

DIVERSITY MEASURES FOR ERROR PREDICTION IN LARGE LANGUAGE MODELS

Noel Ngu, Computer Science

Mentor: Prof. Paulo Shakarian, PhD

School of Computing and Augmented Intelligence

Motivation and Background

Why is it important? The performance issues of language models in reasoning tasks accentuates the need for robust and reliable error prediction models to be used as introspective tools for language models.

Where did we start? This project builds upon the research team's previous work with error-prediction in large-language models for math word problems. While the research team previously used the presence of math symbols as features for error prediction, this model instead uses various measures of diversity to predict errors in large-language models. This generalizes the research team's previous work and allows it to be applicable to various other reasoning tasks.

Methodology

- The language model used in this work is GPT-3.5. This model is prompted with questions from publicly available datasets: DRAW-1K, CSQA and Last-Letter Concatenation. These datasets contain not only the question, but also the answer key by which responses can be evaluated. 20 responses are collected for each question and the model is also prompted at various settings, modifying a hyperparameter known as temperature.
- The 20 responses to each question are embedded using an open-source embedding model known as **Sentence Transformer**. The resulting vector of numbers produced by the embedding model is averaged. This average is called the centroid.
- Using the 20 sampled paths, the entropy, Gini impurity and average distance from the centroid is calculated.


 Given two whole numbers, let's call them the first number and the second number. The first number is three times the second number. When 20 is added to the second number, it equals the first number plus 6.

Figure 1: GPT-3.5's response to the question: One whole number is three times a second. If 20 is added to the smaller number, the result is 6 more than the larger.

$$H = - \sum_{e \in U} P(e) \log(P(e)) \quad (1)$$

$$G = 1 - \sum_{e \in U} P(e)^2 \quad (2)$$

Figure 2: The equations for Shannon entropy (1) and Gini impurity (2)

- The machine-learning libraries used for error-prediction are Scikit-Learn and Tensorflow. The following models are implemented: XGBoost, AdaBoost as well as Multi-Layer Perceptron models of different number of layers. The features used to predict errors in large-language models are the LLM's entropy, Gini impurity and average distance from the centroid using the 20 samples from each question.

Conclusion

The machine-learning models are able to predict cases in which ChatGPT fails to answer certain questions with reasonable accuracy, precision and recall. This model can be used in large-language model evaluation and can be used to further improve large-language models in the future or be used to improve reliability in large-language model responses.

Acknowledgement

This project was done under the supervision of Prof. Paulo Shakarian. This project builds off of the research team's previous work with predicting ChatGPT performance, using the presence of math symbols as features in order to predict performance.

Objective

To implement **accurate error-prediction model for large-language models** that:

- Demonstrate generalizability across diverse reasoning tasks.
- Enhance the reliability of large-language model on reasoning tasks.

Results

- A positive correlation between the entropy, Gini impurity and average distance from centroid and probability of failure is observed. This pattern is noticed at various temperature settings and across all the datasets tested during the research team's experiments.

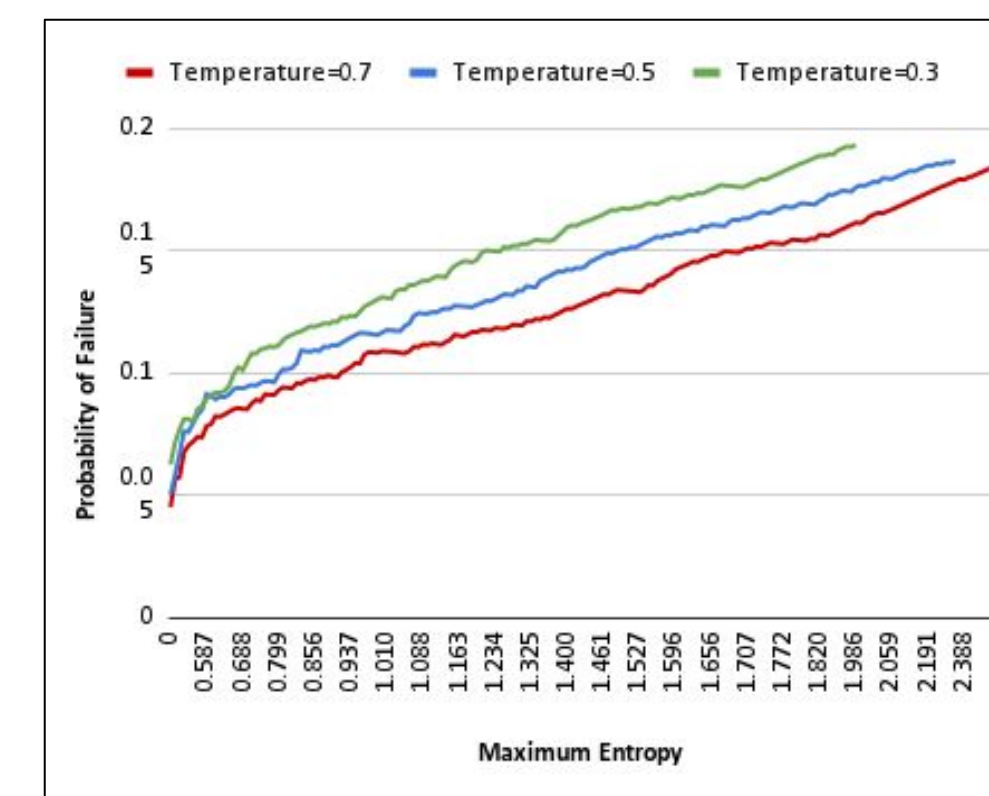


Chart 1: Minimum Entropy vs Probability of Failure

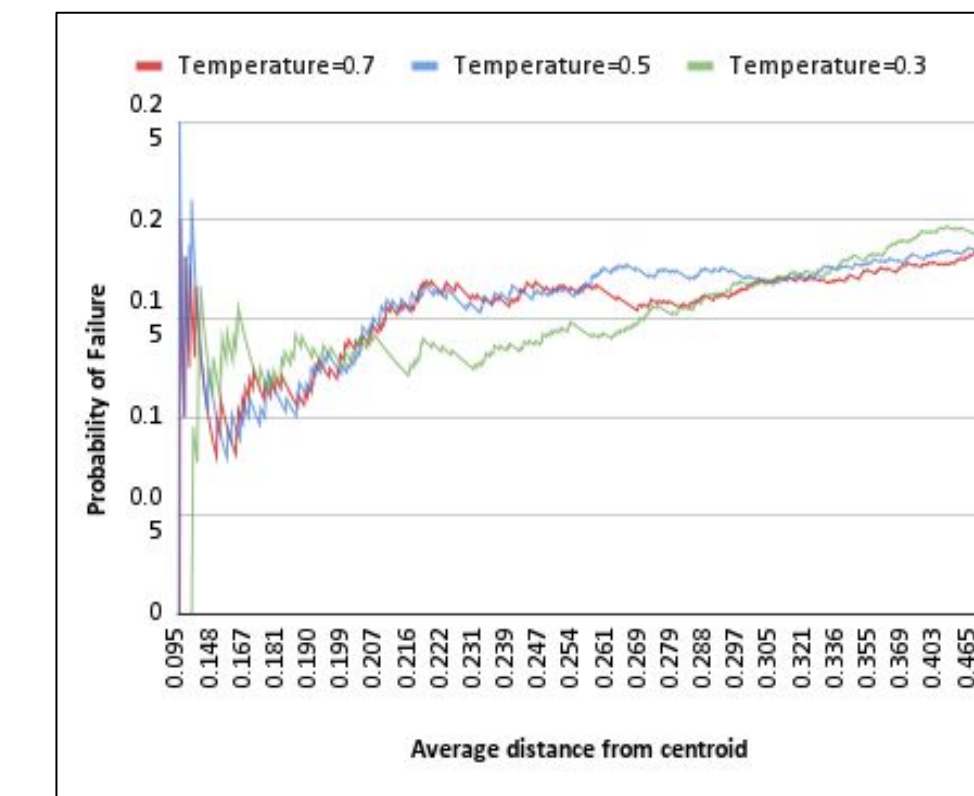


Chart 2: Avg. Distance From Centroid vs Probability of Failure

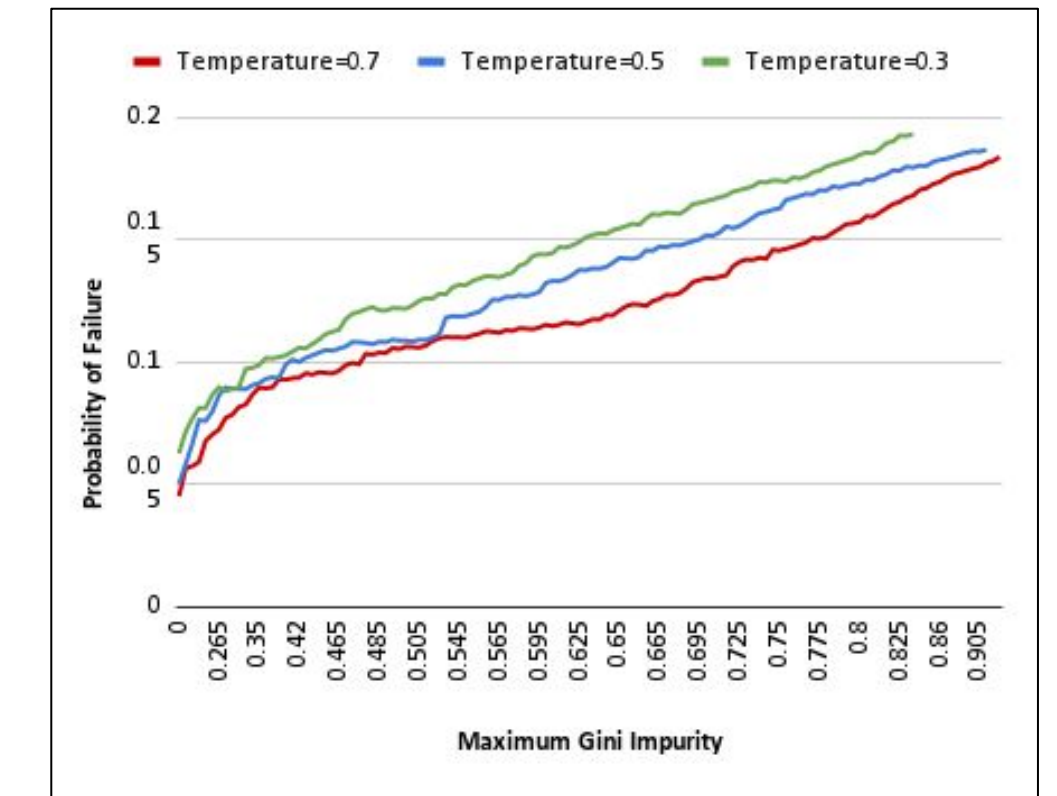


Chart 3: Minimum Gini Impurity vs Probability of Failure

Figure 3: Graphs of various measures of diversity plotted against the model's probability of failure. There is a notable positive correlation between the probability of the language model's responses being incorrect and these diversity measures.

- Models are able to predict errors in language models with good precision and recall for both classes using the features that were generated: entropy, Gini impurity and average distance from centroid.

name	Accuracy	Incorrect		Correct	
		Precision	Recall	Precision	Recall
5 Layer MLP	0.6583333	0.5983843	0.6773185	0.7202482	0.6435338
10 Layer MLP	0.6826666	0.6239400	0.7033225	0.7441588	0.6665741
15 Layer MLP	0.692	0.6241188	0.7424556	0.7676899	0.6530502
25 Layer MLP	0.6766666	0.6026953	0.8004123	0.7947384	0.5808451
30 Layer MLP	0.6543333	0.5960767	0.7583428	0.7937190	0.5738300
AdaBoost	0.5423333	0.2998840	0.5655172	0.3903725	0.5238938
XGBoost	0.4613333	0.3486666	0.8	0.1126666	0.2

name	Accuracy	Incorrect		Correct	
		Precision	Recall	Precision	Recall
XGBoost	0.6249056	0.5888387	0.5013175	0.6892052	0.7036809
AdaBoost	0.6652287	0.5893148	0.4598704	0.6984473	0.7959055
5 Layer MLP	0.6651261	0.5883610	0.4638313	0.6992691	0.7932187
10 Layer MLP	0.6096846	0.5471638	0.5738563	0.5599483	0.6323767
15 Layer MLP	0.6085843	0.5474060	0.5656786	0.5576379	0.6358729
25 Layer MLP	0.6580418	0.4780304	0.3700431	0.6830162	0.8412460
30 Layer MLP	0.5222165	0.1555441	0.4	0.3666723	0.6

name	Accuracy	Incorrect		Correct	
		Precision	Recall	Precision	Recall
XGBoost	0.6363880	0.2775904	0.6090090	0.8800852	0.6425931
AdaBoost	0.6923631	0.3332741	0.6630630	0.9018539	0.6989226
5 Layer MLP	0.7493034	0.4026896	0.6740240	0.9132778	0.7663474
10 Layer MLP	0.7483084	0.3942774	0.6629129	0.9109294	0.7674996
15 Layer MLP	0.6262985	0.3415213	0.7602102	0.7347488	0.5956980
25 Layer MLP	0.7472587	0.3934971	0.6578078	0.9095330	0.7674846
30 Layer MLP	0.6423034	0.1719450	0.4702702	0.6892325	0.6801062

Figure 4: Error prediction models for the datasets used in the experiment. The training dataset is undersampled and the language models are prompted at a temperature setting of 0.7. A color of green represents a better score whereas a color of red represents a worse score.