	bAbI Tasks 1-10									
Dataset	1	2	3	4	5	6	7	8	9	10
True dataset	**100%**	**100%**	39%	**100%**	**99%**	**100%**	**94%**	**97%**	**99%**	**98%**
Question only	18%	17%	22%	22%	34%	50%	48%	34%	64%	44%
Passage only	53%	86%	**60%**	59%	31%	48%	85%	79%	63%	47%
$\Delta(min)$	-47	-14	$+21$	-41	-65	-52	-9	-18	-35	-51
	bAbI Tasks 11-20									
	11	12	13	14	15	16	17	18	19	20
True dataset	**94%**	**100%**	**94%**	**96%**	**100%**	**48%**	**57%**	**93%**	**30%**	**100%**
Question only	17%	15%	18%	18%	34%	26%	48%	91%	10%	70%
Passage only	71%	74%	**94%**	50%	64%	**47%**	48%	53%	21%	**100%**
$\Delta(min)$	-23	-26	0	-46	-36	-1	-9	-2	-9	0

Table 1: Accuracy on bAbI tasks using our implementation of the Key-Value Memory Networks

Task	Full	Q-only	P-only	$\Delta(min)$
Key-Value Memory Networks				
CBT-NE	**35.0%**	29.1%	24.1%	-5.9
CBT-CN	**37.6%**	32.4%	24.4%	-5.2
CBT-V	52.5%	**55.7%**	36.0%	$+3.2$
CBT-P	55.2%	**56.9%**	30.1%	$+1.7$
Gated Attention Reader				
CBT-NE	**74.9%**	50.6%	40.8%	-17.5
CBT-CN	**70.7%**	54.0%	36.7%	-16.7
CNN	**77.8%**	25.6%	38.3%	-39.5
WdW	**67.0%**	41.8%	52.2%	-14.8
WdW-R	**69.1%**	50.0%	50.6%	-15.6

Table 2: Accuracy on various datasets using KV-MemNets (window memory) and GARs

Task	Complete passage	Last sentence
CBT-NE	22.6%	**22.8%**
CBT-CN	**31.6%**	24.8%
CBT-V	**48.8%**	45.0%
CBT-P	34.1%	**37.9%**

Table 3: Accuracy on CBT tasks using KV-MemNets (sentence memory) varying passage size.

Metric	Full	Q-only	P-only	$\Delta(min)$
EM	**70.7%**	0.6%	10.9%	-59.8
F1	**79.1%**	4.0%	14.8%	-64.3

Table 4: Performance of QANet on SQuAD

suppressed baselines and 5 additional baselines reported by Onishi et al. (2016). We suspect that the models memorize attributes of specific entities, justifying the entity-anonymization used by Hermann et al. (2015) to construct the CNN dataset.

SQuAD Our results suggest that SQuAD is an unusually carefully-designed and challenging RC task. The span selection mode of answering requires that models consider the passage thus the abysmal performance of the Q-only QANet (Table 4). Since SQuAD requires answering by span selection, we construct Q-only variants here by placing answers from all relevant questions in random order, filling the gaps with random words. Moreover, Q-only and P-only models achieve F1 scores of only 4% and 14.8% resp. (Table 4), significantly lower than 79.1 on the proper task.

5 Discussion

We briefly discuss our findings, offer some guiding principles for evaluating new benchmarks and algorithms, and speculate on why some of these problems may have gone under the radar. Our goal is not to blame the creators of past datasets but instead to support the community by offering practical guidance for future researchers.