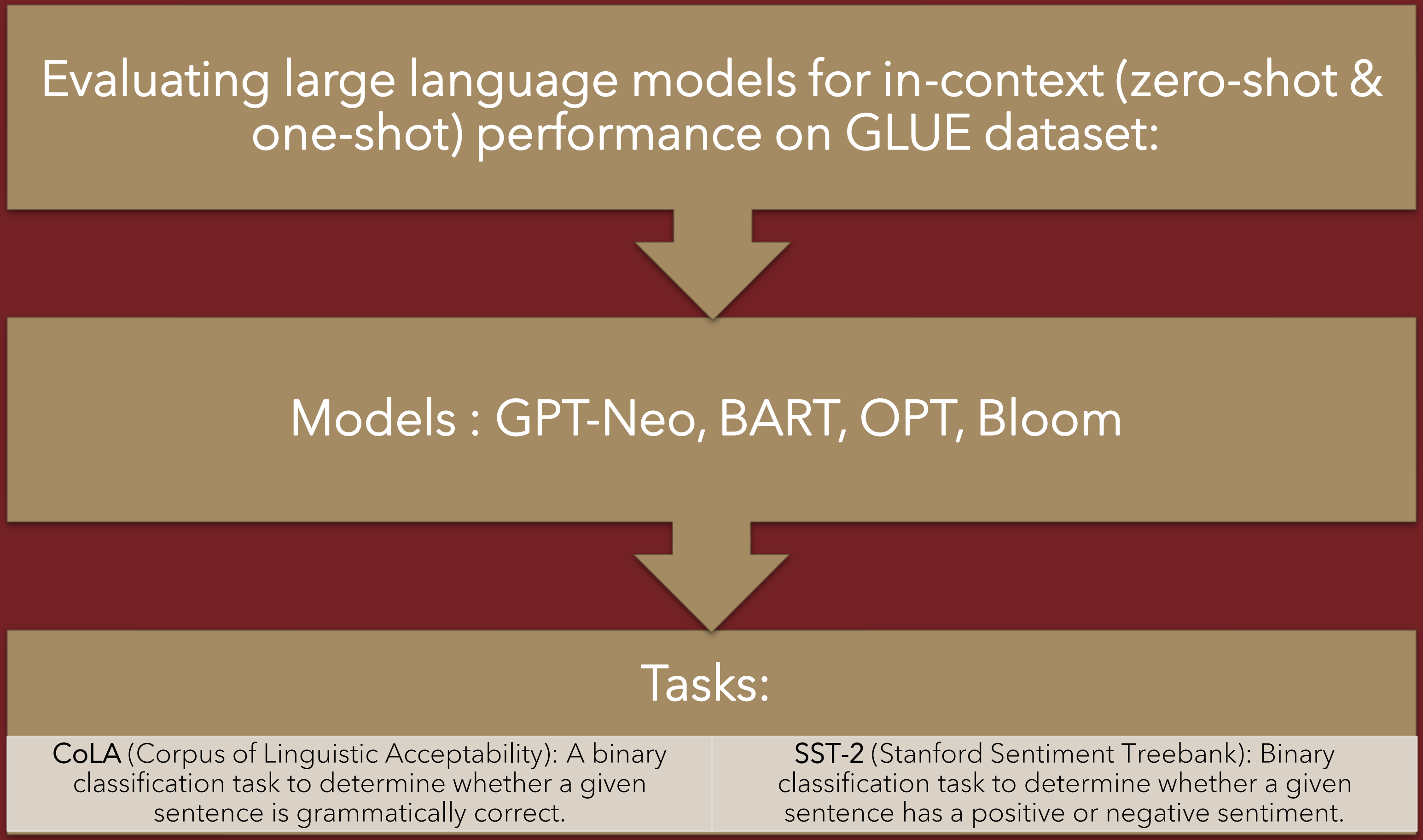
COMPARATIVE ANALYSIS: EVALUATING THE PERFORMANCE OF LARGE LANGUAGE MODELS ON NLP TASKS

Final Report – Abhinav, Arshia, Sai, Sushumna

PROBLEM STATEMENT



Evaluating large language models for in-context (zero-shot & one-shot) performance on GLUE dataset:

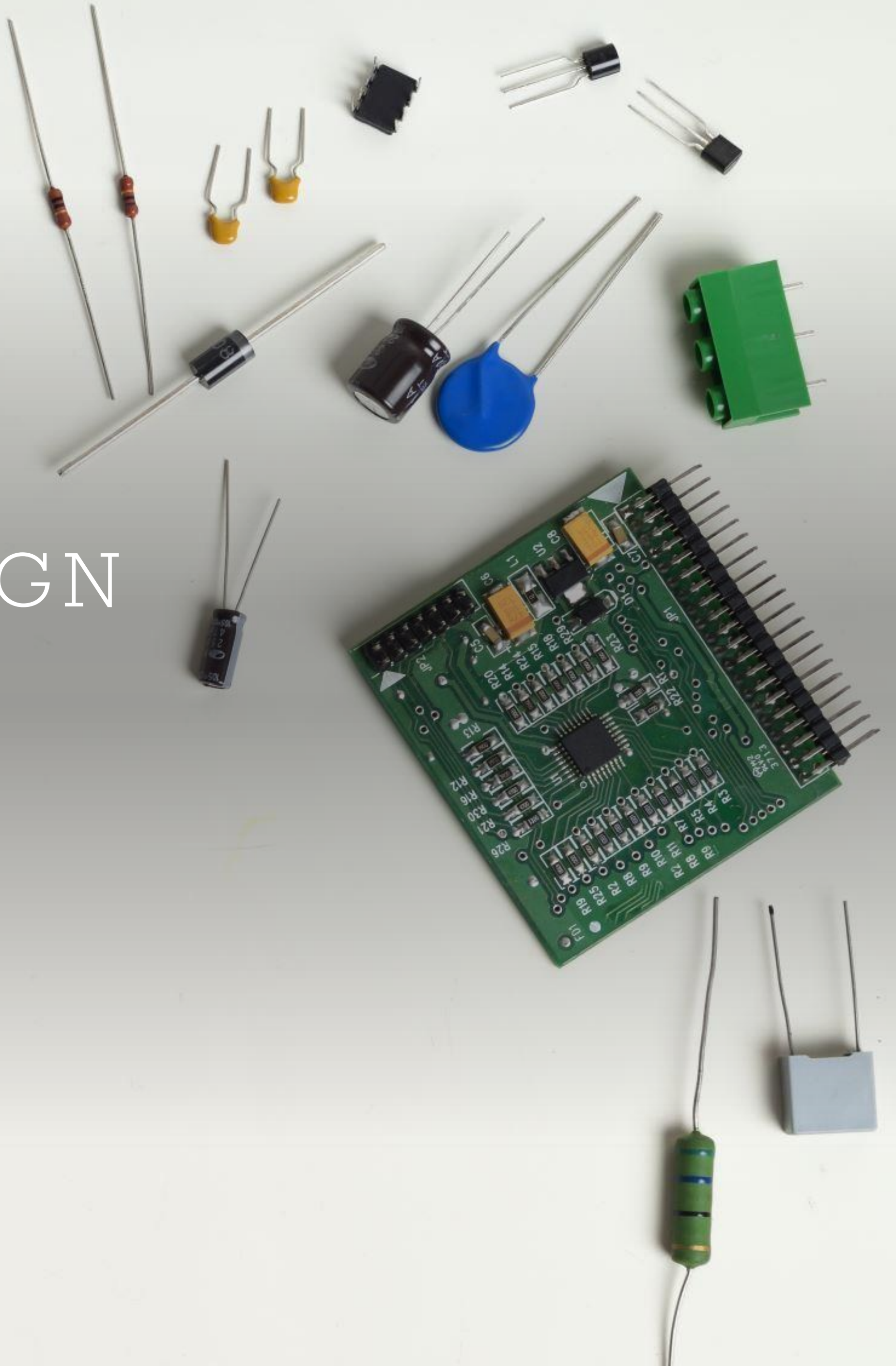
Models : GPT-Neo, BART, OPT, Bloom

Tasks:

CoLA (Corpus of Linguistic Acceptability): A binary classification task to determine whether a given sentence is grammatically correct.

SST-2 (Stanford Sentiment Treebank): Binary classification task to determine whether a given sentence has a positive or negative sentiment.

EXPERIMENT DESIGN



EXPERIMENT DESIGN

ZeroShot Prompts

CoLA: "Determine if the following sentence is grammatically correct: Sentence: '{sentence}':

SST-2: "Determine the sentiment of the following sentence: Sentence: '{sentence}', Sentiment: "

OneShot Prompts

CoLA: "Given the example: Sentence: 'He went to the store.'\n- Grammatically correct: yes

- prompt: Determine if the following sentence is grammatically correct:\n- Sentence: '{sentence}'\n- Grammatically correct: "

SST-2: "Given the example:\n- Sentence: 'I love this movie!'\n- Sentiment: positive.

- prompt: Determine the sentiment of the following sentence:\n- Sentence: '{sentence}'\n- Sentiment: "

Few Shot Prompts

CoLA: k: 3 examples:

- "Sentence: 'He went to the store.'\n Grammatically correct: yes"
- "Sentence: 'The children was playing.'\n Grammatically correct: no"
- "Sentence: 'She is writing an essay.'\n Grammatically correct: yes"

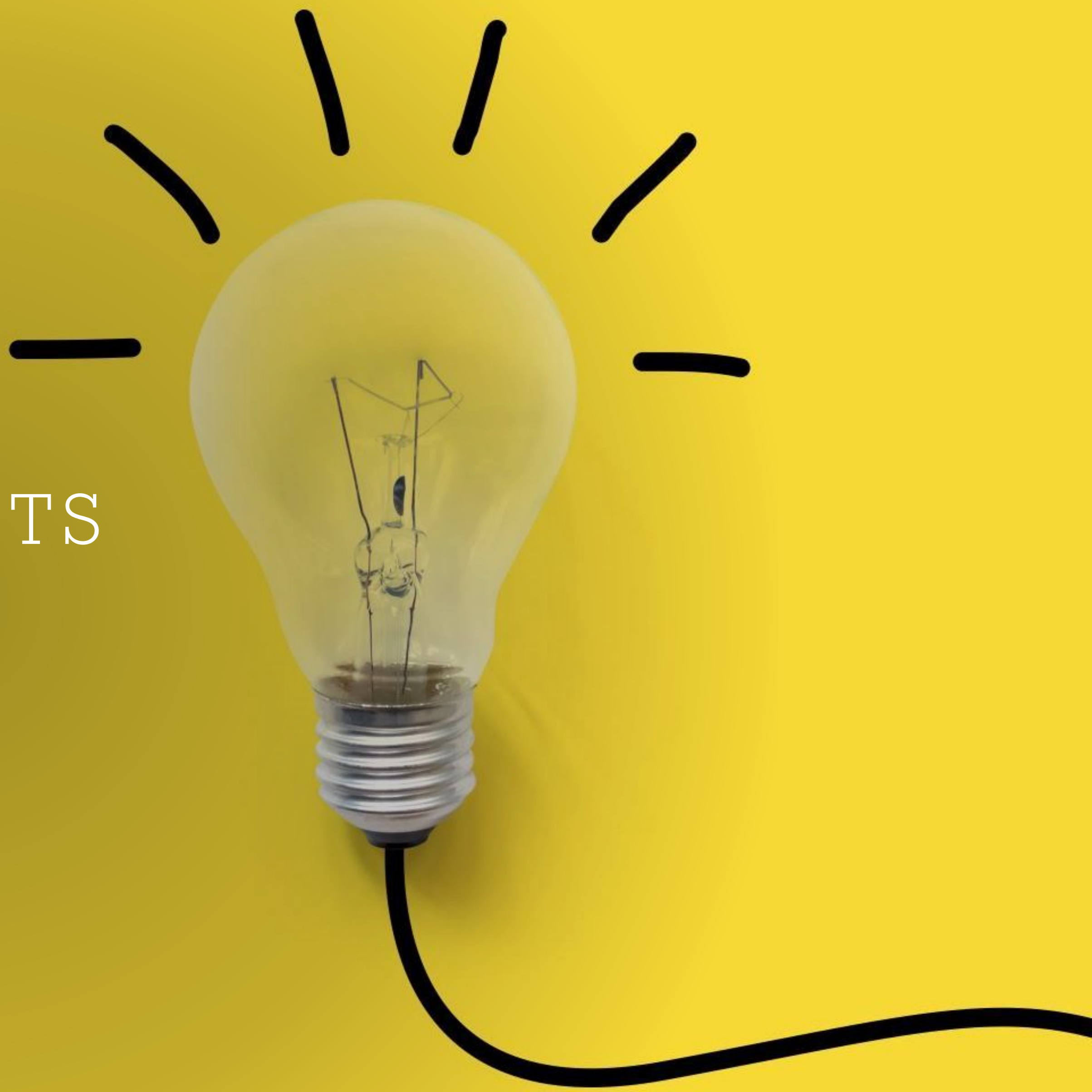
prompt: "Given the examples:\n{examples}\n\nDetermine if the following sentence is grammatically correct:\n- Sentence: '{sentence}'\n- Grammatically correct: "

SST-2: k: 3 examples:

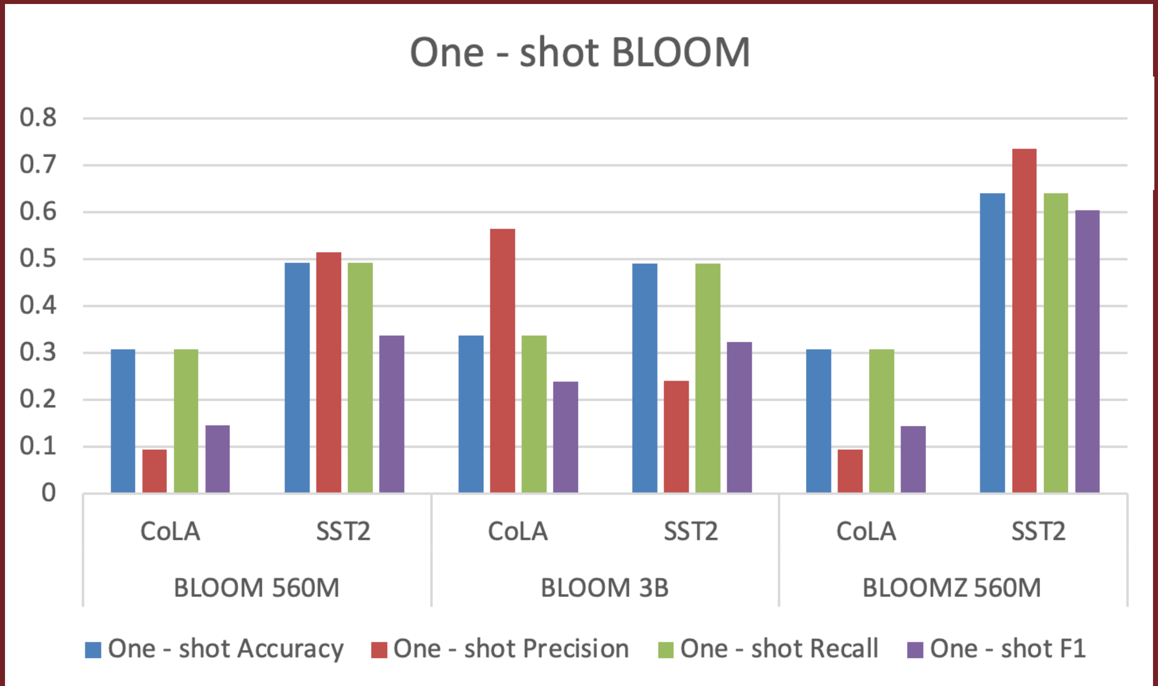
- "Sentence: 'I love this movie!'\n Sentiment: positive"
- "Sentence: 'The food was terrible.'\n Sentiment: negative"
- "Sentence: 'This book is really boring.'\n Sentiment: negative"

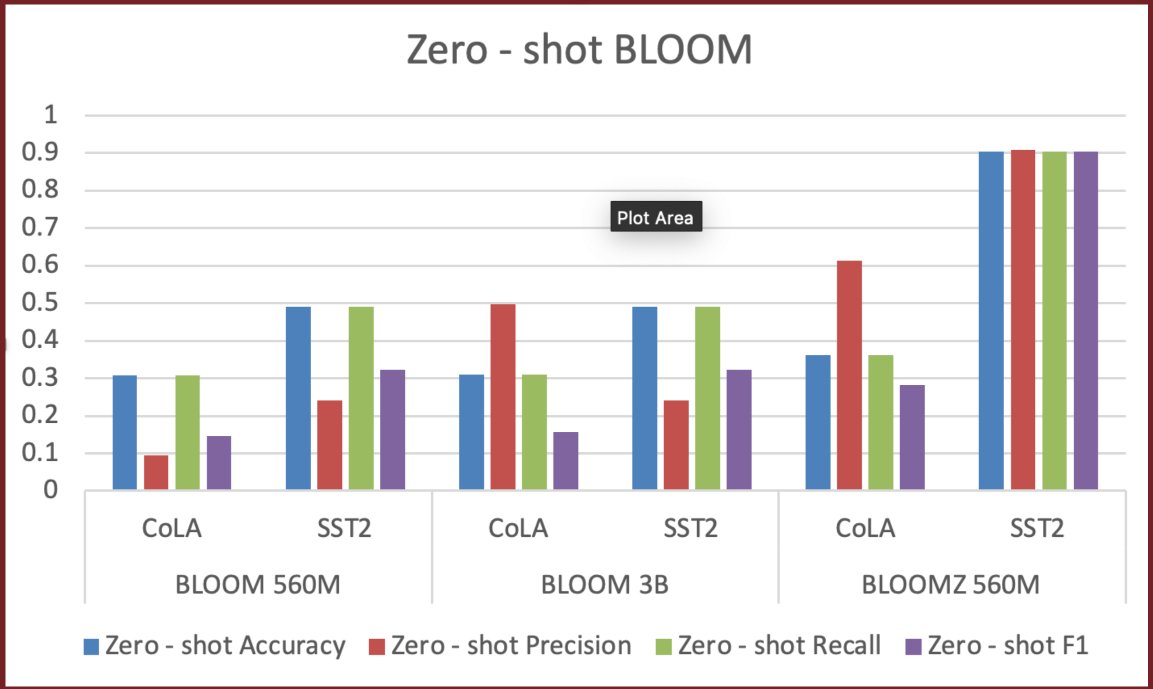
prompt: "Given the examples:\n{examples}\n\nDetermine the sentiment of the following sentence:\n- Sentence: '{sentence}'\n- Sentiment: "

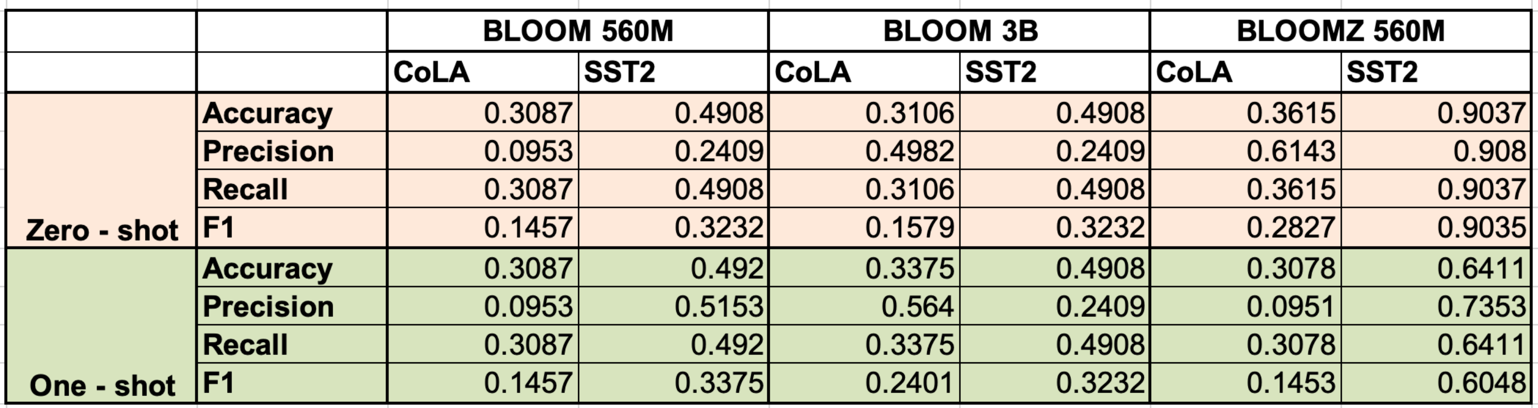
RESULTS



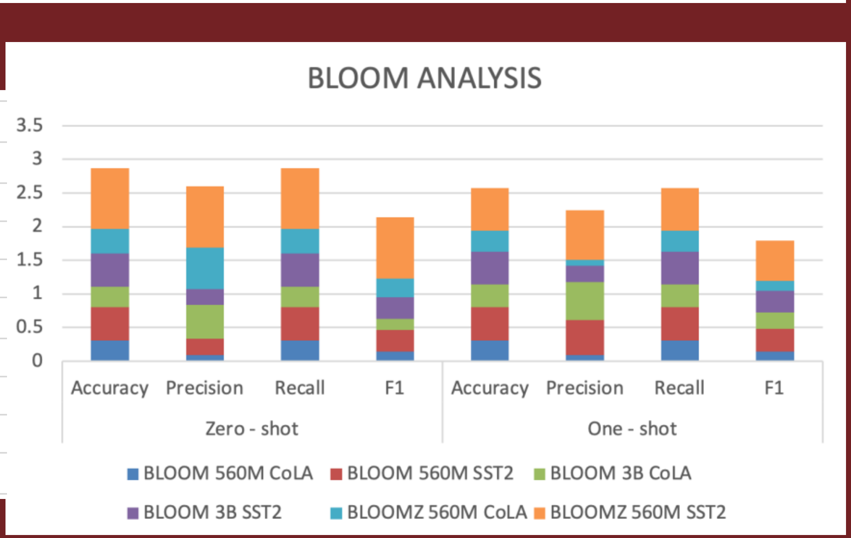
BLOOM RESULTS





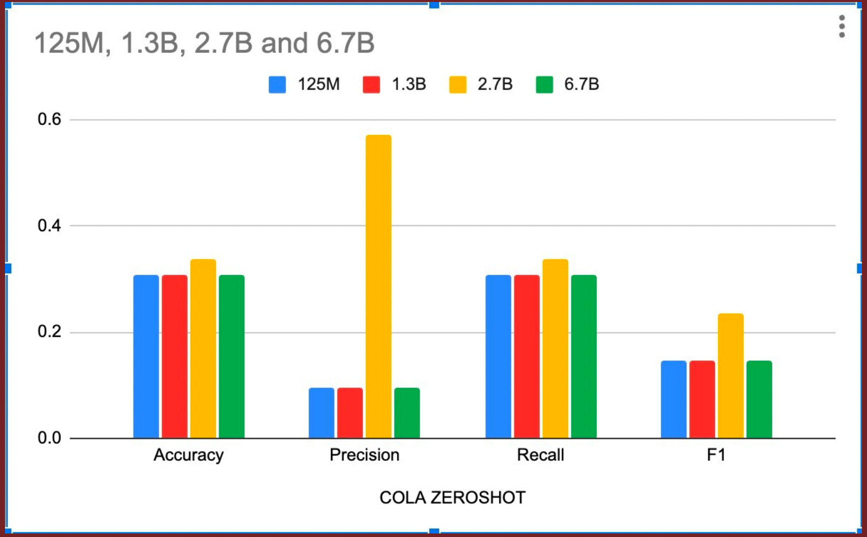


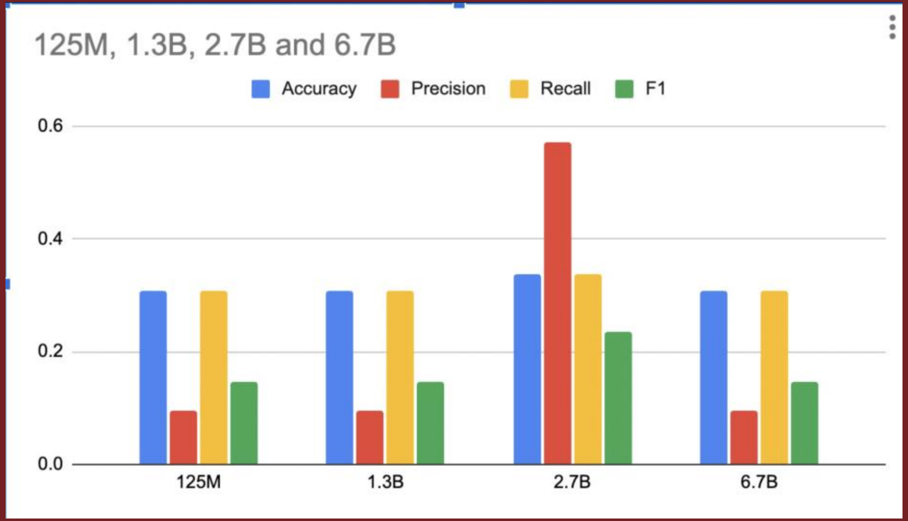
		BLOOM 560M		BLOOM 3B		BLOOMZ 560M	
		CoLA	SST2	CoLA	SST2	CoLA	SST2
Zero - shot	Accuracy	0.3087	0.4908	0.3106	0.4908	0.3615	0.9037
	Precision	0.0953	0.2409	0.4982	0.2409	0.6143	0.908
	Recall	0.3087	0.4908	0.3106	0.4908	0.3615	0.9037
	F1	0.1457	0.3232	0.1579	0.3232	0.2827	0.9035
One - shot	Accuracy	0.3087	0.492	0.3375	0.4908	0.3078	0.6411
	Precision	0.0953	0.5153	0.564	0.2409	0.0951	0.7353
	Recall	0.3087	0.492	0.3375	0.4908	0.3078	0.6411
	F1	0.1457	0.3375	0.2401	0.3232	0.1453	0.6048



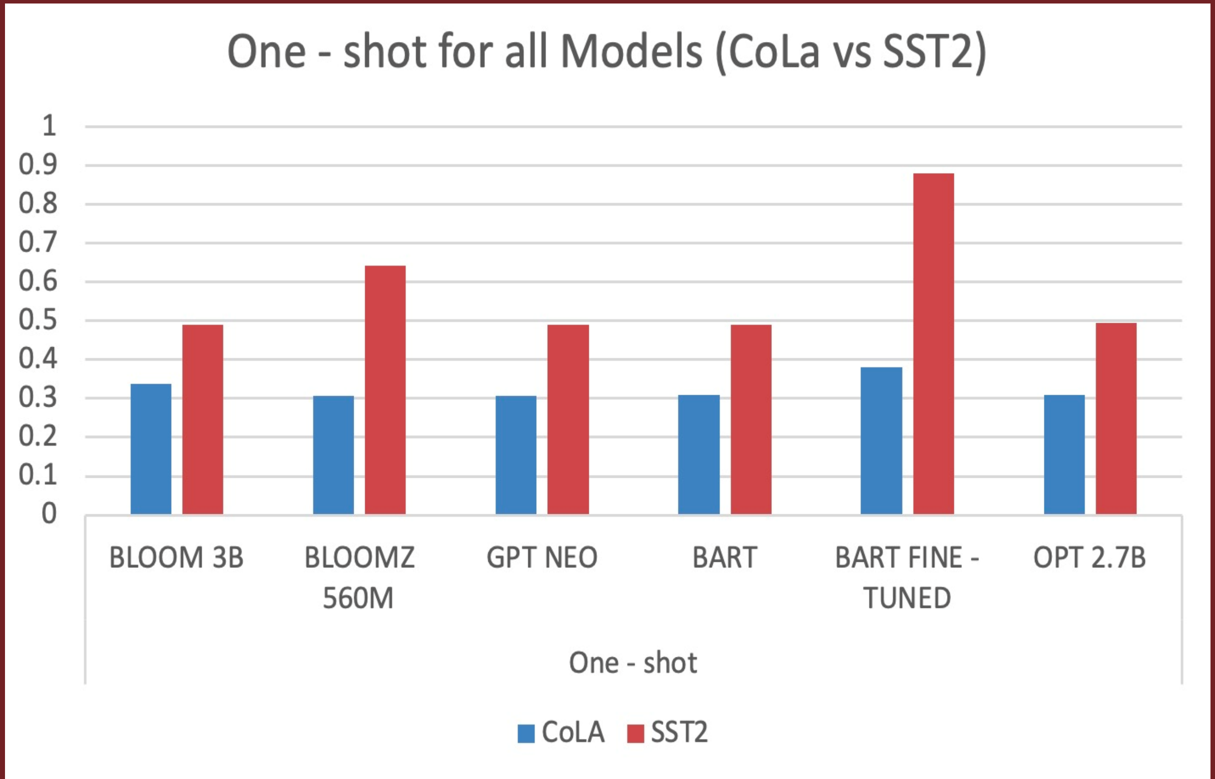
OPT RESULTS

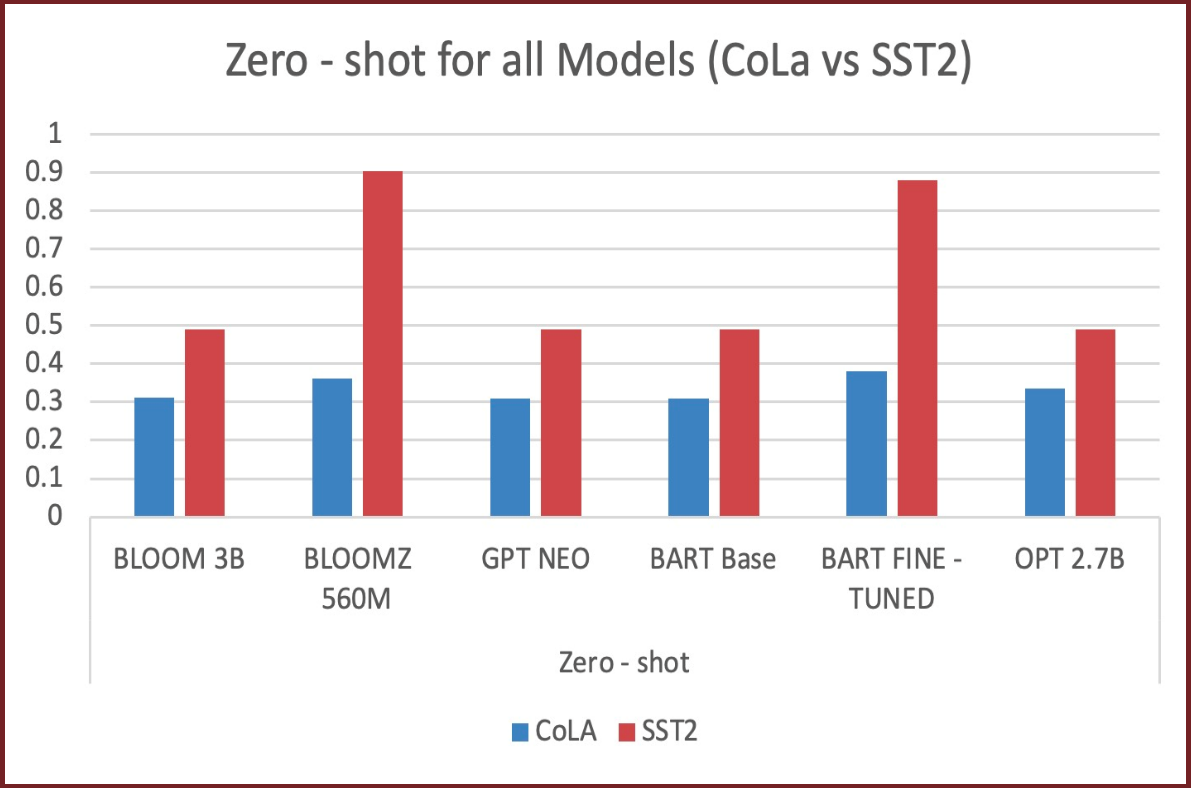
		125M		1.3B		2.7B		6.7B	
		CoLA	SST2	CoLA	SST2	CoLA	SST2	CoLA	SST2
Accuracy	zero shot	0.3087	0.4908	0.3087	0.4908	0.3365	0.4908	0.3087	0.4908
	one - shot	0.3087	0.4908	0.3087	0.4897	0.3087	0.4943	0.3087	0.4759
Precision	zero shot	0.0953	0.2409	0.0953	0.2409	0.5711	0.2409	0.0953	0.2409
	one - shot	0.0953	0.2409	0.0953	0.4103	0.0953	0.555	0.0953	0.4763
Recall	zero shot	0.3087	0.4908	0.3087	0.4908	0.3365	0.4908	0.3087	0.4908
	one - shot	0.3087	0.4908	0.3087	0.4897	0.3087	0.4943	0.3087	0.4759
F1	zero shot	0.1457	0.3232	0.1457	0.3232	0.2347	0.3232	0.1457	0.3232
	one - shot	0.1457	0.3232	0.1457	0.3247	0.1457	0.3405	0.1457	0.4603





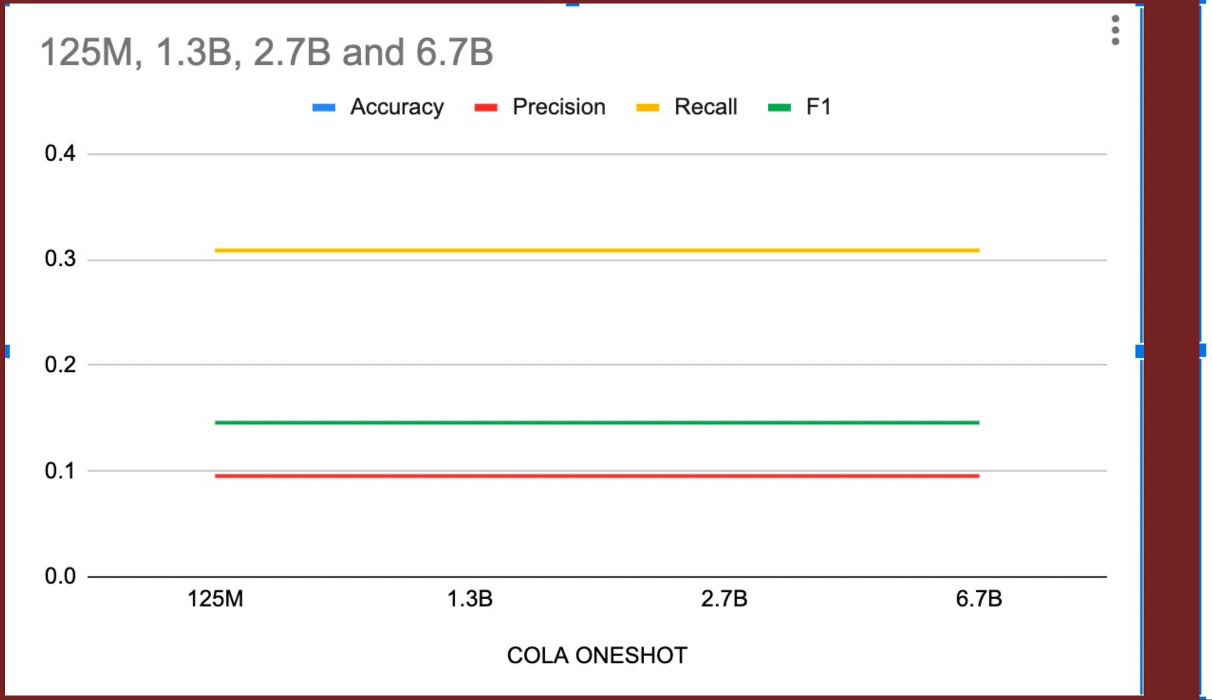
RESULTS: ONE SHOT | ZERO SHOT

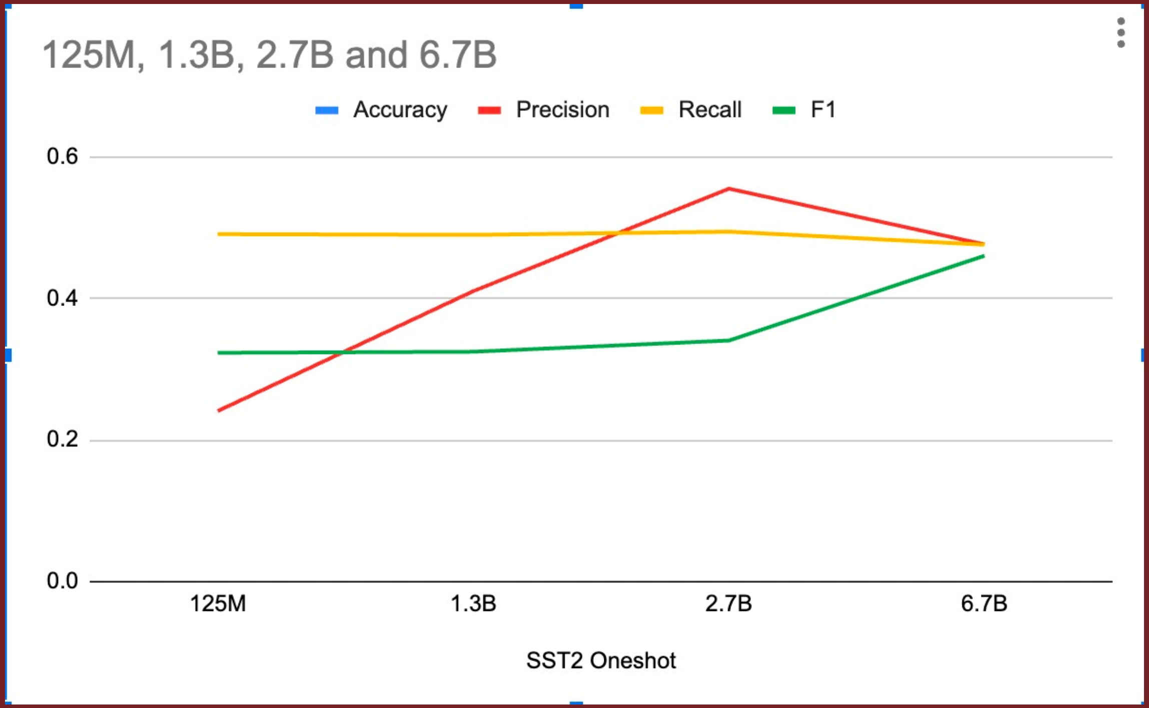




RESULTS: PARAMETER INCREASE

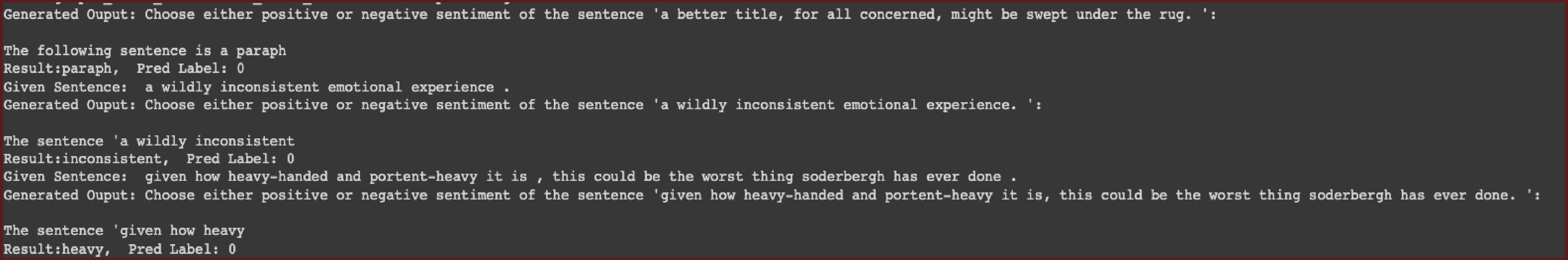
Model Size/Parameter increase doesn't show any increase in the Performance of the tasks.





INTERPRETATION OF RESULTS

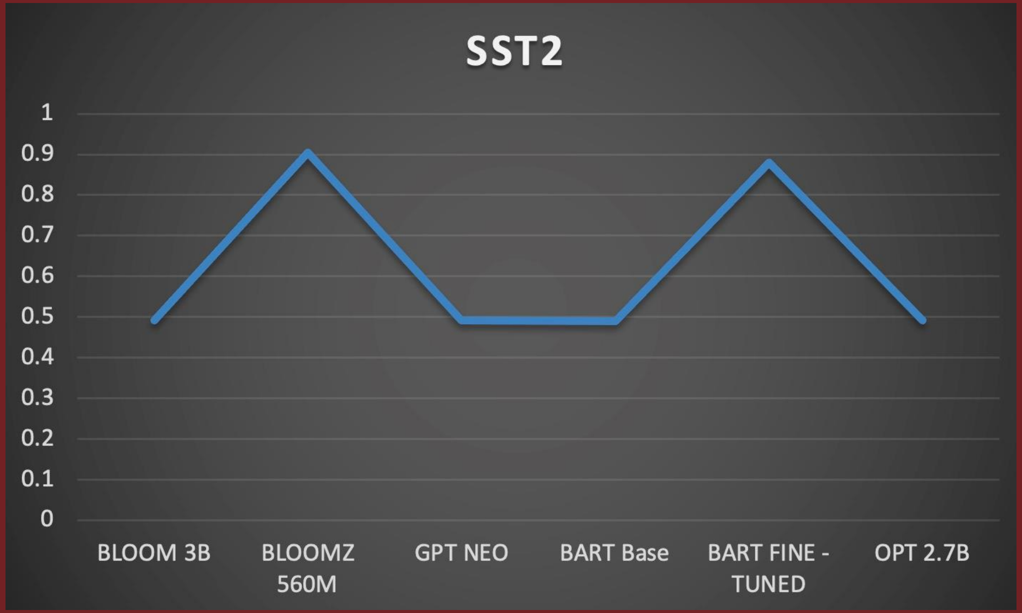
- All selected Pre-Trained models selected have shown to have similar scores for all classification metrics.
- This is because these models are not able to generate meaningful results due to lack of contextual understanding.

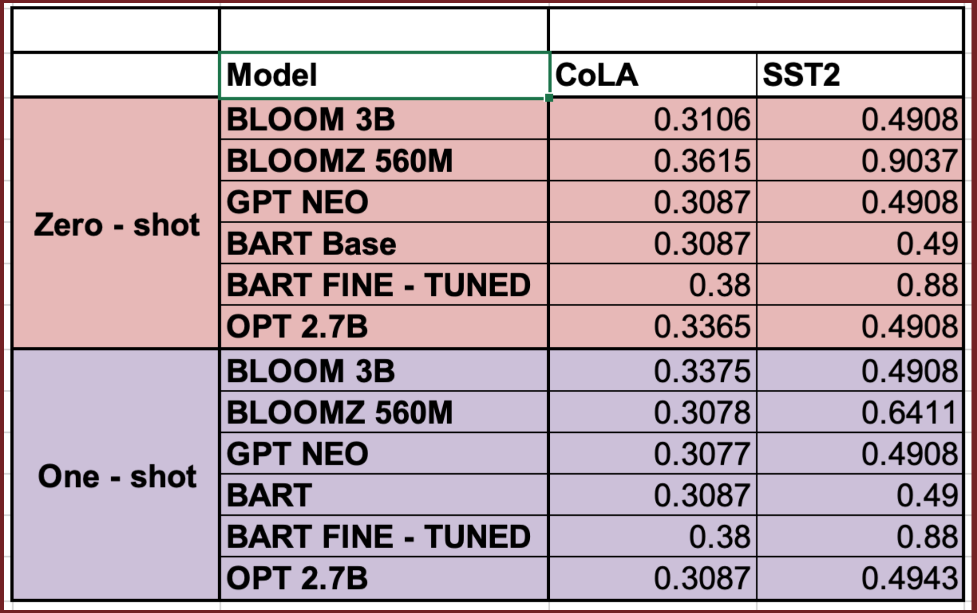


```
Generated Ouput: Choose either positive or negative sentiment of the sentence 'a better title, for all concerned, might be swept under the rug. ':

The following sentence is a paraph
Result:paraph,  Pred Label: 0
Given Sentence:  a wildly inconsistent emotional experience .
Generated Ouput: Choose either positive or negative sentiment of the sentence 'a wildly inconsistent emotional experience. ':

The sentence 'a wildly inconsistent
Result:inconsistent,  Pred Label: 0
Given Sentence:  given how heavy-handed and portent-heavy it is , this could be the worst thing soderbergh has ever done .
Generated Ouput: Choose either positive or negative sentiment of the sentence 'given how heavy-handed and portent-heavy it is, this could be the worst thing soderbergh has ever done. ':

The sentence 'given how heavy
Result:heavy,  Pred Label: 0
```

INTERPRETATION RESULTS :ACCURACY OF ALL MODELS

- Fine Tuning & Instruction Tuned Model shows a significant increase in the accuracy compared to just Pretrained Language Modelling.


SST2



	Model	CoLA	SST2
Zero - shot	BLOOM 3B	0.3106	0.4908
	BLOOMZ 560M	0.3615	0.9037
	GPT NEO	0.3087	0.4908
	BART Base	0.3087	0.49
	BART FINE - TUNED	0.38	0.88
	OPT 2.7B	0.3365	0.4908
One - shot	BLOOM 3B	0.3375	0.4908
	BLOOMZ 560M	0.3078	0.6411
	GPT NEO	0.3077	0.4908
	BART	0.3087	0.49
	BART FINE - TUNED	0.38	0.88
	OPT 2.7B	0.3087	0.4943

OBSERVATIONS AND CONCLUSION

Observations & Conclusion

1

Limited capacity and training data: Smaller models do not have the extensive capacity and training data of larger models like GPT-3, which hinders their ability to **generalize effectively in zero-shot learning tasks.**

2

Fine-tuning and instruction tuning: Models like **BART** and **BLOOMZ** provide better zero-shot performance by adapting the model to specific tasks or enabling better understanding of instructions.

Conclusion

3

Contextual understanding: Zero-shot, few-shot, and one-shot learning might produce similar performance if the **model cannot effectively leverage the context provided,** which is more prevalent in smaller models with limited contextual understanding.

4

Compute resource limitations: Smaller models might not be as effective as zero-shot learners in scenarios with limited computational resources. **They may require fine-tuning to achieve satisfactory performance on specific tasks.**

Conclusion

5

Model architecture and capacity: The architecture of a model may not play a significant role in zero-shot learning. Instead, the capacity of the model, **the quality of training data, and fine-tuning or instruction tuning strategies are more critical.**

6

Parameter increase and performance: A slight **increase in model parameters** might not result in a significant improvement in zero-shot performance. The relationship between model size, performance, and computational resources is complex and may require further analysis to determine optimal trade-offs.

Thank You!

Feel Free to ask any
Questions