A Details for Datasets

ReClor ReClor contains 6,138 multiple-choice questions modified from standardized tests such as GMAT and LSAT, which are randomly split into train/dev/test sets with 4,638/500/1,000 samples respectively. It contains multiple logical reasoning types. The held-out test set is further divided into EASY and HARD subsets based on the performance of BERT-based model [**?**].

LogiQA LogiQA consists of 8,678 multiple-choice questions collected from National Civil Servants Examinations of China and are manually translated into English by experts. The dataset is randomly split into train/dev/test sets with 7,376/651/651 samples correspondingly. LogiQA also contains various logical reasoning types.

MuTual MuTual has 8,860 dialogues annotated by linguist experts and high-quality annotators from Chinese high school English listening comprehension test data. It is randomly split into train/dev/test sets with 7,088/886/886 samples respectively. There more than 6 types of reasoning abilities reflected in MuTual. MuTual^{plus} is an advanced version, where one of the candidate responses is replaced by a safe response (e.g., *"could you repeat that?"*) for each example.