1 Abstract

A major issue with Large Language Models (LLMs) is their tendency to hallucinate incorrect answers. For this reason, several metrics have been devised in the attempt to measure how certain a model is in its answer, or how likely a answer is correct. One such measure is the token distribution entropy. In this work, we propose a new measure, the **S**emantic **To**ken Top-**P** **S**imilarity (STOPS). We observe a moderate to strong inverse correlation between entropy and STOPS. Furthermore, we evaluate the suitability of classifiers based on entropy and STOPS for predicting model answer accuracy. We support all of our findings through extensive experiments conducted on a multitude of datasets, models and prompting approaches.

2 Introduction

3 Related Work

4 Methodology

5 Results

6 Correlation between token entropy and semantic similarity of tokens

The entropy of token probability distributions has been proposed as a measure for LLM uncertainty. We propose a new uncertainty measure, which is the average pairwise cosine similarity of the top-p tokens of a probability distribution. In the following, we will show that there is a inverse relationship

Model	Pearson r	Spearman r
Qwen3-8B	-0.37	-0.39
Mistral-7B-v0.1	-0.50	-0.67
Llama-3.1-8B-Instruct	-0.57	-0.70
Llama-3.1-8B	-0.62	-0.76
deepseek-llm-7b-base	-0.42	-0.65

		Pearson r	Pearson p
Model	Dataset		
Qwen3-8B	gsm8k	-0.37	0.00
	writingprompts	-0.38	0.00
	xsum	-0.37	0.00
Mistral-7B-v0.1	gsm8k	-0.50	0.00
	writingprompts	-0.49	0.00
	xsum	-0.51	0.00
Llama-3.1-8B-Instruct	gsm8k	-0.57	0.00
	writingprompts	-0.58	0.00
	xsum	-0.57	0.00
Llama-3.1-8B	gsm8k	-0.62	0.00
	writingprompts	-0.61	0.00
	xsum	-0.63	0.00
deepseek-llm-7b-base	gsm8k	-0.42	0.00
	writingprompts	-0.43	0.00
	xsum	-0.45	0.00

7 Future Work

Especially for the accuracy analysis, it is still open to perform it on more datasets and perhaps find a dataset where the cosine is a stronger indicator than entropy.