Day-1

- ▼ Massive Multitask Language Understanding
  - It is a dataset used to evaluate the model on a zero or few-shot setting similar to human-like evaluation. This dataset has 57 subjects such as STEM, humanities, and social sciences.
    - **It ranges in difficulty from an elementary level to an advanced professional level, and it tests both world knowledge and problem-solving ability. Subjects range from traditional areas, such as mathematics and history, to more specialized areas like law and ethics. The granularity and breadth of the subjects make the benchmark ideal for identifying a model's blind spots.**
  - ▼ STEPS
    - ▼ DATA LOADING

      the dataset is stored in JSON or CSV format.
    - ▼ Preprocessing
      - firstly the data is read and sent through normalization where all the NLP processing takes place
      - tokenization occurs which creates embeddings
    - ▼ Model Inference
      - The model is fed the questions one by one or in batches(for efficiency) and in turn it outputs an answer (prediction) on what it believes to be correct
    - ▼ Evaluation
      - The models predictions are compared with a ground answer list using an evaluation script and then accuracy or such metrics are generated.
    - ▼ Reporting

    - The scores which are the results of evaluation are reported which are the benchmarks for the said model

- ▼ Analysis
    - Researchers analyze the benchmarks to decide wether the model excelled or suffered in this evaluation and this is used to improve the performance of the model based on the MMLU tasks.