Common Sense Reasoning

Evaluating zero-shot and fine-tuned models on BIG-Bench

By Visalakshi Iyer — Submitted to INFO 555: Applied NLP
visalakshiiyer@arizona.edu

Abstract

This project investigates the ability of language models to perform common sense reasoning tasks using the BIG-Bench data set [1](bench authors2023), a comprehensive benchmark comprising over 200 diverse tasks designed to evaluate the generalization capabilities of large language models beyond traditional benchmarks. Among these tasks, we focus on three representative challenges: **Riddle-Sense**, **Causal-Judgement**, and **Odd-One-Out**, which test models' ability to reason about riddles, causality, and category mismatches, respectively. To further explore model adaptability and improvement potential, we incorporate the CommonsenseQA dataset [2] (Talmor et al.2018), a multiple-choice question-answering benchmark focused on social, logical, and common sense reasoning. This dataset complements the BIG-Bench tasks, providing a fine-tuning ground to examine whether targeted training can enhance performance on related tasks.

This project is uploaded on:
GitHub Link:
https://github.com/visalakshi2001/applied-nlp-project2 [3]

1 Introduction

Making inference through common-sense requires interpreting questions and prompts with real-world knowledge. In most cases of common sense reasoning, the language model (LM) is given with little to no context document. In this project, we have two datasets, one to benchmark, and one to pre-train a language model, then further assess their ability on common sense tasks. The Beyond the Imitation Game Benchmark (BIG-bench) is a comprehensive evaluation suite designed to assess and extrapolate the capabilities of large language models. It encompasses over 200 diverse tasks contributed by a global community of researchers, aiming to probe areas such as linguistics, common sense reasoning, and logical deduction. Our chosen tasks, riddle-sense (multiple choice riddles that tests model capacity on solving puzzles based on text descriptions), causal-judgement (multiple choice questions that contain a narrative and a yes-no judgment questioning the cause of the result), and odd-one-out (prompts the model to select the odd one out of the choices of given objects, based on the model's inherent understanding of the object's properties) provide question prompts that can evaluate the ability of a language model on reasoning in an easily evaluable metric with the use of question-answering style.

The purpose of this project was to analyse how user prompt influences the output of language models in zero-shot prompting, and whether pre-training a language model on a similar dataset makes any difference (anticipated improvement) in the model performance (accuracy).

The models used in this experiment are DeBERTa-v3-base (He et al.2021), GPT-3.5, GPT-4o-mini (OpenAI2023), and CommandR (from Cohere c4ai models)(Cohere For AI2024). And after evaluation the performance of the models on zero-shot responses, DeBERTa-v3-base is further fine-tuned on CommonSenseQA dataset with over **9741** examples of training dataset, evaluated on **1221** sample of validation examples. The accu-

[1]Beyond the Imitation Game:. Quantifying and extrapolating the capabilities of language models

[2]CommonsenseQA: A Question Answering Challenge

[3]GitHub: https://github.com/visalakshi2001/applied-nlp-project2

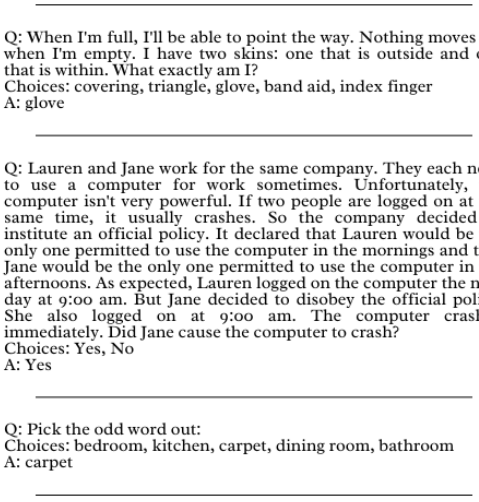

Figure 1: Samples from bigbench dataset (top: riddle-sense, middle: causal judgement, bottom: odd-one-out)

racy of **77.7%** was achieved on the validation set.

The results of these techniques are discussed on the proceeding sections of this report.

2 Current Benchmark on the Dataset

Out of all the reported scores on Papers-with-Code and the official GitHub repository of BIG-bench dataset, current benchmark for the riddle-sense, causal-judgement, odd-one-out tasks are as follows: [4]

Task	Model	Accuracy
RiddleSense	GPT3-200B	44.8%
CausalJudgement	PaLM-535B	54.7%
OddOneOut	PaLM-64B	45.3%

Table 1: Model Accuracy Across Tasks

The total questions in riddle-sense are **49** questions, for odd-one-out it is **86** questions and for causal-judgement there are **188** questions. There are several other models benchmarked on few-shot examples, but we will consider the benchmark relating to zero-shot examples. All the scores depicted above are accuracies achieved by the highest performing model in 0-shot prompt examples.

[4]List of all benchmark tasks, with results in results folder: github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks

3 Methodology

3.1 Zero-Shot Prompting

Zero-shot prompting is a technique where a language model generates responses to a given task without any prior task-specific fine-tuning. In this project, models including DeBERTa-v3-base, GPT-3.5, GPT-4o-mini, and CommandR were evaluated on three tasks from the BIG-Bench dataset.

Input questions and answer choices were formatted into a standardized template to prompt the models effectively. For example, questions were paired with corresponding choices using predefined patterns that clarify the task (e.g., "Question: [Q] Choice: [C]"). For each task, responses were generated from the selected models using appropriate APIs or frameworks, such as Hugging Face, Cohere and OpenAI. The responses were compared to ground truth answers, and model accuracy was calculated. This served as the primary performance metric.

Model_Eval	CJ	OOO	RIDDLE
DeBERTa	52.6%	26.7%	22.5%
Correct	99	23	11
GPT-3.5	58.5%	66.8%	69.4%
Correct	110	60	34
GPT-4o	66.5%	84.9%	81.6%
Correct	125	73	40
CommandR	54.3%	61.6%	55.1%
Correct	102	53	27
Total	188	86	49

Table 2: Results of Zero-Shot Prompting

Among the models tested, GPT-4o demonstrated the strongest performance across all three tasks, achieving 66.5% accuracy in CausalJudgement, 84.9% in OddOneOut, and 81.6% in RiddleSense. These results significantly exceed the benchmark values for all tasks, indicating that GPT-4o's advanced architecture enables robust generalization and reasoning capabilities even without fine-tuning. On the other hand, DeBERTa, a model designed for high contextual understanding, fell short of the benchmark on all tasks. It achieved accuracies of 52.6% in CausalJudgement, 26.7% in OddOneOut, and 22.5% in RiddleSense. While these results showcase DeBERTa's

limitations in zero-shot reasoning scenarios, they provide a baseline for measuring improvements after fine-tuning.

3.2 Fine-Tuning on CommonSenseQA

To explore whether fine-tuning can enhance model performance on related reasoning tasks, the DeBERTa-v3-base model was fine-tuned on the CommonsenseQA dataset. This dataset is a multiple-choice question-answering dataset designed to test common sense, social reasoning, and logical inference. Its questions align well with the reasoning patterns evaluated in BIG-Bench tasks, making it an ideal dataset for this fine-tuning exercise.

1. The dataset was preprocessed to match the model's input requirements. Each question and its corresponding choices were tokenized using the DeBERTa tokenizer, with padding and truncation applied to maintain consistent input lengths. Labels for correct answers were encoded numerically.
2. Fine-tuning was conducted using Hugging Face's Trainer API, for **3 Epochs** and validation performance was monitored at the end of each epoch.
3. The model achieved accuracy of **77.72%** on the validation set, yet it seems that the model was underfitting the data. Since the training was computationally expensive, the training was stopped after 3 epochs and model was retrieved with a goof training accuracy, to use its weights on testing the benchmark.

To track the training process, Weights-and-Biases API(Biewald2020) [5] was used, the results from the training steps are as follows:

4 Evaluation

After fine-tuning, the DeBERTa-v3-base model was re-evaluated on the BIG-Bench tasks. The objective was to measure performance improvements and compare the fine-tuned model's results against its zero-shot baseline. The results of fine-tuning are as follows:

The performance of the DeBERTa model highlights the effectiveness of enhancing the weights through fine-tuning (pre-training) on the given datasets.

[5]Weights & Biases Experiment Tracking

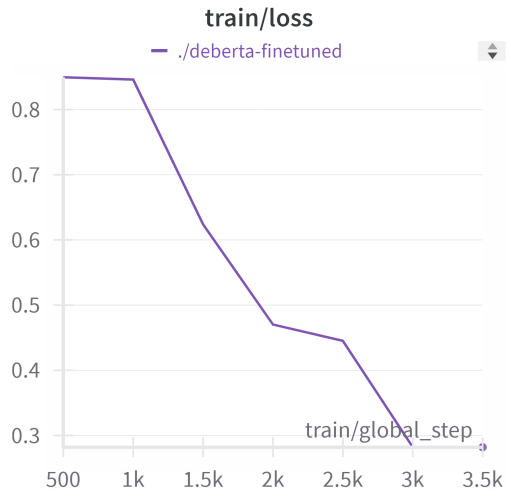

Figure 2: Training Loss: evaluated on 9741 samples)

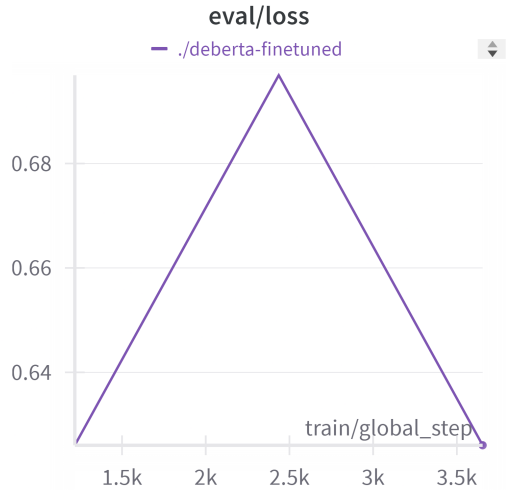

Figure 3: Validation Loss: evaluated on 1221 samples

Model_Eval	CJ	OOO	RIDDLE
DeBERTa (Baseline)	52.6%	26.7%	22.5%
Correct	99	23	11
DeBERTa (FineTuned)	54.8%	32.6%	38.8%
Correct	103	28	19
Total	188	86	49

Table 3: Results of Pre-trained weights of CommonSenseQA

The training process went for over three epochs—a process and took approximately three hours to complete. The model's performance improved across all tasks. These improvements, though modest in CausalJudgement, are substantial in OddOneOut and RiddleSense, where the model struggled the most before fine-tuning. This suggests that fine-tuning helped the model better align its knowledge with the reasoning patterns required for these tasks.

5 Conclusion

This project sought to explore and enhance the common sense reasoning capabilities of language models using tasks from the BIG-Bench dataset and fine-tuning on the CommonsenseQA dataset. The evaluation focused on three challenging tasks: Riddle-Sense, Causal-Judgement, and Odd-One-Out, assessing models in both zero-shot and fine-tuned settings.

The zero-shot evaluation revealed significant differences in performance among the models. GPT-4o-mini emerged as the most capable model, outperforming its counterparts in all three tasks. On the other hand, DeBERTa-v3-base struggled in the zero-shot setting.

Fine-tuning DeBERTa on the CommonsenseQA dataset demonstrated the potential for performance improvements through targeted training. After fine-tuning, DeBERTa showed notable gains, particularly in tasks where it initially struggled. For instance, its accuracy improved to 54.8% in Causal-Judgement, 32.6% in Odd-One-Out, and 38.8% in Riddle-Sense. These improvements, while modest, are significant given the inherent difficulty of the tasks and the zero-shot baseline performance. Overall, the study highlights the potential of fine-tuning to align pre-trained models more effectively with specific reasoning tasks.

6 Limitations and Future Directions

Fine-tuning and evaluation, particularly for large-scale models, were computationally intensive. Due to resource constraints, fine-tuning on CommonsenseQA was limited to three epochs, potentially restricting the model's ability to fully adapt to the dataset. The study evaluated only three tasks from the BIG-Bench dataset. Expanding the analysis to include additional tasks could provide a more comprehensive understanding of model capabilities.

References

2023“protectbench authors, 2023 BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

2020“protectBiewald, 2020 Lukas Biewald. 2020. Experiment tracking with weights and biases. *Software available from wandb.com*, 2:233.

2024“protectCohere For AI, 2024 Cohere For AI. 2024. Command-r: A large language model with open weights optimized for reasoning, summarization, and question answering. `https://huggingface.co/CohereForAI/c4ai-command-r-v01`.

2021“protectHe et al., 2021 Pengcheng He, Xiaodong Liu, Jianfeng Wang, and Jianfeng Gao. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. `https://huggingface.co/microsoft/deberta-v3-base`.

2023“protectOpenAI, 2023 OpenAI. 2023. Gpt-3.5 and gpt4o. `https://openai.com/api/`.

2018“protectTalmor et al., 2018 Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.