**CS\_685 NLP Course Project**

***Breaking Bias: An Approach to Quantify and Reduce Bias of LLMs In Multiple Choice Answering***

**Project Overview**

This project aims to address the issue of intrinsic bias in Large Language Models (LLMs) when answering Multiple Choice Questions (MCQs). Specifically, we focus on mitigating bias towards certain option tokens (e.g., consistently favoring option "A") and position tokens (e.g., preferring options presented earlier in the question).

This repository provides the code and resources used for our research, allowing for reproducibility and further exploration of LLM de-biasing techniques in the context of MCQ answering.

**Documentation:**

1. **mmlu01:** Contains train and evaluation sets tailored for different experiments.

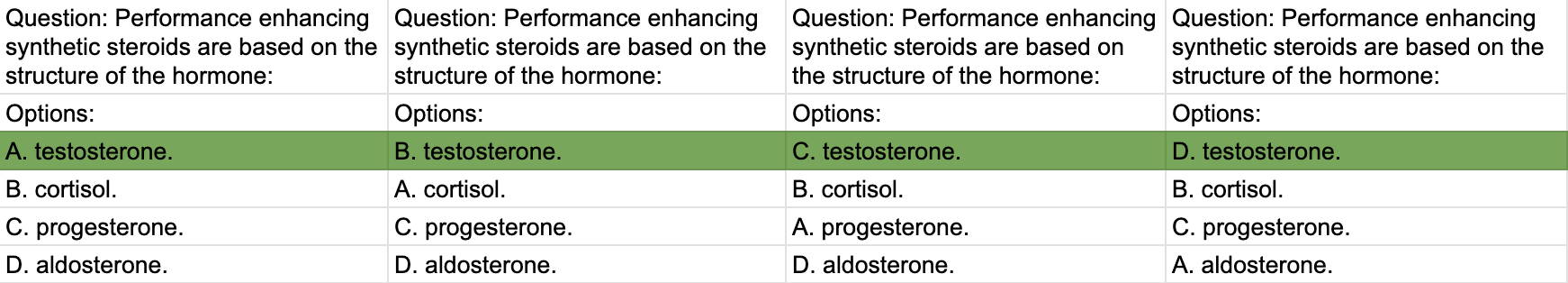
* **Domain-specific datasets(ID / In Domain) :** Focusing on specific subject areas (e.g., science, history, etc.)
* **Out-of-domain datasets( OOD / Out of Domain):** Evaluating generalization capabilities on unseen subjects.

**Datasets for training :**

* **trainset:** The dataset used for training the models in experiment1.
* **valset:** A validation dataset used to evaluate the model's performance during training, helping to avoid overfitting.
* **permuted\_train\_set\_8k, permuted\_val\_set\_8k:** Training and validation sets where the order of options and their positions for each question have been shuffled. These sets contain 500 questions with 16 permutations
* **permuted\_train\_set\_16k, permuted\_val\_set\_16k:** Similar to the 8k sets, but with 16,000 questions
* **permuted\_train\_set\_32k, permuted\_val\_set\_32k:** Again, similar to the previous sets but containing 32,000 questions.

**mmlu01 / varying\_option:**

Test Dataset created by changing ground\_truth answer's option token to A,B,C,D (4 permutations to a question)



**Out-of-Domain Subjects:**

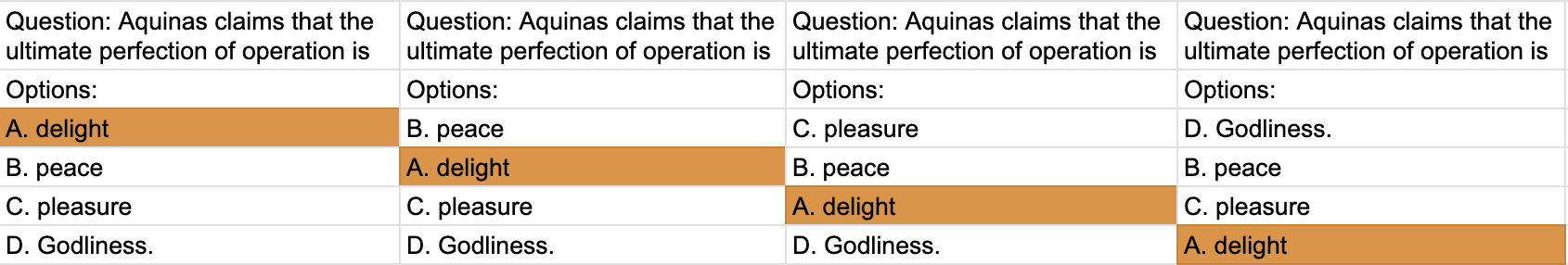
* professional\_law
* prehistory
* philosophy
* high\_school\_mathematics
* conceptual\_physics
* college\_medicine
* abstract\_algebra

**In Domain Subjects:**

* testset : Dataset used for testing in experiment1
* permuted\_testset\_8k,permuted\_testset\_16k,permuted\_testset\_32k: Permuted testset used for experiments on 8k,16k,32k sets

**Similar structure in mmlu01/varying\_position:**

Test Dataset created by changing ground\_truth answer's position (4 permutations to a question)



1. **mmlu02:**

**Contains inferences generated by following models on test datasets of mmul01 varying\_option and varying\_position**

**Baseline Models used for Bias Analysis:** llama-7b, llama-2-7b,llama-2-13b,llama-3-8b,mistral-7b,mistral-7b-instruct,vicuna-7b,vicuna-13b,gemma-2b,gemma-2b-it,gemma-7b-it,gemma-7b

**Models of experiments1 :** llama-2-7b-lora-tuned , llama-7b-lora-tuned , gemma-2b-it-lora-tuned , gemma-7b-it-lora-tuned

**Models of experiments2 :** llama-2-7b-lora-tuned-per-8k , llama-7b-lora-tuned-per-8k , gemma-2b-it-lora-tuned-per-8k

**Models of experiments3 :** llama-2-7b-lora-tuned-per-16k , llama-7b-lora-tuned-per-16k , gemma-2b-it-lora-tuned-per-16k

**Models of experiments4 :** llama-2-7b-lora-tuned-per-32k

**other experiments models :** llama-7b-1shot-cot

1. **mmlu03:**
   1. Contains Evaluation results on inferences in mmlu02 generated by models.
   2. Contains Comparative analysis plots of evaluation results

* Evaluation metrics were calculated on inferences generated by a each model and stored results csv files in "mmul03/model-name/" folder

**(error\_bins, ppa\_scores , recall\_imbalance , incorrect\_likelihood for both in-domain and out-domain test datasets)**

1. **out\_dir : (Trained models )**

This folder contains trained models . These models can be loaded for further usage.

1. **logs:** **(Training logs)**

This folder stores the logs generated while training models .

1. **prompt\_tuning :**

This folder has scripts related to prompt tuning

1. **COT : (Chain-of-thought prompting)**

Contains scripts related to chain-of-thought prompting experiments

**8) plots:**

Contains scripts for plotting evaluation metrics results .

**Scripts:**

1. **Evaluation.ipynb**

1 ) Script logic to generate evaluation results and stores it in corresponding .csv files

1. **Inference.ipynb**

1) Script to generate inference results and stores it in corresponding .csv files

2) We can load pretrained model from SDK of unsloth or huggingface

3) We can load trained models from out\_dir

(Corresponding code is commented in notebook)

1. **Create datasets.ipynb**
2. Script to create datasets of mmlu01

**d) Error\_analysis .ipynb**

1. Script to required to do error analysis for report   
    [ prompts where baselines and our trained models are failing are stored in mmlu/<model-name>/varying\_option/error\_analysis\_cases

mmlu/<model-name>/varying\_position/error\_analysis\_cases ]

**e) lora\_tuning.ipynb**

1. Script to train models using loRA methods

**Getting Started**

1. **Download the code repository**

1. **Install dependencies:**
   1. Create a conda virtual environment
      1. conda create -n myenv
      2. conda activate myenv
   2. pip install -r requirements.txt
   3. Ensure to have good compute resources like NVIDIA L4 GPU and 32GB ram to train the models

**Experiment Setup:**

You'll likely use these datasets to compare model performance under different training conditions:

**Tuning without permutations :** Train on the original trainset, evaluate on valset in mmlu01.

**Tuning with permutations experiments:** Train on the permuted training sets (8k, 16k, 32k) and evaluate on their corresponding permuted validation sets in mmul01.

**3) LoRA tuning , fine tuning an LLM using a specific configuration:**

Run the lora\_tuning.ipynb with correct paths to training and validation sets

**4) Evaluate the LLM’s Bias using ID and OOD datasets:**

Run the evaluation.ipynb

**Contact**

For any questions or inquiries, please contact [rpulishetty@umass.edu](mailto:rpulishetty@umass.edu), [akannapu@umass.edu](mailto:akannapu@umass.edu), [sphurthiagarw@umass.edu](mailto:sphurthiagarw@umass.edu),[rssharma@umass.edu](mailto:rssharma@umass.edu),[bponnam@umass.edu](mailto:bponnam@umass.edu) .

**Acknowledgement**

We acknowledge the creators and contributors of the MMLU benchmark datasets and the various LLM architectures used in this project.