimental Psychology*, 80(3p2):1.

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.

Ira Fischler, Paul A Bloom, Donald G Childers, Salim E Roucos, and Nathan W Perry Jr. 1983. Brain potentials related to stages of sentence verification. *Psychophysiology*, 20(4):400–409.

Namrata Shivagunde, Vladislav Lialin, and Anna Rumshisky. 2023. Larger probes tell a different story: Extending psycholinguistic datasets via in-context learning. *arXiv preprint arXiv:2303.16445*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.