Analyzing the reasoning and self explaining capabilities of pretrained language models using prompts

- **Supervisor**: Lütfi Kerem Şenel
- **Examiner**: Prof. Schütze
- **BSc, MSc, Open**: M.Sc.
- **General Topic Area**: Language Model Analysis, Interpretability.
- **Prerequisites**: Moderate experience in Python.. Enthusiasm and curiosity to explore new methods. Familiarity with the Huggingface library is a plus.
- **Details**: Interpretability (being able to explain the mechanisms and reasons behind a model's or an algorithm's decisions) is one of the key desired properties in machine learning and artificial intelligence. Several methods such as training probing models, investigating attention weights or computing saliency scores have been proposed to investigate model behavior. As models became more powerful and capable in understanding language, they started to be used more and more in a text-to-text format where the target task is formulated using natural language and the model is asked to fill-in-the-blank or generate a continuation to the input prompt. Several studies used language modeling objective to investigate the knowledge stored in a model.

  The aim of this project is to investigate various language models' ability to justify their decisions by generating plausible reasoning alongside their decisions. The investigation can be performed at three different levels: i) models (i.e., investigate different PLMs' responses), ii) tasks (i.e., investigate a model's response on various downstream tasks/datasets), iii) prompts (i.e., investigate a model's response on a dataset using different prompts). At the first stage the investigations can be performed by manual investigations, searching for automated evaluations can be considered at later stages. Investigating the relationship between generalization and reasoning can be a possible extension. Another possible extension is investigating the effect of finetuning PLMs for explicit reasoning.
- **Project Takeaways**: Hands-on experience with state-of-the-art pretrained Language Models; gaining insights about the strengths and weaknesses of these models.