Class Structure

1. Data / NLP tasks overview
2. task brainstorming session
3. description of selected tasks
4. assign people to groups by task
5. team work (data)
6. discussion
7. evaluation overview
8. team work (evaluation)
9. discussion

Data / NLP Tasks overview

- How much data do we need for training an LLM?
    - The more the better.
    - Short answer: Trillions of words (https://github.com/togethercomputer/RedPajama-Data)
        - this is 10^12
    - https://arxiv.org/pdf/2402.18041.pdf has 181-page survey (incl. 100 pages Appendices+References)
- Several dataset types:
    - Data for pre-training (often a mixture of the following)
        - Webpages, News, Books, Academic materials, Code (!), Parallel corpora, Social media, Encyclopedias,
        - (there are also domain-specific datasets, medical, financial, legal, …)
    - Instruction fine-tuning
        - instruction datasets (Figure 10 in the survey shows various types)
        - constructed by humans or computers, either genuine or artificial
        - again, there is a large variety in terms of the domain (IT, Code, Math, Education, ..)
    - Preference datasets (RLHF)
        - Figure 16 show different approaches for voting/recording preferences
        - Voting, Sorting, Scoring (some can be done by both humans and other models)
    - Evaluation datasets
        - Specialized datasets to test various aspects of model behavior (Fig 18)
    - NLP (task-specific) datasets
    - Data preprocessing:
        - Language ID, filtering, deduplication

Task brainstorming session

Before concentrating on specific tasks, we had a quick brainstorming session about tasks we can use - and evaluate - LLMs on.

Description of selected tasks

We split to 6 groups and each group was discussing one of the following tasks:

1. Machine translation
2. Question answering
3. Summarization
4. Sentiment analysis
5. Hate speech detection
6. Code generation

Assign people to groups

take the right-most digit of your university ID which is between 1 and 6 inclusive

Team work - data

Each group needed to:

1. Find data for their task
2. Gather data details (size, quality, usage)

Evaluation overview

How to evaluate?

- Automatically, using a (deterministic) algorithm, aka “metric”
    - BLEU, ROUGE, F1-score, exact match, accuracy, …
    - Very cheap, and simple.
    - String-based metrics may not correlate with human judgments
- Manually by humans
    - Expensive, not straightforward to define well
    - If successful, provides very high-quality estimates of model performance
- Using a different model
    - LLMs used as judges (typical in chat/dialogs, used in machine translation as well)
    - Not interpretable, domain-specific, more expensive than automatic metrics, but way less expensive than human annotation.
    - When set up properly, correlates well with human judgment

Which method to choose?

- This very much depends on the task - it could be very simple (anything where we can measure accurracy) or open-ended (many different correct answers)

Team work - evaluation

Same groups, worked out how would they evaluate their model:

- Find a test set
- Find an evaluation metric
- Find what a “reasonable” score could look like