**UnifiedQA-v2:**
**Stronger Generalization via Broader Cross-Format Training**

Daniel Khashabi\(^1\)  
Yeganeh Kordi\(^2\)  
Hannaneh Hajishirzi\(^1,3\)

\(^1\)Allen Institute for AI  \(^2\)Tehran Polytechnic  \(^3\)University of Washington

**Abstract**

We present **UnifiedQA-v2**, a QA model built with the same process as **UnifiedQA**, except that it utilizes more supervision – roughly 3× the number of datasets used for **UnifiedQA**. This generally leads to better in-domain and cross-domain results.\(^1\)

**1 Introduction**

An earlier work by Khashabi et al. (2020b) showed transfer between different QA variants, such as extractive questions (Rajpurkar et al., 2016) and multiple-choice questions (Lai et al., 2017). Based on this finding, Khashabi et al. (2020b) presented **UnifiedQA**, a single QA model trained on datasets of 4 different QA variants, and showed its empirical strength across a variety of benchmarks. Most notably, this approach showed remarkable generalization to unseen datasets compared to models specialized to individual datasets. Followup works have also replicated similar successes on more recent datasets that did not exist at the time of **UnifiedQA**’s construction (Hendrycks et al., 2020; Dasigi et al., 2021; Bragg et al., 2021; Wu et al., 2021; Zhong et al., 2021b) – further strengthening its empirical strength.

A question left open by Khashabi et al. (2020b) is whether there are gains in further broadening the supervision set; and if so, how sizable they are. We revisit this question by training a model (named **UnifiedQA-v2**) with the same supervision set used in building **UnifiedQA**, except that we use 20 datasets for training it (as opposed to 8 datasets used for supervising **UnifiedQA**). Experimental results show that **UnifiedQA-v2** leads to 1-4% performance improvements over **UnifiedQA**, on average (columns indicated with ‘v2 - v1’ in Table 1). The highest average gains correspond to mid-sized ‘large’ models (4.2% for in-domain and 4.5% for out-of-domain setups.

\(^1\) The models are accessible online.

**2 Setup**

We construct a model using the same dataset encoding and hyperparameters as Khashabi et al. (2020b), except that we use larger supervision. Below we outline the relevant construction differences.

The datasets user belong to one of the following four QA formats: extractive, abstractive, multiple-choice and yes/no questions. We henceforth follow this color-coding to indicate our perceived format of each dataset.

**2.1 Training the Models**

Like **UnifiedQA**’s construction, we use the T5 architecture (Raffel et al., 2020) for **UnifiedQA-v2** and train it for 350k steps.\(^2\) We do not perform any dataset-specific model selection since the goal of the study is to have a single model that works well across a variety of datasets.

**Training datasets.** **UnifiedQA** is trained on the following 8 datasets;\(^3\) **SQuAD 1.1, SQuAD 2, NarrativeQA, MCtest, RACE, OpenBookQA, ARC, BoolQ.**\(^4\)

**UnifiedQA-v2** is trained on the following 20 datasets: **SQuAD 1.1, SQuAD 2, NewsQA, Quoref, ROPES, NarrativeQA, DROP, NaturalQuestions, MCtest, RACE, OpenBookQA, ARC, CommonsenseQA, QASC, PhysicalIQA, SocialIQA, Winogrande, BoolQ, MultiRC (yes/no), BoolQ-NP.\(^5\)

**2.2 Evaluation**

We discuss the evaluation setup for comparing **UnifiedQA** and **UnifiedQA-v2**. For each of these models, we evaluate a fixed checkpoint across all the target datasets: checkpoint 250k for **UnifiedQA-v2** and checkpoint 100k for **UnifiedQA**, while **UnifiedQA was trained for 120k steps.**

**The relevant citations are listed in Appendix A.**

**ARC and OpenBookQA datasets are used as-is without any retrieved evidence documents.**
(a) in-domain evaluation

| Eval on → | Size ↓ | AdversarialQA dBiDAF dev (F1) | AdversarialQA dBERT dev (F1) | AdversarialQA dRoberta dev (F1) | ReCorD dev (F1) | QAConv dev (F1) | TweetQA dev (BLEU) | StrategyQA (F1) | PuebloQA (F1) | BosuQ-CS (F1) |
|------------|--------|-------------------------------|-------------------------------|-------------------------------|----------------|----------------|-------------------|----------------|--------------|--------------|
| Small      | 79.4   | 80.7                          | 67.6                          | 66.4                          | 77.1           | 78.9           | 54.5              | 53.7           | 55.2         | 50.4         |
| Base       | 88.2   | 86.8                          | 78.1                          | 74.7                          | 82.5           | 82.4           | 49.8              | 60.4           | 70.3         | 69.0         |
| Large      | 91.1   | 90.8                          | 85.9                          | 86.3                          | 86.1           | 85.8           | 51.6              | 63.2           | 78.0         | 78.8         |
| 3B         | 93.2   | 93.3                          | 87.1                          | 87.1                          | 87.1           | 87.1           | 81.6              | 66.7           | 84.3         | 84.6         |
| 11B        | 93.4   | 93.0                          | 89.6                          | 89.6                          | 90.2           | 90.3           | 67.4              | 66.7           | 84.6         | 85.7         |
| best       |        |                               |                               |                               |                |                |                   |                |              |              |
| specialized |       |                               |                               |                               |                |                |                   |                |              |              |
| model      | 95.6   | 95.6                          | 96.0                          | 95.0                          | 95.0           | 95.0           | 95.0              | 95.0           | 95.0         | 95.0         |

(b) out-of-domain evaluation

| Eval on → | Size ↓ | RACE-C (Accuracy) | HeurQA (Accuracy) | MMMLEU (Accuracy) | RecClor dev (Accuracy) | QuAl (Accuracy) | OneStopQA elementary (Acc) | OneStopQA intermediate (Acc) | OneStopQA advanced (Acc) | Aggregated |
|------------|--------|-------------------|-------------------|--------------------|------------------------|-----------------|------------------------------|-----------------------------|---------------------------|------------|
| Small      | 51.3   | 49.3              | 37.4              | 38.0               | 35.8                   | 31.4            | 40.9                         | 40.1                        | 42.6                      | 39.2       |
| Base       | 63.8   | 67.1              | 30.5              | 31.2               | 32.7                   | 33.2            | 57.4                         | 57.5                        | 59.1                      | 58.2       |
| Large      | 71.5   | 74.0              | 26.0              | 35.3               | 37.6                   | 38.5            | 63.5                         | 63.5                        | 65.5                      | 65.5       |
| 3B         | 77.5   | 80.0              | 40.5              | 40.0               | 44.0                   | 45.0            | 65.4                         | 65.4                        | 67.5                      | 67.5       |
| 11B        | 80.8   | 85.0              | 45.4              | 45.6               | 49.5                   | 50.1            | 71.3                         | 71.3                        | 73.1                      | 73.1       |
| best       |        |                   |                   |                    |                        |                 |                              |                             |                           |            |
| specialized |       |                   |                   |                    |                        |                 |                              |                             |                           |            |
| model      | 48.1   | 50.0              | 44.4              | 50.0               | 50.0                   | 50.0            | 50.0                         | 50.0                        | 50.0                      | 50.0       |

Table 1: Evaluation of UNIFIEDQA and UNIFIEDQA-v2 models, indicated with ‘v1’ and ‘v2’ headers, respectively. Neither of these models is specialized for individual datasets (i.e., one model evaluated across all datasets). The results are split across ‘in-domain’ datasets (observed during training) and ‘out-of-domain’ datasets (not observed during training). UNIFIEDQA-v2 generally leads to better results, as indicated by the aggregate results in the last column. For comparison, the bottom row indicates the best prior models that are often specialized (limited) to individual datasets. The colors indicate the formats: extractive, abstractive, multiple-choice and yes/no.

following Khashabi et al. (2020b)’s setup for unseen dataset.

We evaluate the model for two settings: (1) in-domain evaluation on datasets seen by both models during training (discussed in §2.1). Summary of the results in Table 1a. (2) out-of-domain evaluation on datasets not seen by either of the models. In this setup: We evaluate the models on the following unseen AdversarialQA, ReCoRD, RACE-C, HeadQA, MMLLEU, RecClor, QuAl, OneStopQA, MCScript, MCScript-2.0, CosmosQA, DREAM, ProcessBank, PROST, StrategyQA, PubmedQA, QACov, TweetQA. The results are summarized in Table 1b. All numbers are reported on the test sets, except for those specified with ‘dev’ which are evaluated on dev splits.

Metrics. We evaluate each dataset via their common metric. The complete list of evaluation metrics is in Appendix B. In addition to per-dataset scores, we provide aggregate scores that contrast the two models via a single number (the right-most columns of Table 1). Specifically, we compute the difference between the average performance of v2 and v1 models of the same size (indicated with ‘v2 > v1’) and the percentage of the datasets on which v2 leads to a better performance than v1 of the same size (indicated with ‘v2 > v1’).
Best prior specialized models. For ease of comparison, we include the best published results on each dataset (last rows in Table 1) which generally correspond to models that are fine-tuned to each dataset. While these are useful to understand the room for improvement on each dataset, they are not directly comparable to our target setup where we evaluate a single model across all datasets.

3 Results

We summarize our takeaways from Table 1. Across all experiments, UNIFIEDQA-v2 leads to 1-4% performance improvements over UNIFIEDQA, on average (columns indicated with ‘v2 - v1’). The highest gains correspond to mid-sized ‘large’ models (2.7% for in-domain and 4.5% for out-of-domain setups). Conversely, the lowest gains correspond to the extreme sizes (‘small’ and ‘11B’).

Similar trends can also be seen with ‘v2 > v1’ metric (percentage of datasets on which UNIFIEDQA-v2 outperforms UNIFIEDQA). On this metric, all the numbers are above 50%, which indicates that v2 models generally lead to higher performance. In particular, UNIFIEDQA-v2 of size ‘large’ outperforms UNIFIEDQA of the same size on 80% and 76% of the datasets, for in-domain and out-domain settings, respectively.

The gains are not quite uniform across datasets. Among the in-domain datasets, those with smaller supervision sets generally benefit more from broader training. For example, the ‘small’ models on ARC-easy dataset (around 2k size) enjoy a gain of 4.2% (42.9 → 47.1). We leave further understanding of these gains as a function of various dataset properties to future work.

4 Related Work

The past few years has seen a lot of activity toward unified model designs (Friedman et al., 2021; Aghajanyan et al., 2021; Gupta et al., 2021; Tafjord and Clark, 2021; Lourie et al., 2021; Jiang et al., 2021; Aribandi et al., 2021; Zhong et al., 2021a; Chen et al., 2021; Zhong et al., 2022). For example, Tafjord and Clark (2021) propose a unified model for simultaneously answering questions and explaining its decisions, Zhong et al. (2021a) propose a zero-shot unified model for solving various classification tasks via an existing QA model, and Aribandi et al. (2021) show positive transfer across a wider variety of NLP tasks.

5 Final Thoughts

In this brief study, we showed the empirical benefits of broader training of QA systems, especially when evaluated for out-of-domain setup. The study leaves out several open questions for future work. One open question is whether broad and unified training of QA systems has any benefits in terms of compositional generalization (Keysers et al., 2019; Gu et al., 2021; Bogin et al., 2022), which is often pursued in the multi-hop QA literature. We welcome more work in this direction.

Another open direction is whether one can build unified models that can produce answers at different granularity (Khashabi et al., 2021). This necessitates models that accept more versatile inputs, ideally those that can encompass some definition of the expected outputs. This is reminiscent of the recent literature on solving tasks given their language instructions (Mishra et al., 2021). We hope to see variants of this literature tailored toward QA.

Last but not least, it is not clear whether the gains here should be counted as meaningful progress toward models with better language understanding abilities or mere memorizing of common questions that exploit the weaknesses of our (the research community’s) evaluation protocols. We hope that the future work addresses such important issues by rethinking our means of measuring progress.

Acknowledgment

TPU machines for conducting experiments were provided by Google.

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. arXiv preprint arXiv:2101.11038.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaxiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. 2021. Ext5: Towards extreme multi-task scaling for transfer learning. arXiv preprint arXiv:2111.10952.

Stéphane Aroca-Ouellelette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. Prost: Physical reasoning of objects through space and time. arXiv preprint arXiv:2106.03634.

“One man’s generalization is another man’s carefully engineered inductive bias” –someone on Twitter
Max Bartolo, Alastair Roberts, JohannesWelbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics* (TACL), 8:662–678.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Britany Harding, Brad Huang, Peter Clark, and ChristopherD Manning. 2014. Modeling biological processes for reading comprehension. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 1499–1510.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. Starc: Structured annotations for reading comprehension. In *Annual Meeting of the Association for Computational Linguistics* (ACL).

Ben Bogin, Shivanshu Gupta, and Jonathan Berant. 2022. Unobserved local structures make compositional generalization hard. *arXiv preprint arXiv:2105.06912*.

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. Flex: Unifying evaluation for few-shot nlp. In *Advances in Neural Information Processing Systems* (NeurIPS).

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2021. Meta-learning via language model in-context tuning. *arXiv preprint arXiv:2110.07814*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Conference for the North American Chapter of the Association for Computational Linguistics* (NAACL).

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *ArXiv*, abs/1803.05457.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Conference on Artificial Intelligence* (AAAI).

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP).

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL).

DheeruDua, YizhongWang, PradeepDasigi, GabrielStanovsky, SameerSingh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL).

Dan Friedman, BenDodge, and DanqiChen. 2021. Single-dataset experts for multi-dataset question answering. *Conference on Empirical Methods in Natural Language Processing* (EMNLP).

MorGeva, DanielKhashabi, EladSegal, TusharKhot, DanRoth, and JonathanBerant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics* (TACL), 9:346–361.

Yu Gu, Sue Kase, Michelle Vanni, BrianSadler, Percy Liang, XifengYan, and YuSu. 2021. Beyond id: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference* (WWW), pages 3477–3488.

TanmayGupta, AmitaKamath, AniruddhaKembhavi, and DerekHoiem. 2021. Towards general purpose vision systems. *arXiv preprint arXiv:2104.00743*.

DanHendrycks, CollinBurns, StevenBasart, AndyZou, MantasMazeika, DawnSong, and JacobSteinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations* (ICLR).

LifuHuang, RonanLe Bras, ChandraBhagavatula, and YejinChoi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 2391–2401.

RongtaoHuang, BoweiZou, YuHong, WeiZhang, AitiAw, and GuodongZhou. 2020. Nut-re: Noisy user-generated text-oriented reading comprehension. In *International Conference on Computational Linguistics* (COLING), pages 2687–2698.

LiweiJiang, JenaD Hwang, ChandraBhagavatula, Ronan Le Bras, MaxwellForbes, JonBorchardt, JennyLiang, OrenEtzioni, MaartenSap, and YejinChoi. 2021. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07374*.

YufanJiang, ShuangzhiWu, JingGong, YahuiCheng, PengMeng, WeiliangLin, ZhiboChen, et al. 2020. Improving machine reading comprehension with single-choice decision and transfer learning. *arXiv preprint arXiv:2011.03292*.

QiaoJin, BhuwanDhindra, ZhengpingLiu, WilliamCohen, and XinghuaLu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP).
Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for squad. In Annual Meeting of the Association for Computational Linguistics (ACL).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Conference on Empirical Methods in Natural Language Processing (EMNLP).

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In Conference on Empirical Methods in Natural Language Processing (EMNLP).

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to ai comprehension for computational linguistics. In AAAI.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WINORANDE: an adversarial winograd schema challenge at scale. In Conference on Artificial Intelligence (AAAI).

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. Transactions of the Association for Computational Linguistics (TACL), 7:217–231.

Yu Sun, Shuhuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137.

Oyvind Tafjord and Peter Clark. 2021. General-purpose question-answering with macaw. arXiv preprint arXiv:2109.02593.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Karen Suleman. 2017. Newsqa: A machine comprehension dataset. In Rep4NLP@ACL.

David Vilares and Carlos Gómez-Rodríguez. 2019. Head-qa: A healthcare dataset for complex reasoning. In Annual Meeting of the Association for Computational Linguistics (ACL), pages 960–966.

Chien-Sheng Wu, Andrea Madotto, Wenhao Liu, Pascale Fung, and Caiming Xiong. 2021. Qaconv: Question answering on informative conversations. arXiv preprint arXiv:2105.06912.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Tweetqa: A social media focused question answering dataset. In Annual Meeting of the Association for Computational Linguistics (ACL), pages 5020–5031.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. In Conference on Empirical Methods in Natural Language Processing (EMNLP).

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2019. Reclor: A reading comprehension dataset requiring logical reasoning. In International Conference on Learning Representations (ICLR).

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. arXiv preprint arXiv:1810.12885.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. In Conference on Artificial Intelligence (AAAI).

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021a. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In Conference on Empirical Methods in Natural Language Processing (EMNLP) - Findings, pages 2856–2878.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021b. Meta-tuning language models to answer prompts better. In Conference on Empirical Methods in Natural Language Processing (EMNLP) - Findings.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. 2022. Summarizing differences between text distributions with natural language. arXiv preprint.

A Dataset References

For ease of readability in the main text, the citations corresponding to each dataset are included here: SQuAD 1.1 (Rajpurkar et al., 2016), SQuAD 2 (Rajpurkar et al., 2018), NewsQA (Trischler et al., 2017) Quoref (Dasigi et al., 2019), ROPEs (Lin et al., 2019), AdversarialQA (Bartolo et al., 2020), ReCoRD (Zhang et al., 2018), DROP (Dua et al., 2019) NarrativeQA/NarQA (Kociský et al., 2018), QAConv (Wu et al., 2021), TweetQA (Xiong et al., 2019), HeadQA (Vilares and Gómez-Rodríguez, 2019), RACE-C (Liang et al., 2019), MCTest (Richardson et al., 2013), RACE (Lai
et al., 2017), OpenBookQA (Mihaylov et al., 2018), ARC (Clark et al., 2018, 2016), QASC (Khot et al., 2019), CommonsenseQA/CQA (Talmor et al., 2019), Winogrande (Sakaguchi et al., 2020), MMMLU (Hendrycks et al., 2020), ReClor (Yu et al., 2019), Quail (Rogers et al., 2020), OneStopQA (Berzak et al., 2020), MCScript (Ostermann et al., 2018), MCScript 2.0 (Ostermann et al., 2019), CosmosQA (Huang et al., 2019), ProcessBank (Berant et al., 2014), DREAM (Sun et al., 2019), PROST (Aroca-Ouellette et al., 2021), BoolQ (Clark et al., 2019), BoolQ-NP (Khashabi et al., 2020a) the binary (yes/no) subset of MultiRC (Khashabi et al., 2018), StrategyQA (Geva et al., 2021), PubmedQA (Jin et al., 2019).

B Dataset Metrics

Here we summarize the metrics that were used for the evaluation. For each dataset, we have picked a metric that is recommended by dataset authors, and whenever possible, we use their script for evaluation. For multiple-choice tasks we always used accuracy: if the correct answer is picked, then the model gains one credit. Otherwise, no credit is given. For extractive tasks, we use F1 token overlap between the answer text and gold answer(s). There are different ways of computing this score, usually based on how the texts are normalized. Therefore, for each dataset, we used its official F1 computation. For NarrativeQA we used ROUGE-L. For TweetQA we used BLEU. For Yes/No questions, we also use SQuAD 1.1’s F1 token overlap: in most cases that the model provides “yes”/“no” answers, F1 acts like binary scoring (correct or incorrect credits). When the predictions contain more than “yes”/“no” (like, “yes it is correct”), it provides a soft measure of correctness.

\footnote{Our evaluation script is public (see Footnote 1).}