THESIS DATA STORAGE STRATEGY

Google Drive

- Raw data dumps (submissions and comments)
- Input data files from filtering raw dumps
- Initial dataset
- Files stored as .zst, can use zst to csv script to uncompress them

Large data files

Github Repository

Dataset Creation

- Script to filter raw dumps
- Notebook to create initial dataset
- Notebook to process initial dataset
- Notebook to create llama2 instruction datasets
- Directory for data set processing output (charts, statistics)
- Notebook to create "high-quality" dataset
- Notebook to create llama2 instructions "high quality" dataset

Fine-Tuning

- Flan-T5 fine-tuning notebook (colab)
- Llama-2 fine-tuning notebook (colab)
-py versions?

Evaluation

- Flan-T5 fine-tuning notebook (colab)
- Llama-2 fine-tuning notebook (colab)
-py versions?

HuggingFace Hub

- 4 datasets (main, llama-2-main, high-quality, llama-2-high-quality)
- 4 models: Flan-T5 xl-main, llama2-7b-chat-main, Flan-T5-xl-high-quality, llama2-7b-chat-high-quality
- Store llama2 models as complete models, store flan-t5 models as peft adapters

Models and tokenizers

Paper content

Overleaf

- LaTeX paper
- Figures, Charts, Tables