A Appendices

A.1 Datasets: Details

We evaluate our UNIFIEDQA on 19 existing datasets that target various formats, as well as various complex linguistic phenomena. Table 2 shows different properties for our datasets (whether it comes with a paragraph, whether the paragraph explicitly contains the answer, whether there are candidate-answers as part of the input, etc.) Most importantly, they are grouped into several formats/categories described below. Table 2 gives summary statistics of these datasets.

Extractive QA (EX). All the datasets in this format require models to extract the answer to a given question as a substring from a context paragraph. SQuAD 1.1 (Rajpurkar et al., 2016) contains questions about Wikipedia paragraphs. A later version of this dataset, SQuAD 2 (Rajpurkar et al., 2018), includes unanswerable questions which empirically makes the task much harder. For our evaluation, we use the development sets of SQuAD 1.1 and SQuAD 2. NewsQA (Trischler et al., 2017) dataset focuses on paraphrased questions with predicate-argument structure understanding collected from news articles from CNN/DailyMail articles. Quoref (Dasigi et al., 2019) contains questions that require coreference resolution in Wikipedia articles and can even have disjoint spans as answers. ROPES (Lin et al., 2019) centers around situation understanding, where the model must under the causes and effects implicit in the given situation.

Abstractive QA (AB). All the datasets in this format require models to produce answers that are often not mere substrings of the given context paragraph. NarrativeQA (Kociský et al., 2018) focuses on understanding various events that happen in a given movie plot, based on summaries of their movie adaptations from various web resources. Many of the answers do not have high overlap with the context. DROP (Dua et al., 2019b) contains questions that involve rudimentary mathematical skills (such as counting, addition, subtraction, maximum, minimum, etc.) and questions query multiple parts of the paragraph. The answer can be either a number or a date that can be inferred from the paragraph, or several spans from the context paragraph. Finally, we use an open-domain version of NaturalQuestions (Kwiatkowski et al., 2019) where the paragraph that was used for creating the question is eliminated, and only the questions with short answers up to five tokens are taken. Instead, we follow (Min et al., 2020) to use a DPR retrieval (Karpukhin et al., 2020) engine to augment each question with an additional context paragraph. We call this dataset NatQA.

Multiple-choice QA (MC). All the datasets in this format contain questions that come with candidate answers. MCTest (Richardson et al., 2013) contains questions about simple, fictional stories. RACE (Lai et al., 2017) is a challenging set of English comprehension multiple choice exams given in Chinese middle and high schools. OpenBookQA (Mihaylov et al., 2018), ARC (Clark et al., 2018, 2016), QASC (Khot et al., 2019) are different MC tests focusing on elementary/high school-style science exams. We use several othern datasets that are often framed as commonsense reasoning benchmarks: CommonsenseQA (Talmor et al., 2019) is geared towards activity/concept questions, PIQA (Bisk et al., 2020) addresses physical interaction reasoning, SIQA (Sap et al., 2019) contains question that require social reasoning (motivations, reactions, event orders) and finally Winogrande (Sakaguchi et al., 2020) which a benchmark for hard pronoun resolution problems (Levesque et al., 2011; Peng et al., 2015).

Other than MCTest and RACE, the rest of the datasets do not come with accompanying paragraphs. On such datasets, occasionally a retrieval system is used to supplement each question with a relevant retrieved context paragraph. For most of this the work, we keep the questions as is with no additional retrieval (unless otherwise mentioned), except in §6.3 where we use IR to get numbers comparable to earlier work. One other variability among these datasets is their number of candidate answers. While many datasets have four candidates (see Figure 2), others have more. Later, in §6.2 we will see that our approach generalizes to datasets with different number of candidates, even if it's not seen during training.

Yes/No QA (YN). All the datasets in this format contain questions that could be responded with yes/no answers. One can think of these as multiple-choice questions with 2 candidates; however, they're usually treated differently. Several examples we use are BoolQ (Clark et al., 2019a) and a version of this dataset

with natural-perturbations, BoolQ-NP (Khashabi et al., 2020), the subset of MultiRC (Khashabi et al., 2018) that have binary(yes/no) answers.

Contrast-sets. Additionally, we use *contrast-sets* (Gardner et al., 2020) for several of our datasets (denoted with "CS"): BoolQ-CS, ROPES-CS, Quoref-CS, DROP-CS. These evaluation sets are expert-generated perturbations that deviate from the patterns common in the original dataset.

A.2 Details on the experiments:

Below is several details on the experiments:

- Models: we use two text-to-text frameworks: T5 and BART.
- Model sizes: Most of the experiments are done on T5(11B) which has 11 billion parameters. We also report experiments with BART (large) with 440 million parameters.
- Input/output size: For all experiments, we use token-limits of size 512 and 100 for inputs and outputs sequences, respectively.
- # of iterations for pretraining on the seed datasets (§3): All models are trained for $100k$ steps on the seed datasets.
- Learning rates: we use 1e-3 and 1e-5, for T5 and BART, following the original works on each framework.
- Batch sizes: We use batches of 8 and 120, for the T5 (11B) and BART models, respectively.
- Infrastructure: In the experiments, we use v3-8 TPUs for T5 models, and eight 32GB GPUs for BART models.
- Time spent to build UnifiedQA: pretraining UnifiedQA approximately takes about 36 and 55 hours, on T5(11B) and BART models, respectively.
- Finetuning on datasets (§6.3): the only hyperparameter we iterated over is the training steps. Each model was fine-tuned for $60k$ steps and checkpoints were saved every $2k$ steps. The model with the highest score on the dev set is our selected model.

A.3 UNIFIEDQA: Different Sizes

For completeness we're also showing the scores of UNIFIEDQA of different sizes on each dataset. For these systems each row is a single system.

Trained on ↓ - Evaluated on →	SQuAD11	SQuAD2	NewsQA	Quoref	Quoref-CS	ROPES	ROPES-CS	NarQA	DROP	DROP-CS	BoolQ	MultiRC	NP-BoolQ	BoolQ-CS
Small	79.4	67.6	51.1	25.6	27.6	31.0	32.9	53.7	14.6	17.2	77.1	46.9	59.4	58.1
Base	88.2	78.1	54.2	40.0	38.5	33.9	28.4	58.7	19.7	23.7	82.5	64.8	66.3	61.9
Large	91.1	85.9	48.5	45.5	42.1	47.7	37.9	60.8	24.6	30.7	86.1	54.2	72.6	73.0
3B	93.2	87.4	59.6	60.4	54.7	48.7	43.1	63.3	28.5	33.9	89.3	62.6	78.4	77.0
11B	93.4	89.6	58.9	63.5	55.3	67.0	45.6	65.2	32.5	40.9	90.2	59.9	81.3	80.4

Trained on ↓ - Evaluated on →	RACE	OBQA	OBQA (w/ IR)	ARC-easy	ARC-easy (w/ IR)	ARC-chal	ARC-hard (w/ IR)	MCTest	QASC	QASC (w/ IR)	CQA
Small	56.0	50.4	35.4	42.9	59.5	35.9	35.8	80.0	19.1	37.9	32.8
Base	70.3	59.0	38.4	53.0	69.4	42.4	44.2	86.9	25.8	50.8	45.0
Large	78.1	68.4	54.6	65.9	77.4	54.4	54.8	90.0	43.3	62.6	60.9
3B	83.2	80.8	63.2	78.7	86.2	66.7	64.8	95.0	62.2	76.6	71.3
11B	87.3	86.0	71.2	85.7	89.2	75.6	74.7	95.0	68.5	80.1	76.2

Table 7: UNIFIEDQA of different sizes on our datasets.

A.4 Comparison with the Dedicated Models: extended results

Here we summarize an extension of the results in §6.1. Table 8 summarizes the results of the relevant experiment. In the top portion of the table we have evaluations of T5 model fine-tuned for individual datasets, followed by UNIFIEDQA. As it can be observed from the table, UNIFIEDQA performs almost as good as the best single dataset experts. In some cases UNIFIEDQA performs even better than than the single-dataset experts (e.g., on OBQA or NQA.) On average (last column) UNIFIEDQA is doing much better dataset/format-specific systems. In conclusion, UNIFIEDQA offers flexibility across multiple QA formats while compromising almost nothing compared to dataset-specific experts.

Seen dataset?	Model ↓ - Evaluated on →	NewsQA	Quoref	Quoref-CS	DROP	DROP-CS	ROPES	ROPES-CS	QASC	Commonsen seQA	NP-BoolQ	BoolQ-CS	MultiRC	Avg
No	T5 (SQuAD11)	**62.5**	**71.5**	**61.0**	31.5	37.0	62.0	39.9	64.5	70.4	1.6	0.0	2.4	42.0
	T5 (SQuAD2)	55.7	54.7	46.0	20.3	20.1	29.4	23.9	39.3	52.6	22.2	18.2	9.5	32.7
	T5 (RACE)	49.9	70.7	56.6	29.2	36.5	**72.1**	**48.2**	64.1	73.1	2.5	4.5	3.3	42.6
	T5 (OBQA)	9.3	20.7	14.3	7.7	9.4	20.6	5.4	52.2	67.4	0.2	0.1	0.1	17.3
	T5 (BoolQ)	0.6	1.7	1.4	0.4	0.1	0.0	0.7	14.8	20.8	79.1	78.6	**91.7**	24.2
	T5 (NarQA)	58.0	68.2	57.6	30.7	36.8	48.1	41.7	54.1	59.0	27.2	39.9	28.4	45.8
	UnifiedQA	58.9	63.5	55.3	**32.5**	**40.1**	67.0	45.5	68.5	**76.2**	**81.3**	**80.4**	59.9	**60.7**
Yes	Previous best	66.8 Retro Reader	70.5 XLNet	55.4 XLNet	89.1 ALBERT	54.2 MTMSN	61.1 ROBERTa	32.5 RoBERTa	85.2 KF+SIR+2Step	79.1 FreeLB-RoBERTa	78.4 RoBERTa	71.1 RoBERTa	-- --	

Table 8: UNIFIEDQA is on-par with systems tailored to individual datasets (the diagonal cells vs the last row) while functioning across a wide range of datasets (the last column).

A.5 Pairwise Mixing: extended results

Here we summarize an extension of the results in §5. The question addressed here is whether there is value in mixing datasets with different formats. We evaluated this by adding one dataset of a different format to four different datasets (one for each format). The results are summarized in Table 9. The goal of each sub-table is to measure the *within-format* generalization one can gain via *out-of-format* training. Each sub-table has an *anchor* dataset, indicated in the first column. For example in the first table the anchor dataset is SQuAD. Rows of the table: Each table combines datasets of other formats with the anchor dataset (e.g., SQuAD + RACE, etc). The columns of the sub-tables contain evaluations on the dataset with the same format as the anchor dataset. For example, on the first table, the evaluation is done on SQuAD 1.1/2.0, NewsQA, Quoref which have the same format as SQuaD 1.1, the anchor dataset. The results show that one can achieve gains for question-answering in a certain format by incorporating resources in other formats. In the first two sub-tables, we see that NarQA (AB) and OBQA (MC) help a SQuAD models generalize better to other EX datasets. In the third table where the anchor dataset is NQA

(AB), EX datasets help a NQA model generalize better to other AB datasets. In the 4th/5th subtable, EX and AB datasets help a RACE/OBQA (MC) models generalize better to other MC datasets. Similarly, in the final sub-table, MC dataset helps improve the scores on a YN datasets.

Anchor Dataset / Format	Trained on ↓ - Evaluated on →	SQuAD11	SQuAD2	NewsQA	Quoref	Quoref-CS	Avg
SQuAD11	**SQuAD11**	**85.9**	42.8	51.7	28.2	28.11	47.4
	SQuAD11 + RACE	85.6	42.6	51.7	26.6	27.43	46.8
	SQuAD11 + OBQA	85.7	42.8	**52.1**	27.7	**29.84**	**47.6**
	SQuAD11 + BoolQ	85.8	42.7	52.1	27.7	29.42	47.5
	SQuAD11 + NarQA	85.6	42.7	51.3	**29.4**	26.56	47.1
SQuAD2	**SQuAD2**	76.5	70.7	46.0	17.7	22.04	46.6
	SQuAD2 + RACE	76.5	70.6	47.9	18.6	20.40	**46.8**
	SQuAD2 + OBQA	**76.7**	70.8	**48.4**	16.9	19.80	46.5
	SQuAD2 + BoolQ	75.9	**72.0**	45.4	16.3	20.35	46.0
	SQuAD2 + NarQA	72.5	70.9	47.3	**20.0**	**23.39**	**46.8**

Anchor Dataset / Format	Trained on ↓ - Evaluated on →	NarQA	DROP	DROP-CS	ROPES	ROPES-CS	Avg
NQA	**NarQA**	51.5	10.2	11.1	22.8	15.3	22.2
	NarQA + SQuAD11	52.7	14.1	14.6	30.5	33.2	29.0
	NarQA + SQuAD2	**53.0**	**14.4**	**14.6**	**31.3**	**33.2**	**29.3**
	NarQA + NewsQA	52.5	10.4	12.3	16.6	15.6	21.5
	NarQA + RACE	52.0	10.7	13.5	20.0	17.9	22.8
	NarQA + OBQA	51.8	10.1	11.3	15.4	17.0	21.1
	NarQA + BoolQ	51.8	10.2	10.9	20.7	10.9	20.9

Anchor Dataset / Format	Trained on ↓ - Evaluated on →	RACE	OBQA	ARC-easy	ARC-hard	MCTest	QASC	CQA	Avg
RACE	**RACE**	55.8	26.6	31.8	28.0	62.5	17.9	28.3	35.8
	RACE + SQuAD11	**59.1**	28.0	**32.4**	28.1	**69.4**	**23.5**	**36.1**	**39.5**
	RACE + NewsQA	57.5	28.0	31.6	**28.4**	65.0	19.9	32.1	37.5
	RACE + BoolQ	57.4	26.8	31.8	27.9	63.1	18.0	29.6	36.4
	RACE + NarQ	55.7	**32.2**	30.6	**28.4**	60.9	17.9	28.1	36.3
OBQA	**OBQA**	28.8	51.8	26.1	**34.8**	33.1	6.9	17.3	28.4
	OBQA + SQuAD11	29.6	51.6	**27.2**	33.3	46.3	**9.5**	**23.3**	31.5
	OBQA + SQuAD2	29.5	**53.2**	27.2	33.5	46.6	9.3	23.1	**31.8**
	OBQA + NewsQA	**30.7**	49.4	26.1	32.3	37.8	8.9	22.9	29.7
	OBQA + BoolQ	25.0	50.4	26.0	34.3	27.2	7.1	18.3	26.9
	OBQA + NarQA	29.7	52.8	25.6	33.0	**49.1**	8.9	19.1	31.2

Anchor Dataset / Format	Trained on ↓ - Evaluated on →	BoolQ	MultiRC	NP-BoolQ	BoolQ-CS	Avg
BoolQ	**BoolQ**	76.36	64.10	51.33	53.37	61.3
	BoolQ + SQuAD11	78.41	51.28	54.33	58.36	60.6
	BoolQ + SQuAD2	**78.93**	56.89	**59.38**	58.06	63.3
	BoolQ + NewsQA	77.61	54.17	55.46	59.82	61.8
	BoolQ + RACE	75.69	61.22	54.59	56.89	62.1
	BoolQ + OBQA	76.42	**66.03**	52.03	57.77	63.1
	BoolQ + NarQA	78.90	59.02	55.33	**61.00**	**63.6**

Table 9: Pairwise mixing of formats: mixing with QA of datasets with different formats helps.

A.6 Extended Results of Fine-tuning on Winogrande

Here we provide extended result for the Winogrande dataset. The results are summarized in Table 10. The table include results of fine-tuning UNIFIEDQA$_{T5}$ and UNIFIEDQA$_{BART}$, as well as fine-tuning of the vanilla language models, T5 and BART. As it can be observed, on this dataset, fine-tuning UNIFIEDQA gives stronger results when the size of the training data is limited. With respect to the overall metric AUC, UNIFIEDQA has a slight edge over fine-tuning the vanilla language models.

Model ↓ - Eval. →	Acc. (XS)	Acc. (S)	Acc. (M)	Acc. (L)	Acc. (XL)	AUC
Previous best published	RoBERTa					
	55.4	62.4	66.7	74.2	78.2	67.5
BART$_{large}$ - FT	54.2	57.8	59.7	68.9	72.0	62.4
UnifiedQA$_{BART}$ - FT	56.0	59.5	61.6	68.6	73.3	63.6
T5 - FT	75.6	79.8	86.4	**90.3**	**90.2**	84.8
UnifiedQA$_{T5}$ - FT	**78.8**	**83.4**	**86.9**	88.5	89.4	**85.7**

Table 10: Extended results on the Winogrande dataset