Generalizability and Robustness of Large Language Models Detecting Alzheimer's Disease from Speech

Jekaterina Novikova
Winterlight Labs | Cambridge Cognition

1. Introduction

Large language model (LLMs) have demonstrated promising performance in many tasks. However, before deploying them in real-world healthcare applications, these models need to be assessed on generalizability and robustness. LLMs being pre-trained on large unlabeled text have shown remarkable *generalization* ability for certain NLP tasks. However, when confronted with carefully designed synthetic samples, their *robustness* - the ability to gracefully deal with small perturbations - suffers significantly.

We evaluate and compare generalizability and robustness of multiple language models of varying size and complexity.

Table 1: Characteristics of the models used in this work

Type	Size	Models	# param	Train data size
Encoder	medium	BERT	110 M	3,300M tokens
	medium	RoBERTa	125 M	160Gb of text
Decoder	medium	GPT-Neo-125M	125 M	300B tokens
	medium	Bloom-560M	560 M	350B tokens
	large	Falcon-40B-Instruct	40 B	1,000B tokens
	large	StableBeluga2	70 B	2 trillion tokens

2. Methodology

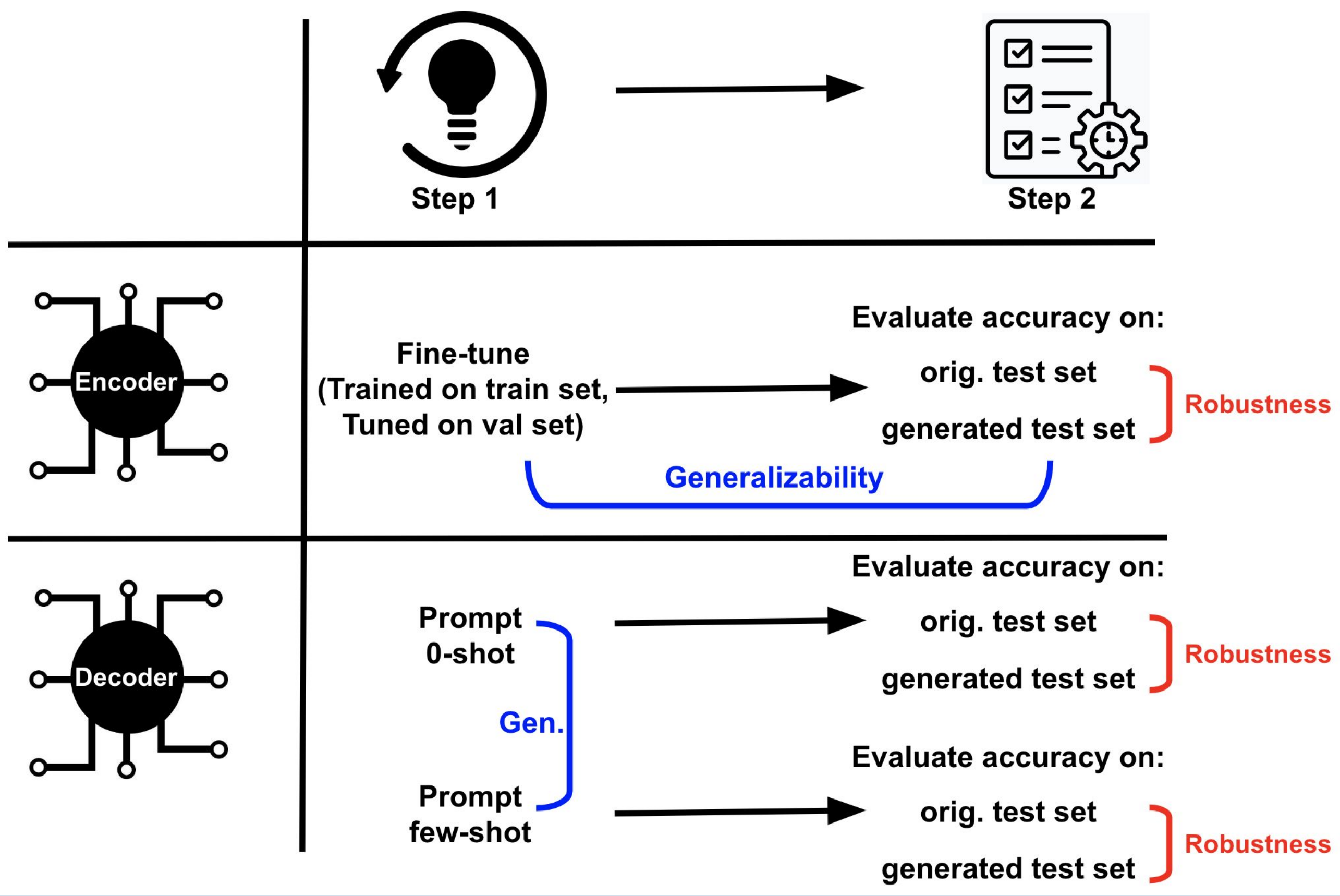
Data

- We use the ADReSS dataset (Luz et al., 2020) with 156 speech samples and associated transcripts from non-AD (N =78) and AD (N =78) English-speaking participants.
- Speech is elicited through the Cookie Theft picture.
- ADReSS dataset is well balanced in terms of age and gender and perfectly balanced between classes.

Models (Table 1)

- BERT
- RoBERTa
- GPT-NEO-125M
- BLOOM-560M
- Falcon-40B-Instruct
- StableBeluga2

Figure 1. Model evaluation process



Step 1
Step 2

Encoder
Fine-tune
(Trained on train set,
Tuned on val set)
Evaluate accuracy on:
orig. test set
generated test set
Robustness
Generalizability

Decoder
Prompt
0-shot
Gen.
Prompt
few-shot
Evaluate accuracy on:
orig. test set
generated test set
Robustness
Evaluate accuracy on:
orig. test set
generated test set
Robustness

2. Generalizability vs robustness

In order to evaluate the models for ***generalizability***, we compare their performances:

- between accuracy results reported on validation and test subsets of data,
- between accuracy results reported for zero- and few-shot settings, whenever applicable.

We consider the model generalizable if test set accuracy is not lower than the validation set accuracy.

To be able to evaluate the models for ***robustness***, we create a synthetic dataset by prompting the StableBeluga2 model to generate a 'noisy' version of the original text that makes it sound more/less like the speech of a person with dementia for the samples labelled as AD/HC. We then compare performance on the original and generated test sets.

We consider the model robust if accuracy on the original test set is similar to that on the generated test set.

3.1. Results

Generalizability (Table 2)

- BERT is able to generalize.
- RoBERTa overfits to the training data.
- Decoder-based models - no claims on their ability to generalize.

Robustness (Table 3)

- BERT and RoBERTa models are susceptible to relatively minor changes in testing data.
- For the decoder-based models, the synthetic generated test set was indeed easier to deal with.

The results of robustness assessment are different from the generalizability evaluation results:

- Most of the models were able to outperform the random baseline on a generated test set.
- Variance of model performance was not as pronounced as in generalizability experiments, indicating sufficient robustness abilities of the models.

Table 3: Robustness evaluation results

	Original test set	Generated test set
BERT	**0.82**	0.77
RoBERTa	0.79	0.77
GPT-Neo-125M	0.50	0.52
Bloom-560M	0.50	0.50
Falcon-40B-Instruct	0.54	0.56
StableBeluga2	0.69	**0.83**

Table 2: Generalizability evaluation results

	Val set acc	Test set acc
BERT	0.73	**0.82**
RoBERTa	**0.91**	0.79
GPT-Neo-125M	0.50 | 0.50	0.50 | 0.50
Bloom-560M	0.36 | 0.50	0.50 | 0.50
StableBeluga2	0.55 | 0.50	0.48 | 0.69

3.2. Results

- Fine-tuned encoder-only models are able to achieve substantially higher performance (accuracy in the range of 0.73-0.91) than decoder-only models of any size, either in zero- or few-shot settings (0.69 accuracy the highest).
- Amongst the decoder models, GPT-NEO-125M and BLOOM-560M are not capable of achieving a higher than random level of performance in either zero- or few-shot settings.
- Larger models though show a significantly higher than random performance level (accuracy of 0.69 for StableBeluga2).
- The task of AD detection is sufficiently challenging and requires either a larger annotated dataset for fine-tuning or a very complex pre-trained model for successful learning.