SETLEXSEMCHALLENGE: Using Set Operations to Evaluate the Lexical and Semantic Robustness of Language Models

Bardiya Akhbari, Manish Gawali, Nicholas A. Dronen[1]
[1]Selling Partner Experience, Amazon

Introduction

Set theory is foundational to mathematics and, when sets are finite, to reasoning about the world.

An intelligent system should perform set operations consistently, regardless of superficial variations in the operands. Initially designed for semantically-oriented NLP tasks, large language models (LLMs) are now being evaluated on algorithmic tasks. Because sets comprise arbitrary symbols (e.g., numbers, words) of unconstrained type, they enable systematic interrogation of LLM robustness along several important dimensions important to real-world applications [1-3].

The SETLEXSEMCHALLENGE is a synthetic benchmark that assesses the robustness of LLMs' instruction-following abilities under various conditions, focusing on set operations and the nature and construction of the set members. Because set operations can be performed on objects of unconstrainted type, the types of the set members can be varied systematically to interrogate LLM robustness in several ways.

We evaluate seven LLMs on our benchmark with at least 12,000 tests each and find they exhibit poor robustness along all dimensions and, notably, that they are susceptible to distinct failure modes along the semantic dimension with SETLEXSEM'S deceptive sets.

Dataset

When constructing SETLEXSEM, we systematically vary the hyperparameters listed in Table 1. For a given hyperparameter set, we create a 50 occurrences of a prompt, each with different samples of the sets **A** and **B**. Outcomes are reported as the average accuracy across all runs.

Table 1. Hyperparameters of SETLEXSEM'S prompts.

Hyperparameter	Values
Operation	$\{\cap, \cup, \setminus, \triangle\}$
Operand size	$\{2, 4, 8, 16\}$
Token type	{number, word}
Token length	{undefined, 1, 2, 3, 4}
Token frequency*	Deciles $\{1, \ldots, 9\}$ of vocabulary by rank frequency
Semantic similarity*	Words in set A share one hypernum, in set B share another
Prompting method	{Simple baseline, Chain of thought (CoT)}
Demonstration phrasing	{natural, formal}
Number of in-context demonstrations	$\{0, 1, 3, 5\}$

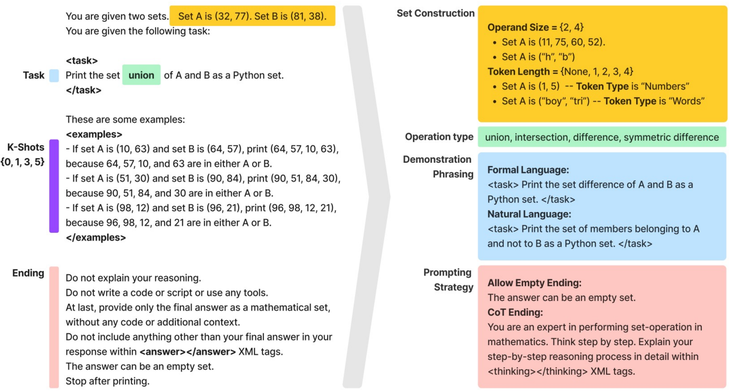

Figure 1. Example of our baseline prompt with sets of size two. Every prompt follows this template: set construction, task definition, demonstrations, and final instructions. Note that the baseline prompt instructs the LLM not to explain its reasoning whereas the chain-of-thought prompt instructs the model to think step by step. In this example, the set members are numbers and each token in a set is two characters long.

Analytical robustness

LLMs exhibit poor robustness along the analytical dimension. With respect to bias, set difference and symmetric difference are consistently more difficult. As expected, accuracy degrades with increasing set size.

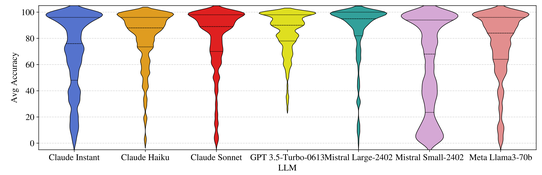

Figure 2. Distribution of accuracies across seven LLMs.

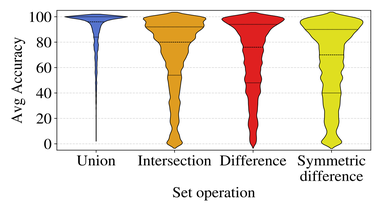

Figure 3. Distribution of accuracies by operation.

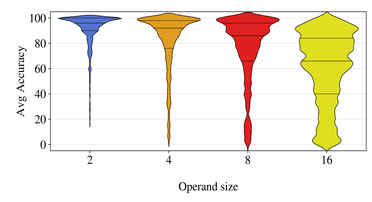

Figure 4. Distribution of accuracy by set size.

Lexical robustness

We use the Google Books N-grams corpus term frequencies to approximate training set frequencies. Accuracy is not invariant to the incidental features term length or term frequency. Terms of length 3 are less frequent than length 5 across all frequencies.

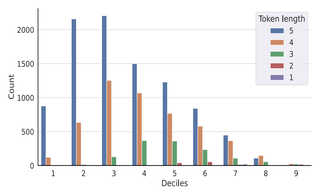

Figure 5. Term length across frequency deciles of vocabulary.

Table 2. Distributions of difference in accuracy of terms of length 5 and 3.

Decile	Mean	Std	Min	Max	N (x 50 run each)
1st	-	-	-	-	0
2nd	16.6	21.27	-8.00	84.00	96
3rd	11.2	17.11	-24.00	74.00	128
4th	6.4	11.40	-28.00	40.00	128
5th	1.7	10.22	-26.00	38.00	128
6th	4.9	11.50	-30.00	40.00	128
7th	4.2	14.50	-46.00	44.00	128
8th	2.5	9.19	-22.00	36.00	128
9th	3.8	23.91	-64.00	90.00	96

Constructing deceptive sets

To test semantic robustness, we construct sets of hyponyms by sampling hypernyms and constructing sets of their hyponyms. Sets with members swapped semantically contradict the set operation a language model is required to perform.

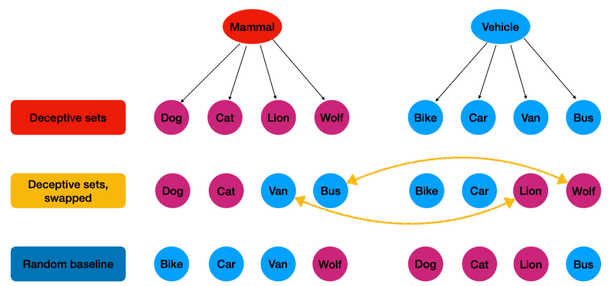

Figure 6. Constructing three types of sets for evaluating semantic robustness.

Semantic robustness

Our hypothesis that sampling hyponyms and swapping them confused the instruction following abilities of LLMs is borne out by Figure 8, which shows that sets with members swapped (orange) has consistently lower average accuracy and higher variance. That LLMs exploit semantic consistency when following instructions to perform set operations is borne out for the non-swapped sets (red), where the robustness on is greater than the random baseline. The exception is set union, for which the random baseline performed worst.

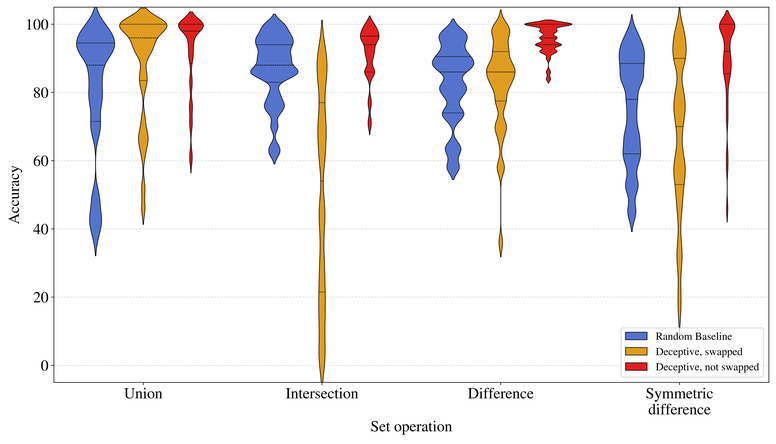

Figure 7. Distributions of accuracy of LLMs on sets comprising "deceptive" words. In the not-swapped case, sets are as they were originally sampled (with the words in a given set having a common hypernym). In the swapped case, half of the deceptive set members are swapped between sets. The random baseline is a random sampling of words from the same vocabulary.

Conclusion

While we have demonstrated here that today's LLMs are not robust to variations of the analytical and lexico-semantic features that SETLEXSEM tests, the long march of science towards greater understanding, and of technology towards greater sophistication, may imply that future systems may indeed be robust to such variations. System 2 thinking may be mechanized. In such a possible future, synthetic datasets like SETLEXSEM could be used to verify that systems that society has become generally confident in are indeed invariant in the ways we desire. In the meantime, our dataset and others like it serve as guideposts to systems designers indicating deficiencies that need to be corrected.

Notably, the failure mode that current LLMs exhibit on the "deceptive" sets of SETLEXSEM demonstrates that the relatedness of entities in the hidden states of an instruction-following neural network can subvert the instruction-following capabilities. To achieve high robustness, then, a model must be either architecturally equipped to, or at least explicitly trained to, balance instruction following and semantics. We hope that the research community sees this challenge as a worthy one to address in future model designs.

References

[1] Cobbe et al.(2021) Training verifiers to solve math word problems, arXiv:2110.14168
[2] Dua et al. (2019) DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraph, In Conference of the North American Chapter of the Association for Computational Linguistics
[3] Hendrycks et al.(2021) Measuring Massive Multitask Language Understanding, ICLR 2021

Conference
NeurIPS 2024 Track Datasets and Benchmarks
*Corresponding author: ndronen@amazon.com