This talk reports the results from two studies on testing how good AI is at Natural Language Understanding and Inferencing (NLU and NLI), which belong to crucial competencies of human communication. The first study tests the NLU and NLI of machines and humans in a semantic task of comparing the semantic similarity of sentences that address 20 questions about the recent covid-related pandemic crisis. We collected answers from 300 participants and asked four human annotators to annotate whether the answers semantically matched "expert" answers. Expert answers are defined as good answers suggested by medical experts. We used the embedding-based method S-BERT (Reimers & Gurevych 2020), a variant of BERT (Devlin et al. 2019), tailored to predict sentence similarity of sentences, in order to measure the semantic similarity of participants' and experts' answers. We then compared embedding-based semantic similarity measures with human annotations. The results from the first study show that the accuracy of the automatic method we tested shows significantly lower results than human annotations. We then used ChatGPT prompts for a similar semantic task that showed better accuracy results.

Discussion: Our question on how good machines at NLU and NLI are can be answered as follows. It depends on the machines. SBERT did not provide good accuracy results in comparison to humans in NLU and NLI tasks, but ChatGPT did. More research is needed in methods comparison in NLU and NLI. Our results of the first experiment are surprising given that BERT-based models achieve usually a good performance in NLI tasks if they are trained on NLI datasets (see Stanford NLI datasets and method comparison (https://nlp.stanford.edu/projects/snli/). We believe that this is due to the missing NLI datasets in the training of SBERT. We will discuss the importance of the datasets used for training the machines for the NLI and NLU tasks.