**Task Description:** The student will benchmark different response generation models. This includes data processing, implementing the response generation pipeline for a few generative models (not all those included on the list, but as many as possible in the research project period), and evaluate the quality of the response using a set of automatic metrics.

**Background:** Neural response generation is a subcategory of text-generation that shares the objective of generating natural-looking text (distinct from any training instance) that is relevant to the prompt.

**1. Data Preparation:** Clean and process the data into the adequate format for training and testing.

**2. Model Benchmark:** Look into and compare the performances on the task of response generation of pre-trained generative models such as: dialoGPT, GPT2, Bart, t5, Blenderbot, xlnet...

**3. Quantitative evaluation:** Evaluate the generated responses using a set of automatic metrics like BLEU, ROUGE, bertscore, chrf… I recommend using huggingface’s metrics library https://huggingface.co/evaluate-metric

**4. Qualitative evaluation:** Compare the quality of the generated content on a grammatical and syntactic level, highlight the differences between the results on the different datasets…

**21/04/2023**

**Agenda**

* daily\_dialog dataset
* Task 🡪 specifications
* Any paper I should read to have more background?
  + DialoGPT: <https://arxiv.org/pdf/1911.00536v3.pdf>
    - Extends GPT-2
    - state-of-the-art results in both automatic and human evaluation
  + GPT-2 : <https://life-extension.github.io/2020/05/27/GPT%E6%8A%80%E6%9C%AF%E5%88%9D%E6%8E%A2/language-models.pdf>
  + BART
* Any reference benchmark for these datasets?
* Agree on a deadline (potentially flexible?)
  + June 24th article, middle of May or end of May 🡪
* Approach: source-target pairs?
* Score per source-target or per dialog?
  + Commonsense questions
  + Multi-turn dialogue
* One Drive folder (instead of Google Drive)