Table 1: UnifiedQA datasets with a description of each and the number of samples in train, dev, and test splits [6].

Name	Description	Train samples	Dev samples	Test samples
narrativeqa	NarrativeQA is an English-language dataset of stories and corresponding questions designed to test reading comprehension, especially on long documents.	65494	6922	21114
ai2_science_middle	The AI2 Science Questions dataset consists of questions used in student assessments in the United States across elementary and middle school grade levels. Each question is 4-way multiple choice format and may or may not include a diagram element. This set consists of questions used for middle school grade levels.	605	125	679
ai2_science_elementary	Same as above. This set consists of questions used for elementary school grade levels.	623	123	542
arc_hard	This dataset consists of genuine grade-school level, multiple-choice science questions, assembled to encourage research in advanced question-answering. The dataset is partitioned into a Challenge Set and an Easy Set, where the former contains only questions answered incorrectly by both a retrieval-based algorithm and a word co-occurrence algorithm. This set consists of "hard" questions.	1119	299	1172
arc_easy	Same as above. This set consists of "easy" questions.	2251	570	2376
mctest_corrected_the_separator	MCTest requires machines to answer multiple-choice reading comprehension questions about fictional stories. The stories and questions are limited to those a young child would understand, reducing the world knowledge that is required for the task.	1480	320	0
squad1_1	This is a reading comprehension dataset consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to each question is a segment of text from the corresponding reading passage.	87599	10570	0
squad2	This dataset combines the original Stanford Question Answering Dataset (SQuAD) dataset with unanswerable questions written adversarially by crowdworkers to look similar to answerable ones.	130319	11873	0
boolq	BoolQ is a question answering dataset for yes/no questions. These questions are naturally occurring; they are generated in unprompted and unconstrained settings. Each example is a triplet of (question, passage, answer), with the title of the page as optional additional context.	9427	3270	0
race_string	Race is a large-scale reading comprehension dataset. The dataset is collected from English examinations in China, which are designed for middle school and high school students.	87866	4887	4934
openbookqa	OpenBookQA contains questions that require multi-step reasoning, use of additional common and commonsense knowledge, and rich text comprehension. OpenBookQA is a new kind of question-answering dataset modeled after open book exams for assessing human understanding of a subject.	4957	500	500