Unify and Conquer

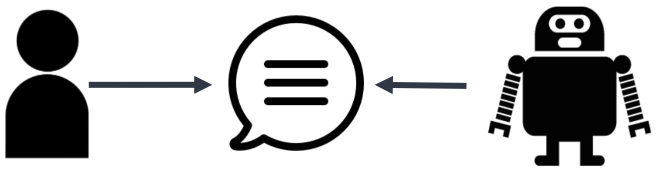
Towards a *Unified* View of Machine Comprehension

Daniel Khashabi
Allen Institute for AI, Seattle

Moving towards NLU, via QA

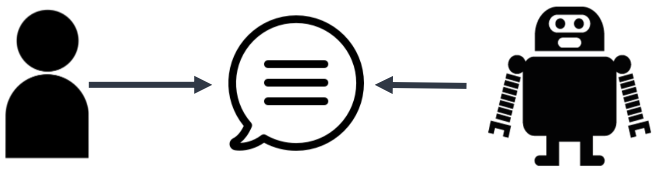
Moving towards NLU, via QA

- Natural Language Understanding:
  - Interpret a given text similar to humans.

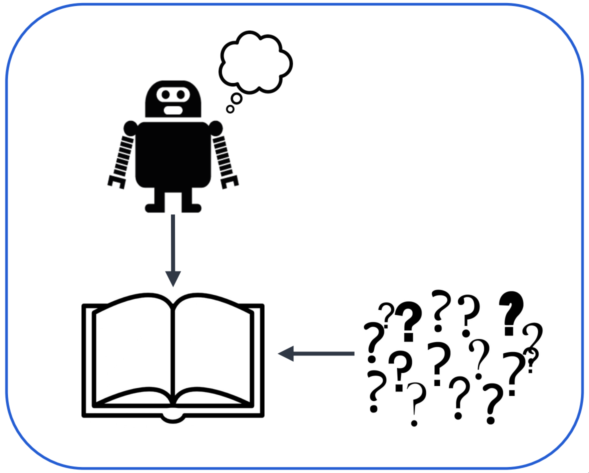


- Measuring the progress by answering questions.
  - A system that is better in understanding language, should have a higher chance of answering these questions.
  - This has been used in the field for many years
    - Question Answering,
    - Reading Comprehension,
    - Machine Comprehension, etc.

Moving towards NLU, via QA



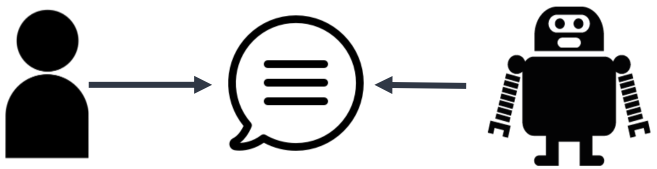
- Natural Language Understanding:
  - Interpret a given text similar to humans.
- Measuring the progress by answering questions.
  - A system that is better in understanding language, should have a higher chance of answering these questions.
  - This has been used in the field for many years
    - Question Answering,
    - Reading Comprehension,
    - Machine Comprehension, etc.



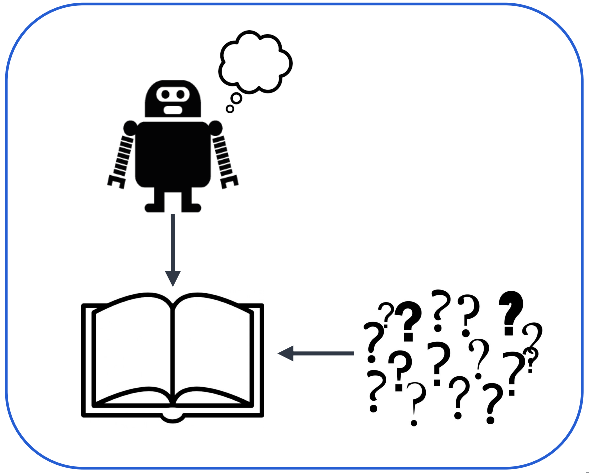
[Winograd, 1972; McCarthy 1976; Lehnert, 1977b; others]

Moving towards NLU, via QA

- Natural Language Understanding:
  - Interpret a given text similar to humans.



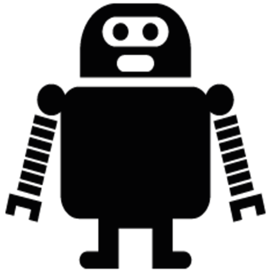
- Measuring the progress by answering questions.
  - A system that is better in understanding language, should have a higher chance of answering these questions.
  - This has been used in the field for many years
    - Question Answering,
    - Reading Comprehension,
    - Machine Comprehension, etc.



[Winograd, 1972; McCarthy 1976; Lehnert, 1977b; others]

QA; a broad definition

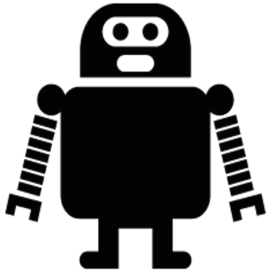
- **Task:** Question Answering (QA)



QA; a broad definition

- **Task:** Question Answering (QA)

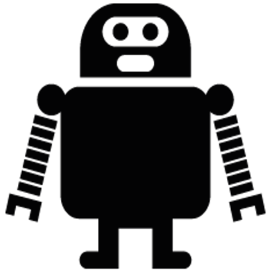
"What does photosynthesis produce that helps plants grow?"



QA; a broad definition

- **Task:** Question Answering (QA)

"What does photosynthesis produce that helps plants grow?"

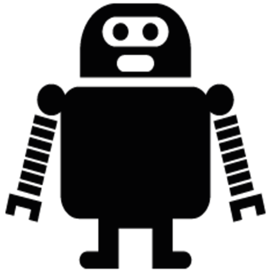


Input: *A question, along with additional information (hints, docs, images, etc.)*

QA; a broad definition

- **Task:** Question Answering (QA)

"What does photosynthesis produce that helps plants grow?"

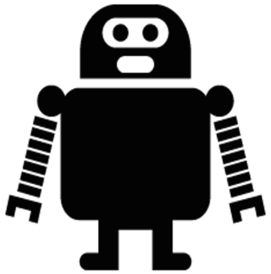


Input: *A question, along with additional information (hints, docs, images, etc.)*

QA; a broad definition

- **Task:** Question Answering (QA)

"What does photosynthesis produce that helps plants grow?"

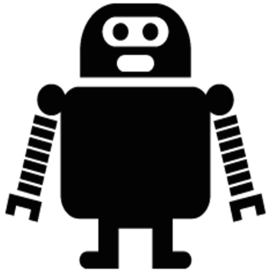


Input: *A question, along with additional information (hints, docs, images, etc.)*

QA; a broad definition

- **Task:** Question Answering (QA)

"What does photosynthesis produce that helps plants grow?"

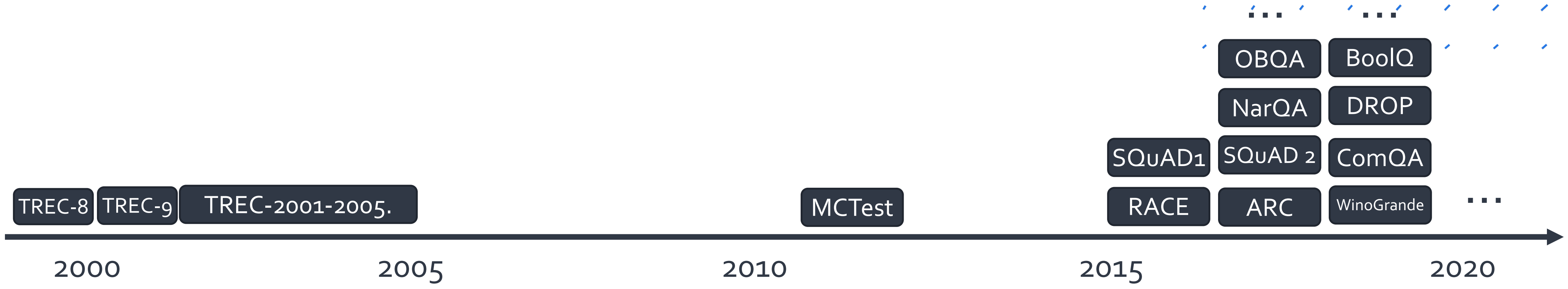


"sugar"

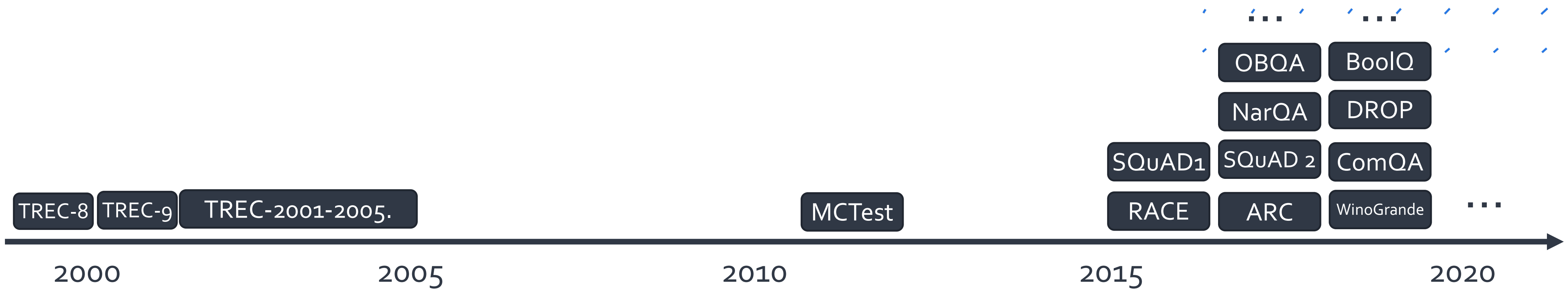
Input: *A question, along with additional information (hints, docs, images, etc.)*

Output: *a string that addresses the input question.*

QA datasets

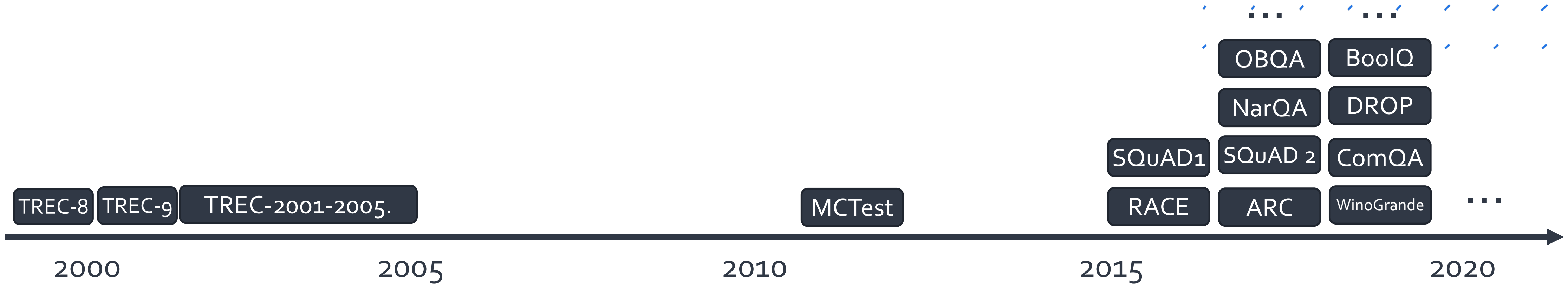

...
...
OBQA
BoolQ
NarQA
DROP
SQuAD1
SQuAD 2
ComQA
TREC-8
TREC-9
TREC-2001-2005.
MCTest
RACE
ARC
WinoGrande
...
2000
2005
2010
2015
2020

QA datasets



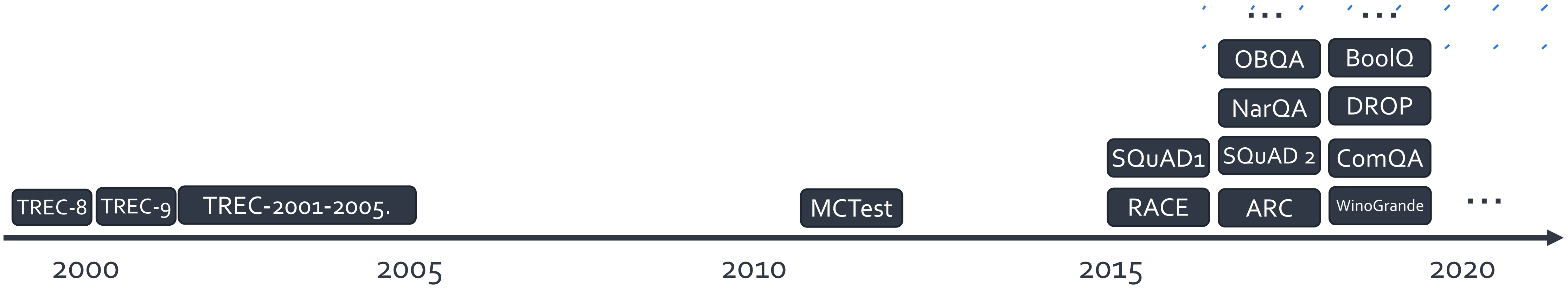
- Motivations for publishing new datasets:
  - Unexplored reasoning challenges
  - Alternate (better?) evaluation protocol (expand)

QA datasets



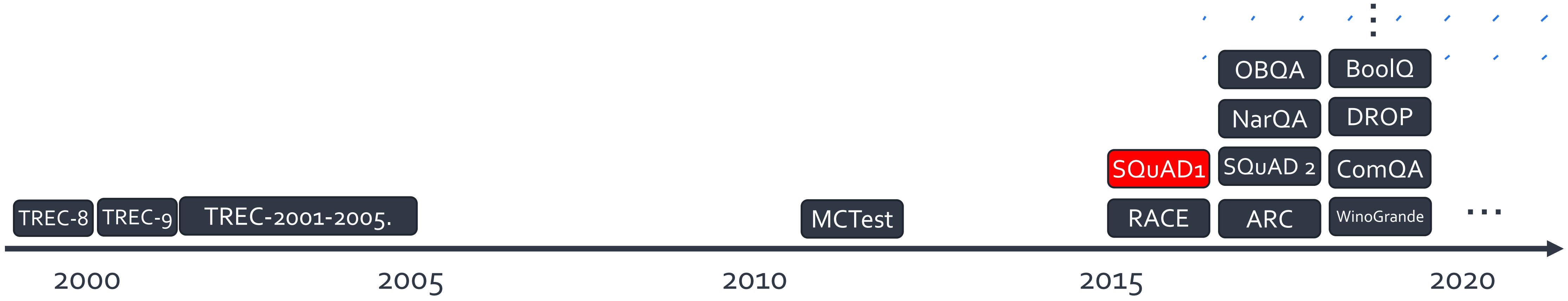
- Motivations for publishing new datasets:
  - Unexplored reasoning challenges
  - Alternate (better?) evaluation protocol (expand)

QA datasets



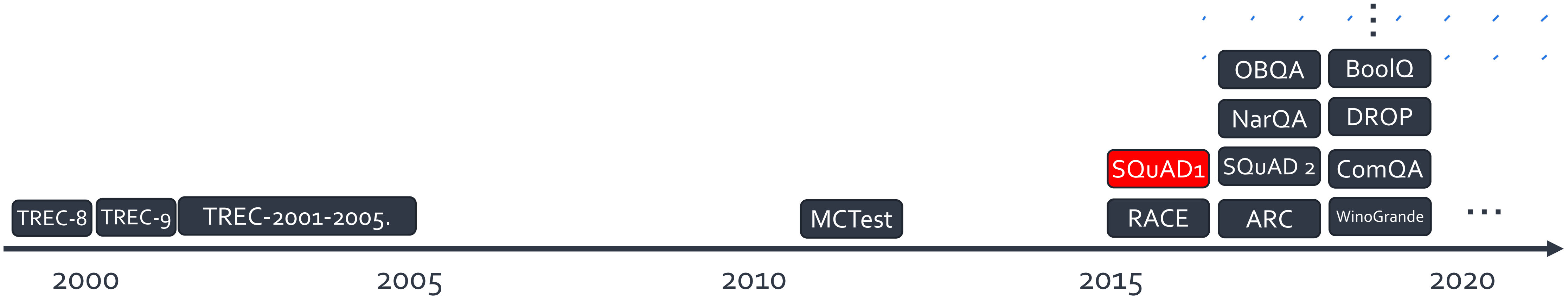
- Motivations for publishing new datasets:
  - Unexplored reasoning challenges
  - Alternate (better?) evaluation protocol (expand)

QA datasets



[Rajpurkar et al, 2016]

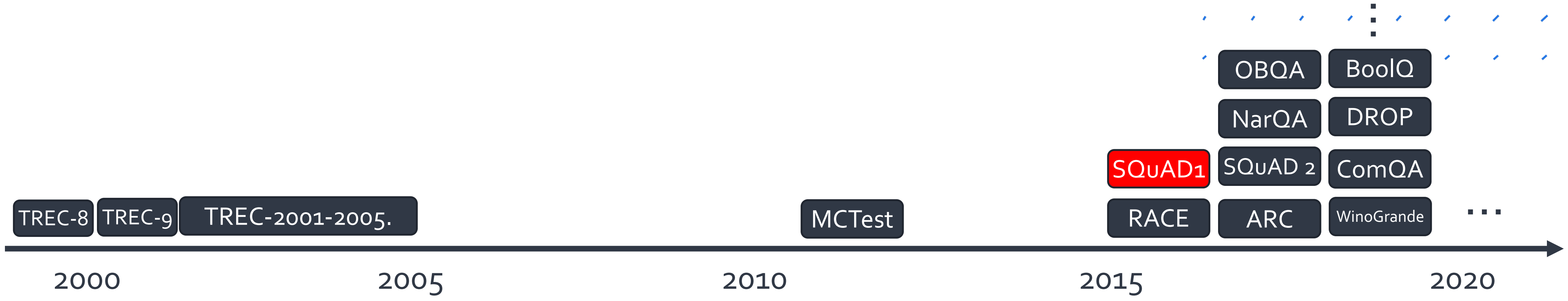
QA datasets



Question: "*At what speed did the turbine operate?*"

[Rajpurkar et al, 2016]

QA datasets

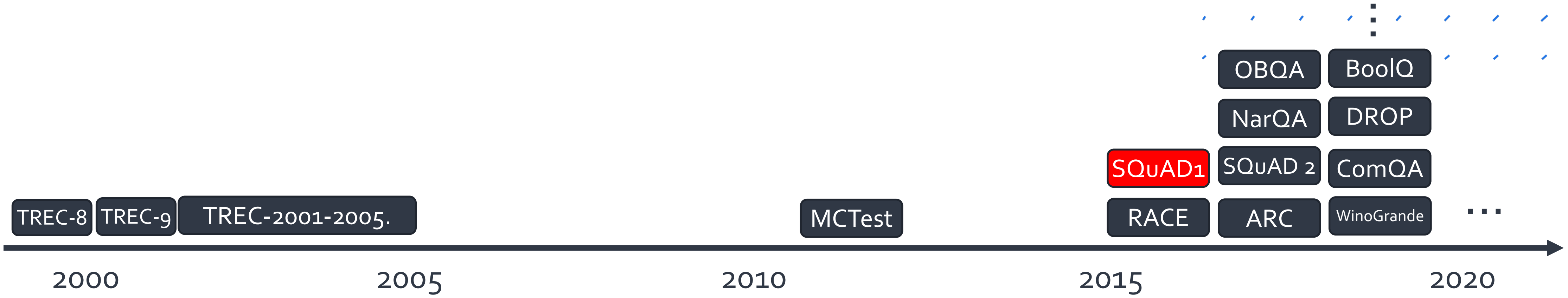


Question: *"At what speed did the turbine operate?"*

Candidates: *(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...*

[Rajpurkar et al, 2016]

QA datasets

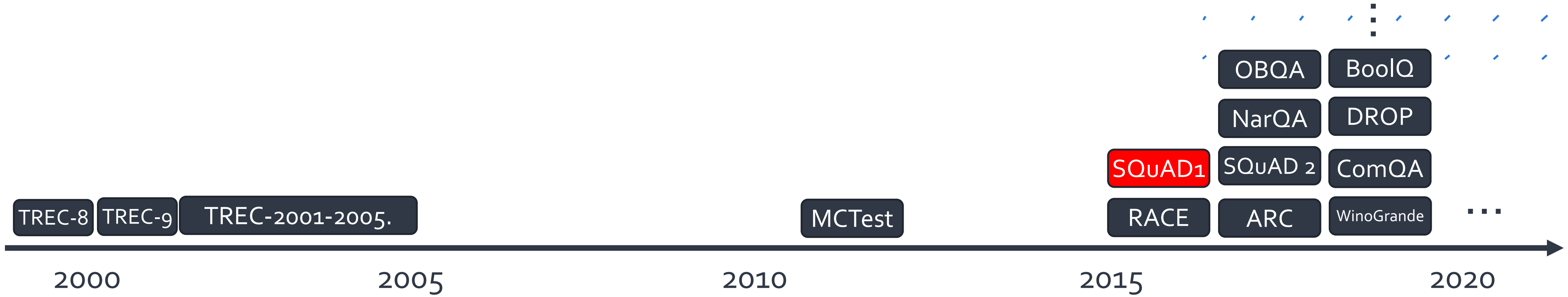


Question: *"At what speed did the turbine operate?"*

Candidates: *(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...*

[Rajpurkar et al, 2016]

QA datasets



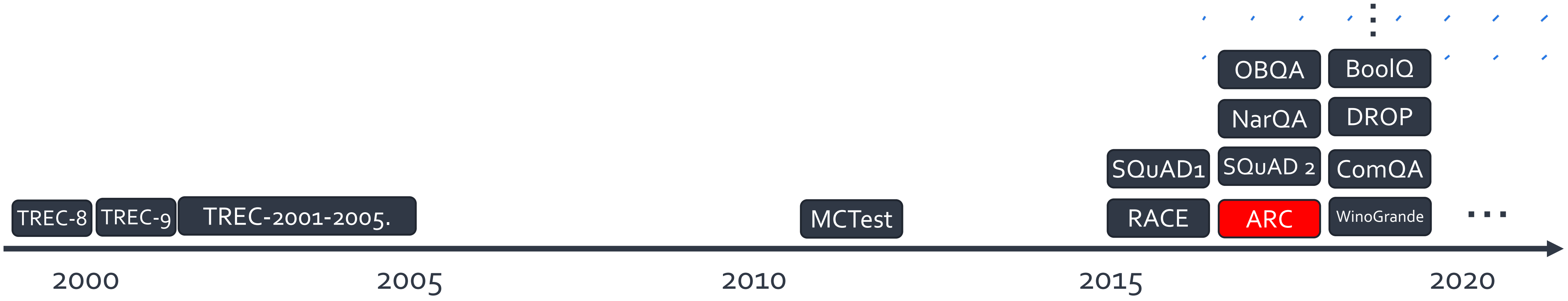
Question: *"At what speed did the turbine operate?"*

Candidates: *(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...*

"16,000 rpm"

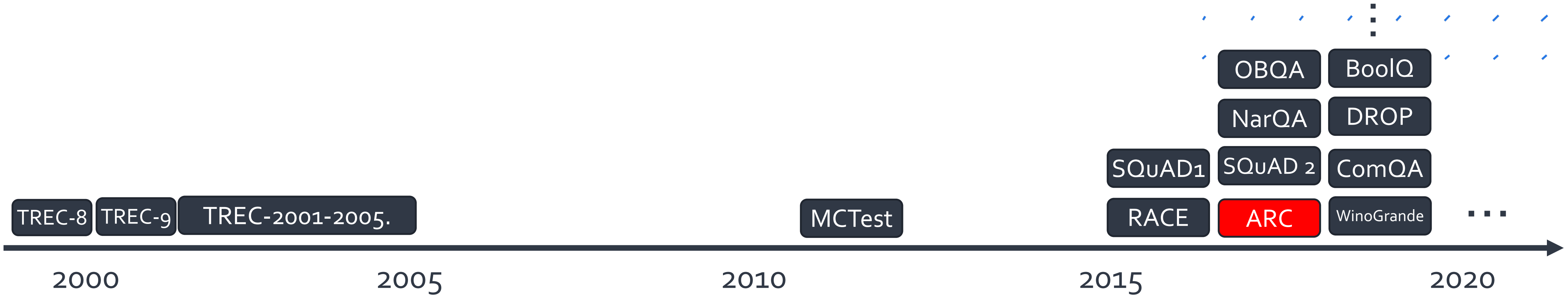
[Rajpurkar et al, 2016]

QA datasets



[Clark et al, 2018]

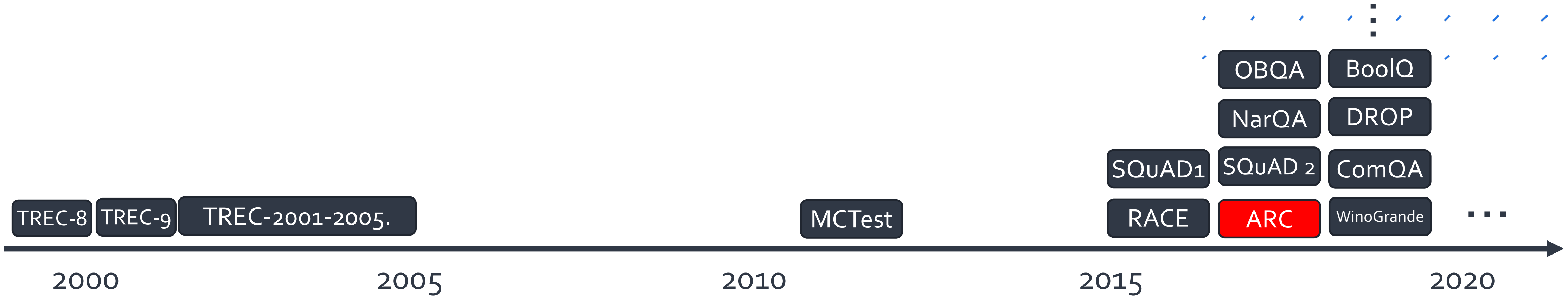
QA datasets



Question: "*What does photosynthesis produce that helps plants grow?* "

[Clark et al, 2018]

QA datasets

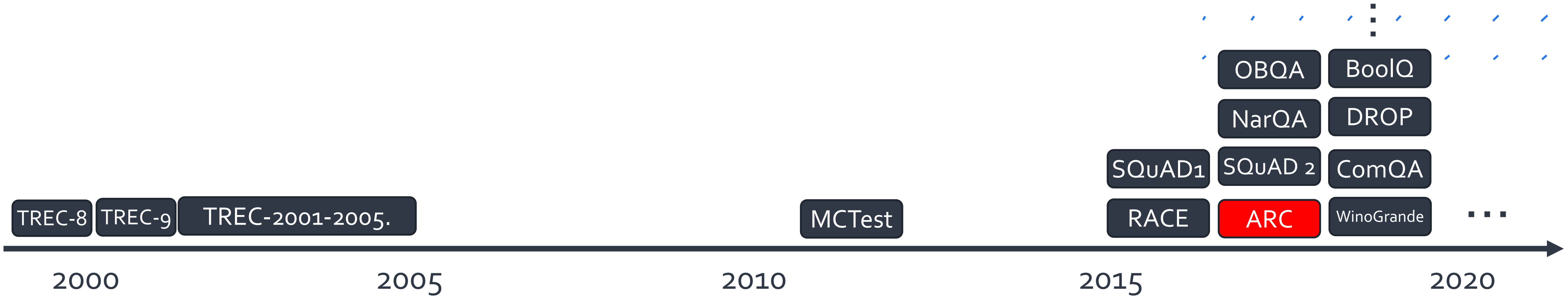


Question: "*What does photosynthesis produce that helps plants grow?"*

Candidates: *(A) water*
(B) oxygen
(C) protein
(D) sugar

[Clark et al, 2018]

QA datasets

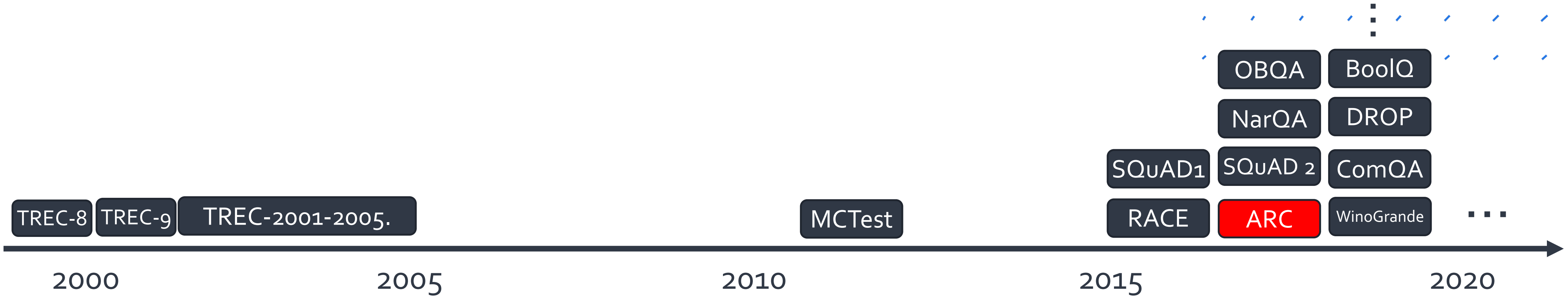


Question: "*What does photosynthesis produce that helps plants grow?* "

Candidates:	*(A) water*
	(B) oxygen
	(C) protein
	(D) sugar

[Clark et al, 2018]

QA datasets



Question: "*What does photosynthesis produce that helps plants grow?* "

Candidates:
(A) water
(B) oxygen
(C) protein
(D) sugar

"The big kid"

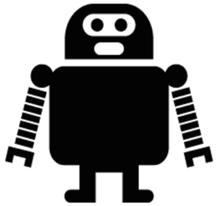
[Clark et al, 2018]

QA Terminology

QA Terminology

- **"Task":** well-formed response for a well-formed question.

Input:
well-formed question



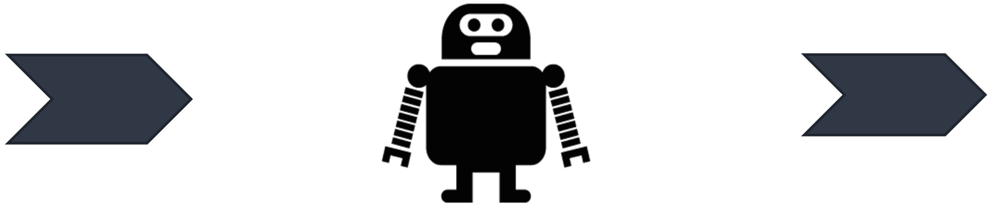
Output:
a well-formed response

- **"Format":** QA with particular **assumptions** about input/output.
  - Defined by datasets.
  - A necessity for automatic evaluation.
  - Depends on the reasoning problem, too.

QA Terminology

- **"Task":** well-formed response for a well-formed question.

Input:
well-formed question



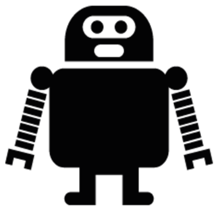
Output:
a well-formed response

- **"Format":** QA with particular **assumptions** about input/output.
  - Defined by datasets.
  - A necessity for automatic evaluation.
  - Depends on the reasoning problem, too.

QA Terminology

- **"Task":** well-formed response for a well-formed question.

Input:
well-formed question



Output:
a well-formed response

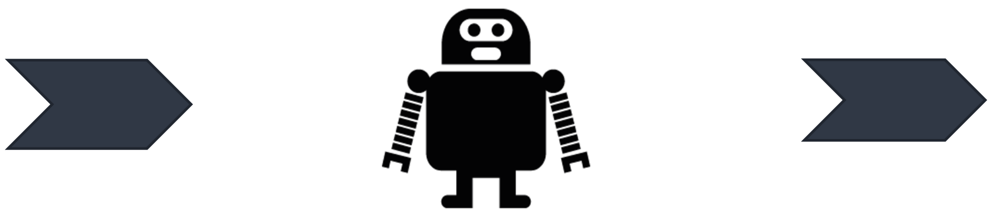
- **"Format":** QA with particular **assumptions** about input/output.
  - Defined by datasets.
  - A necessity for automatic evaluation.
  - Depends on the reasoning problem, too.

Format	Example dataset
Multiple-choice	CommonsenseQA [Talmor et al'19]
YesNo	BoolQ [Clark et al'19]
extractive	SQuAD [Rajpurkar et al'16]
abstractive	NarrativeQA [Kociský et al'18]

QA Terminology

- **"Task":** well-formed response for a well-formed question.

Input:
well-formed question



Output:
a well-formed response

- **"Format":** QA with particular **assumptions** about input/output.
  - Defined by datasets.
  - A necessity for automatic evaluation.
  - Depends on the reasoning problem, too.

Format	Example dataset
Multiple-choice	CommonsenseQA [Talmor et al'19]
YesNo	BoolQ [Clark et al'19]
extractive	SQuAD [Rajpurkar et al'16]
abstractive	NarrativeQA [Kociský et al'18]

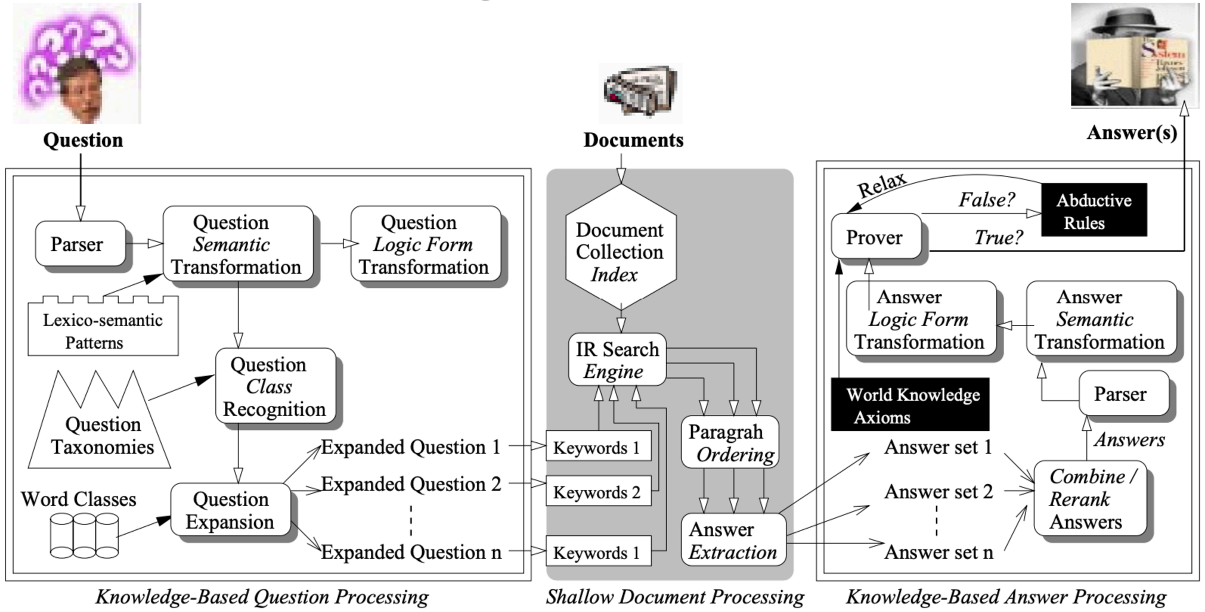
Our progress in QA: the good

- More general language representations.

Our progress in QA: the good

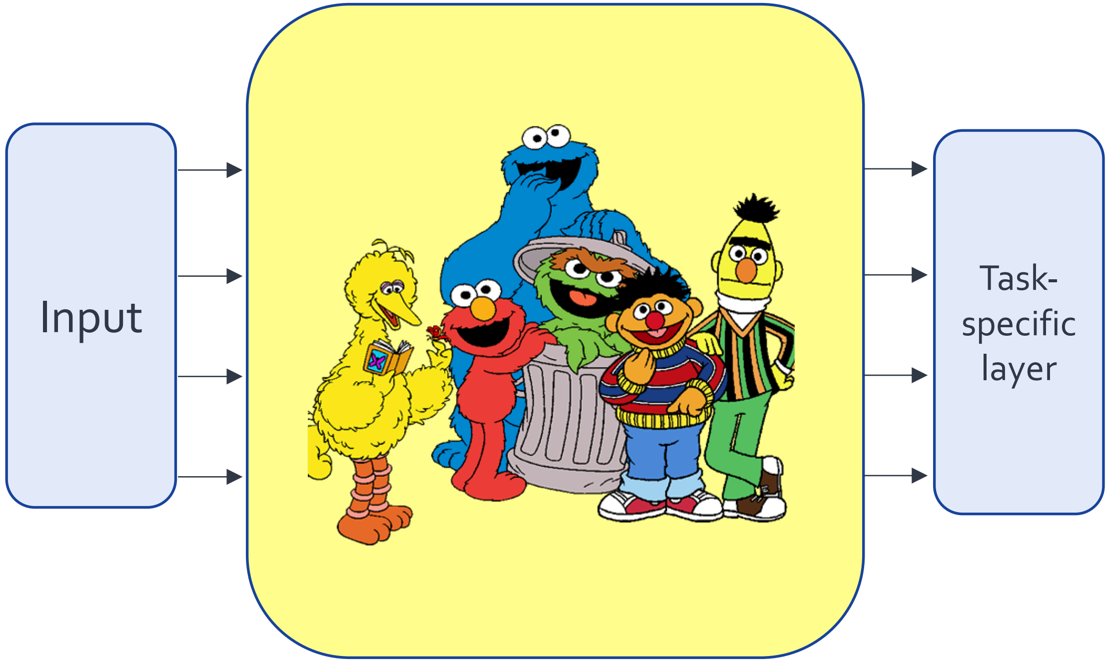
- More general language representations.

[Harabagiu et al, 2000; others]



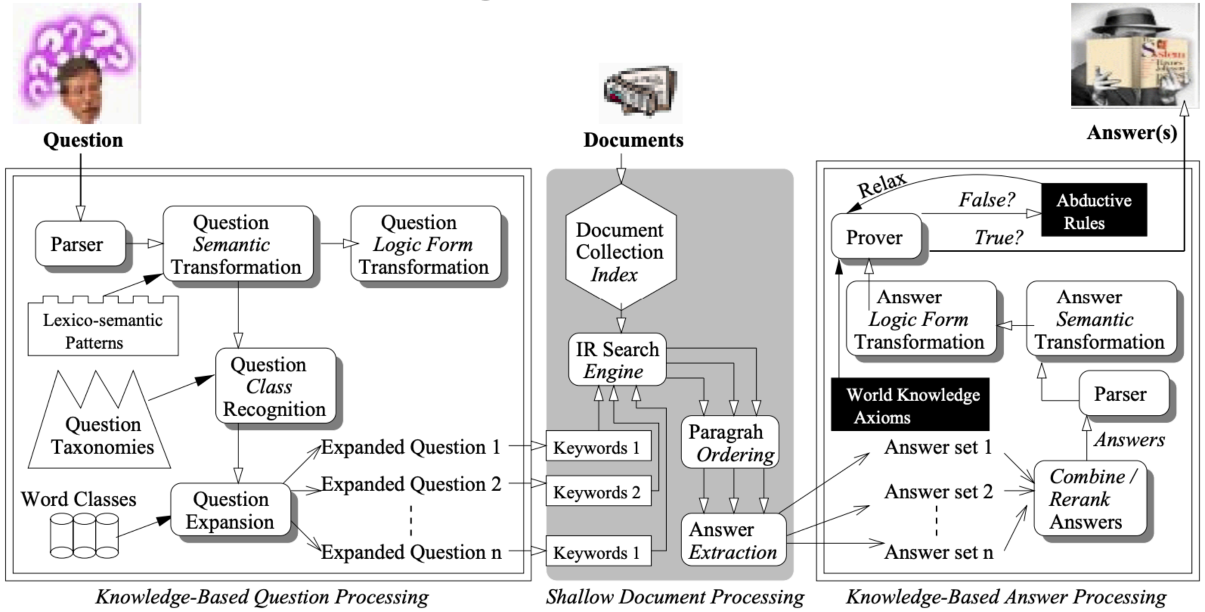
Our progress in QA: the good

- More general language representations.

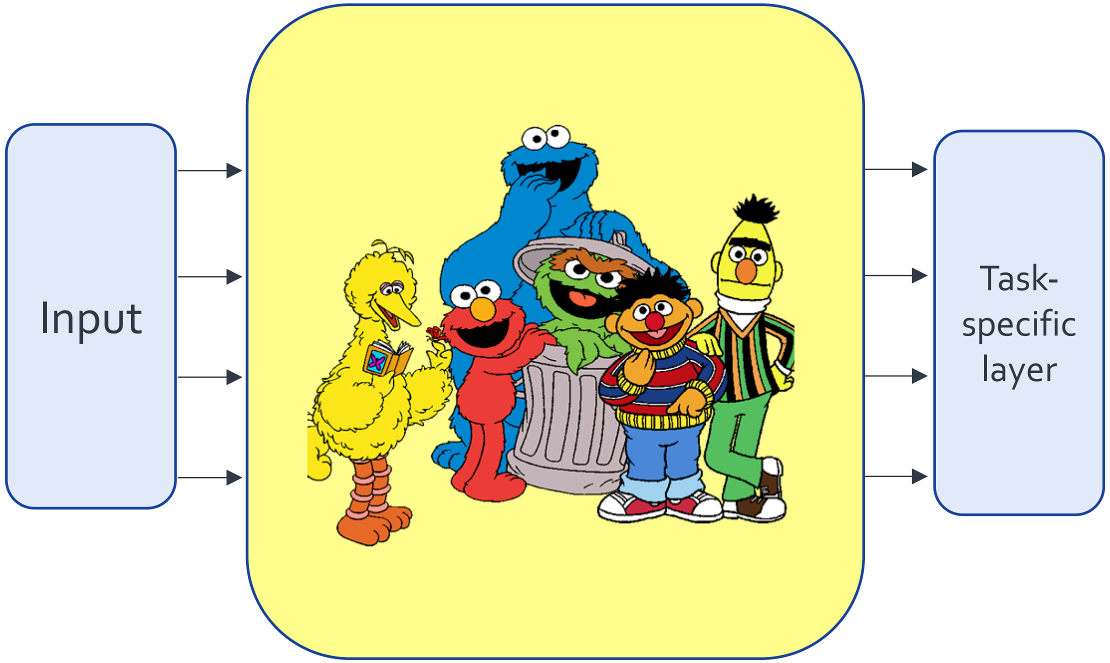


[Peters et al; Devlin et al; others]

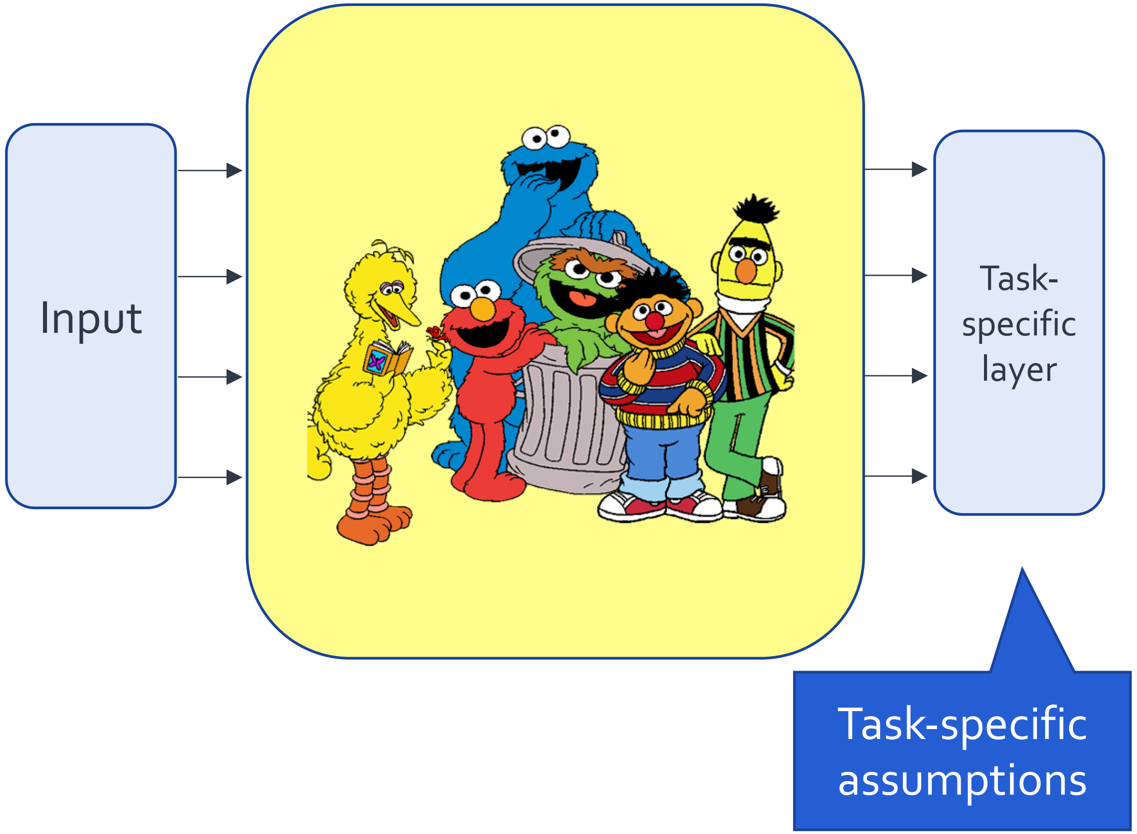
[Harabagiu et al, 2000; others]



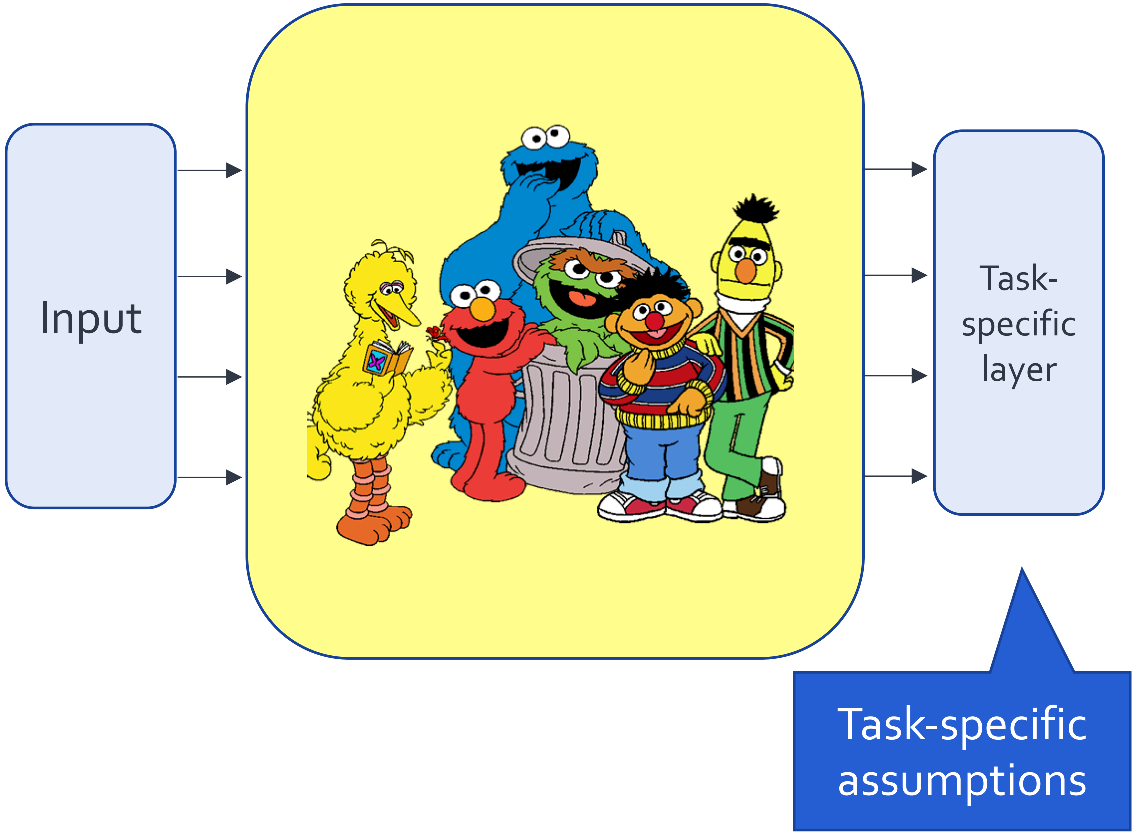
Our progress in QA: the bad



Our progress in QA: the bad

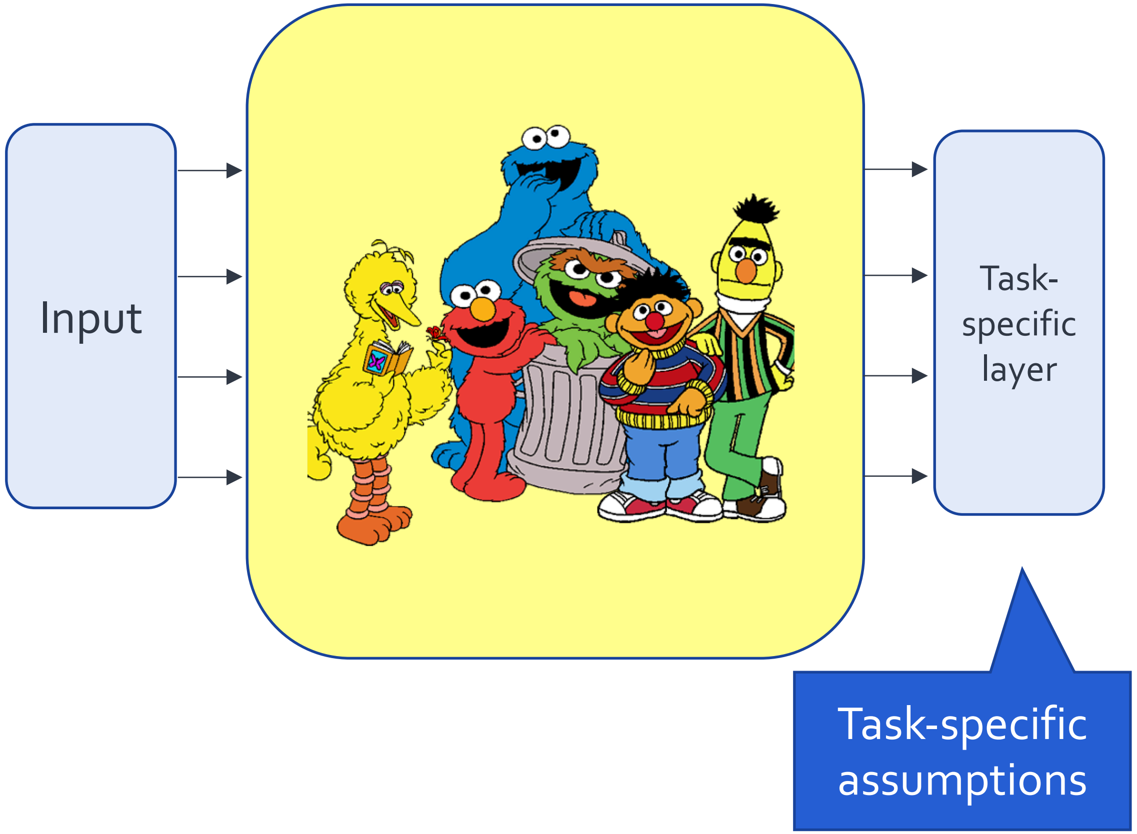


Our progress in QA: the bad



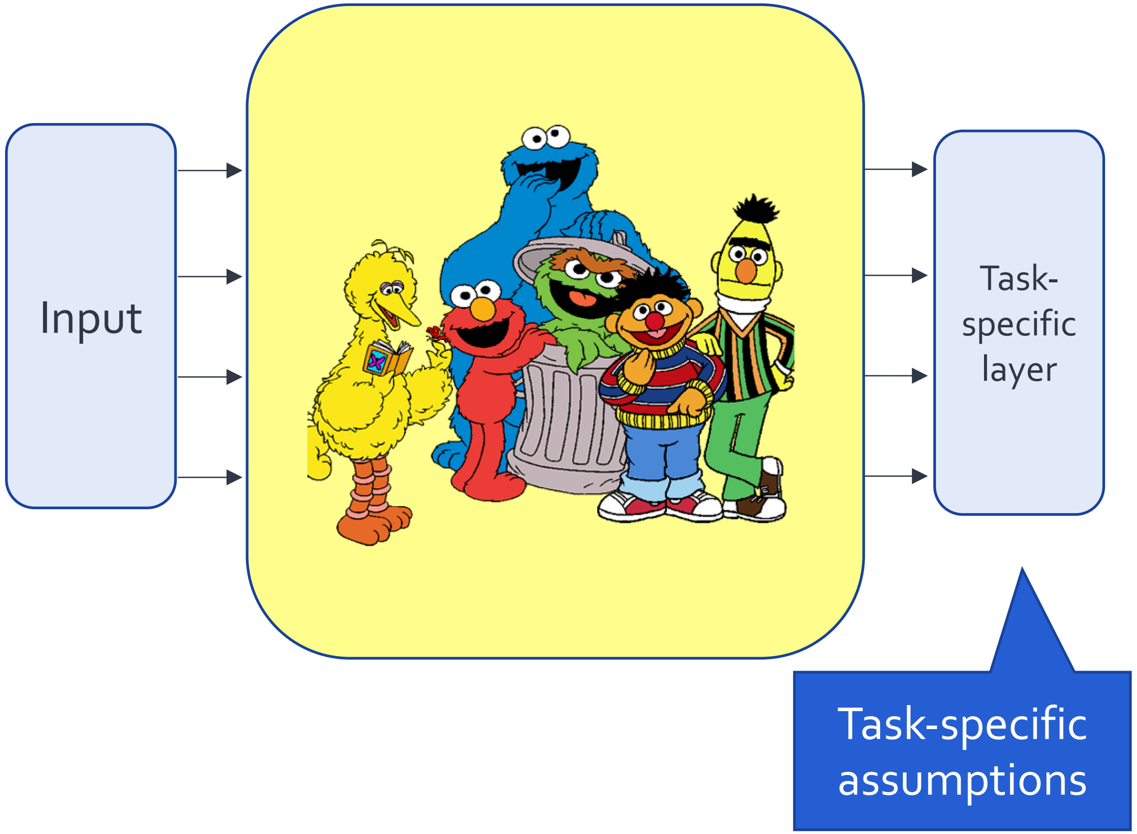
format	assumption
Yes/No QA	
Multiple-choice QA	
Extractive QA	
Abstractive QA	

Our progress in QA: the bad



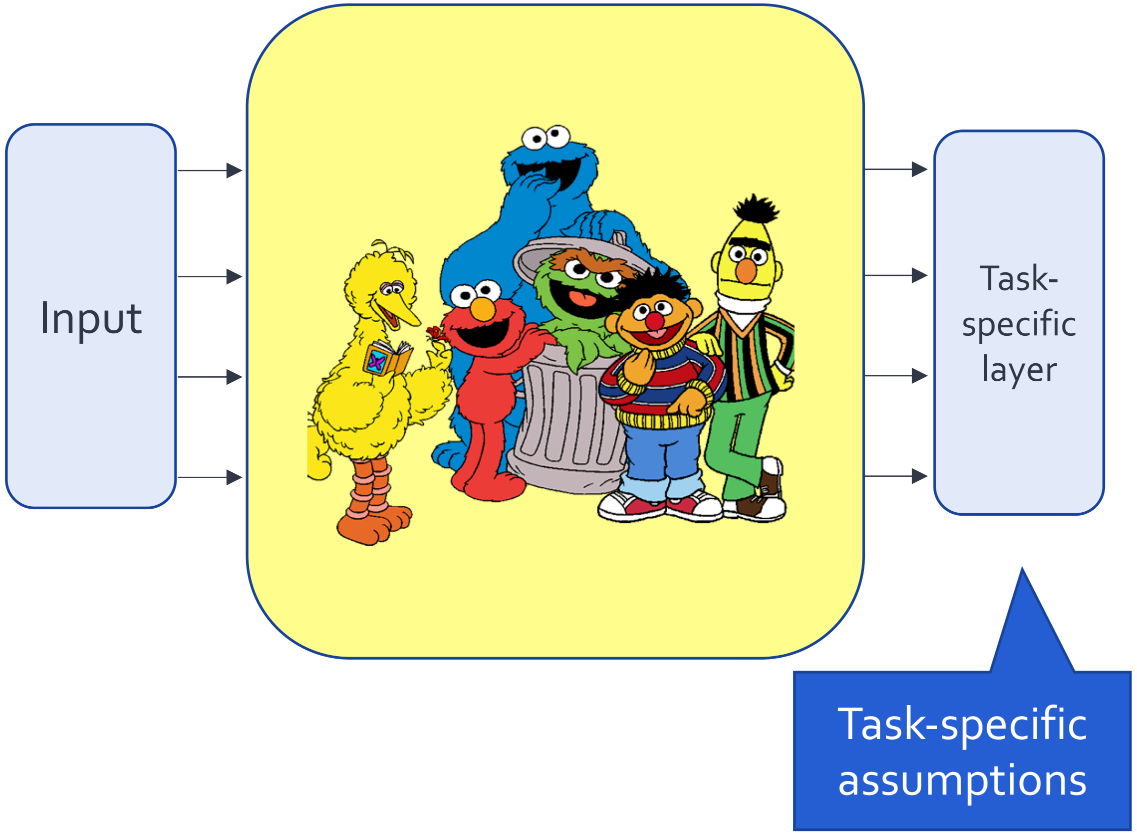
format	assumption
Yes/No QA	*binary output*
Multiple-choice QA	
Extractive QA	
Abstractive QA	

Our progress in QA: the bad



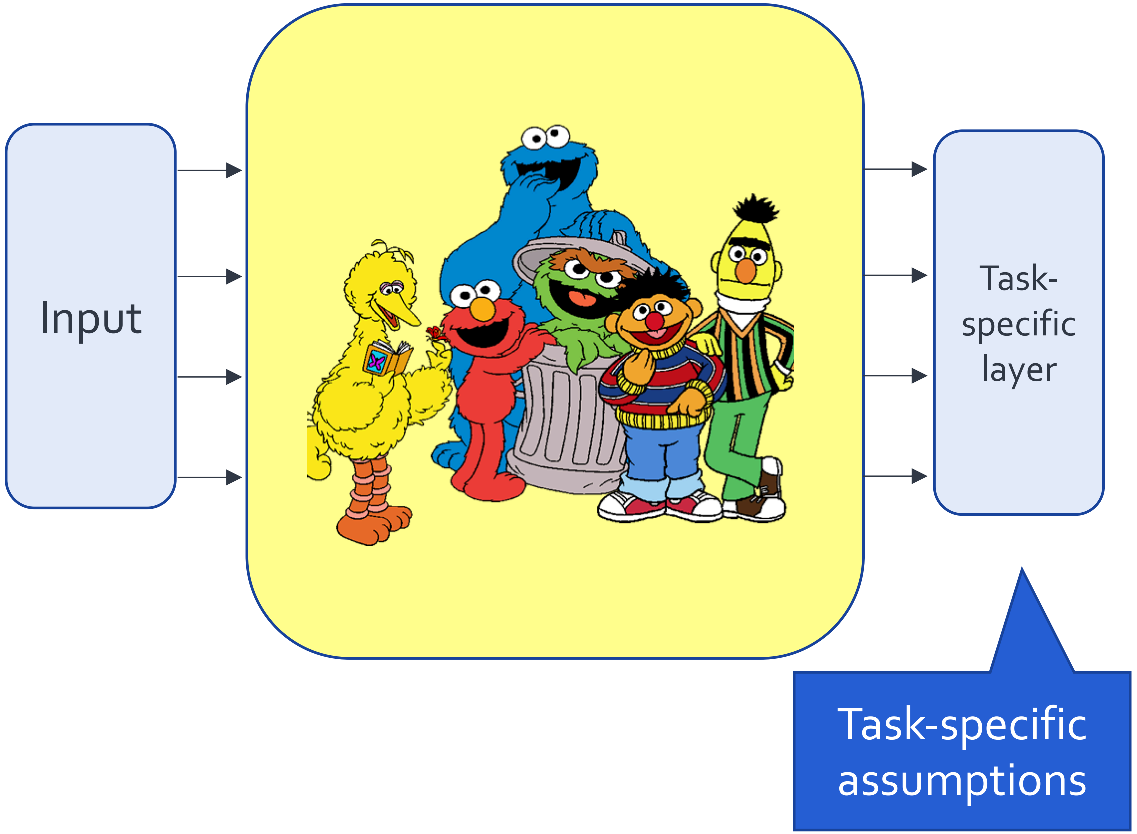
format	assumption
Yes/No QA	*binary output*
Multiple-choice QA	*exactly one of the candidate answers is correct.*
Extractive QA	
Abstractive QA	

Our progress in QA: the bad



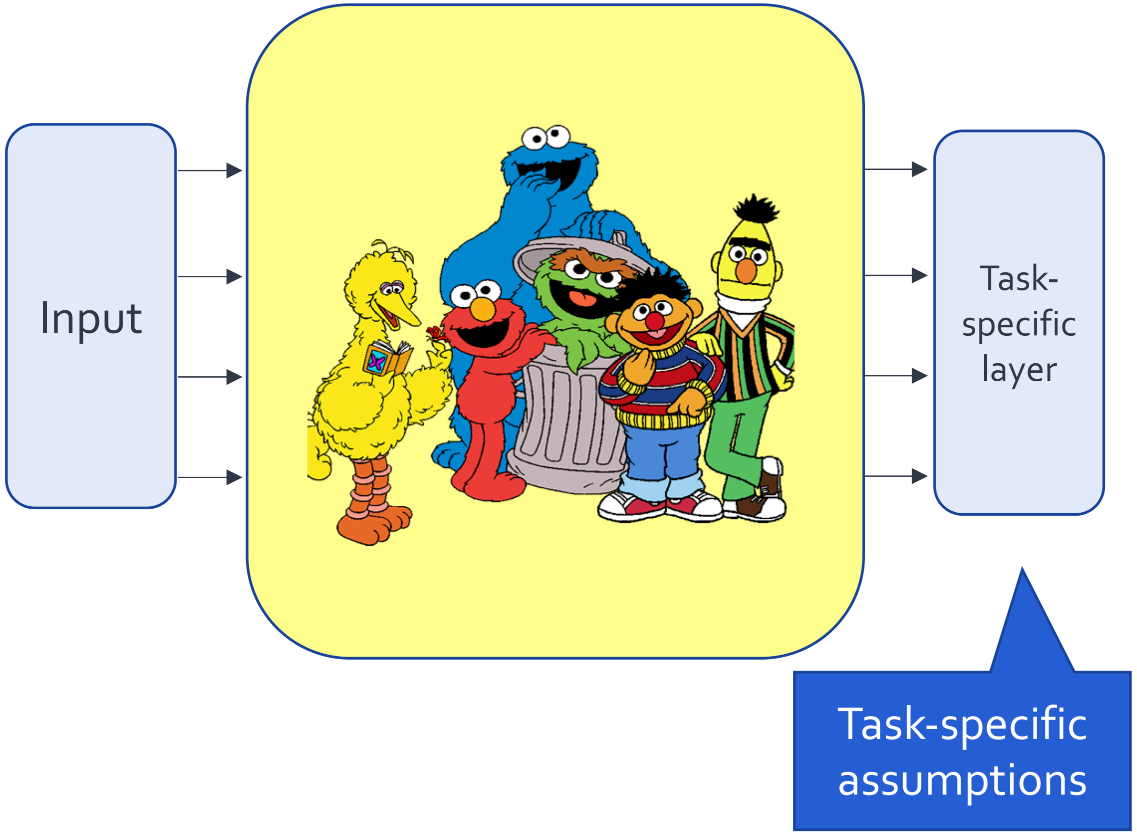
format	assumption
Yes/No QA	*binary output*
Multiple-choice QA	*exactly one of the candidate answers is correct.*
Extractive QA	*answer is a subset of a given paragraph*
Abstractive QA	

Our progress in QA: the bad



format	assumption
Yes/No QA	*binary output*
Multiple-choice QA	*exactly one of the candidate answers is correct.*
Extractive QA	*answer is a subset of a given paragraph*
Abstractive QA	*answer is a mixture of what is given and items not given.*

Our progress in QA: the bad

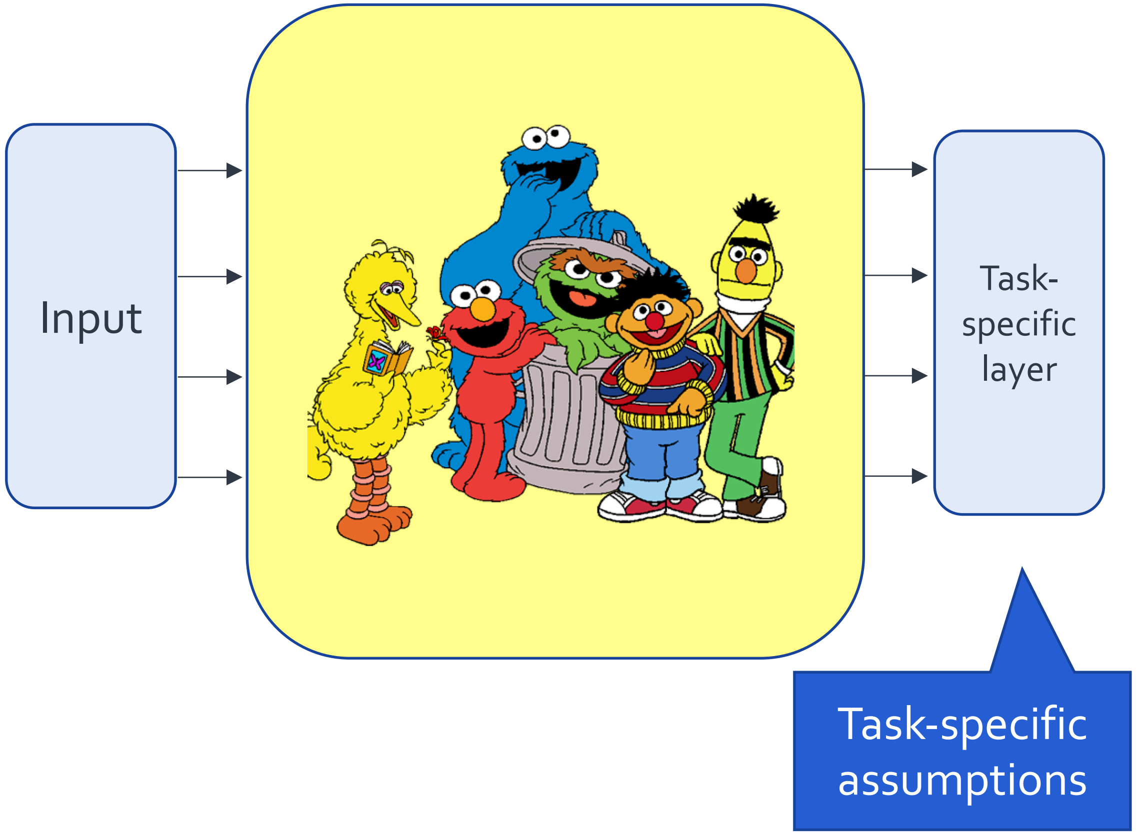


Consequences of format-specific design:

- Prevent generalization across formats
- Don't benefit from labeled data of other formats.

format	assumption
Yes/No QA	*binary output*
Multiple-choice QA	*exactly one of the candidate answers is correct.*
Extractive QA	*answer is a subset of a given paragraph*
Abstractive QA	*answer is a mixture of what is given and items not given.*

Our progress in QA: the bad

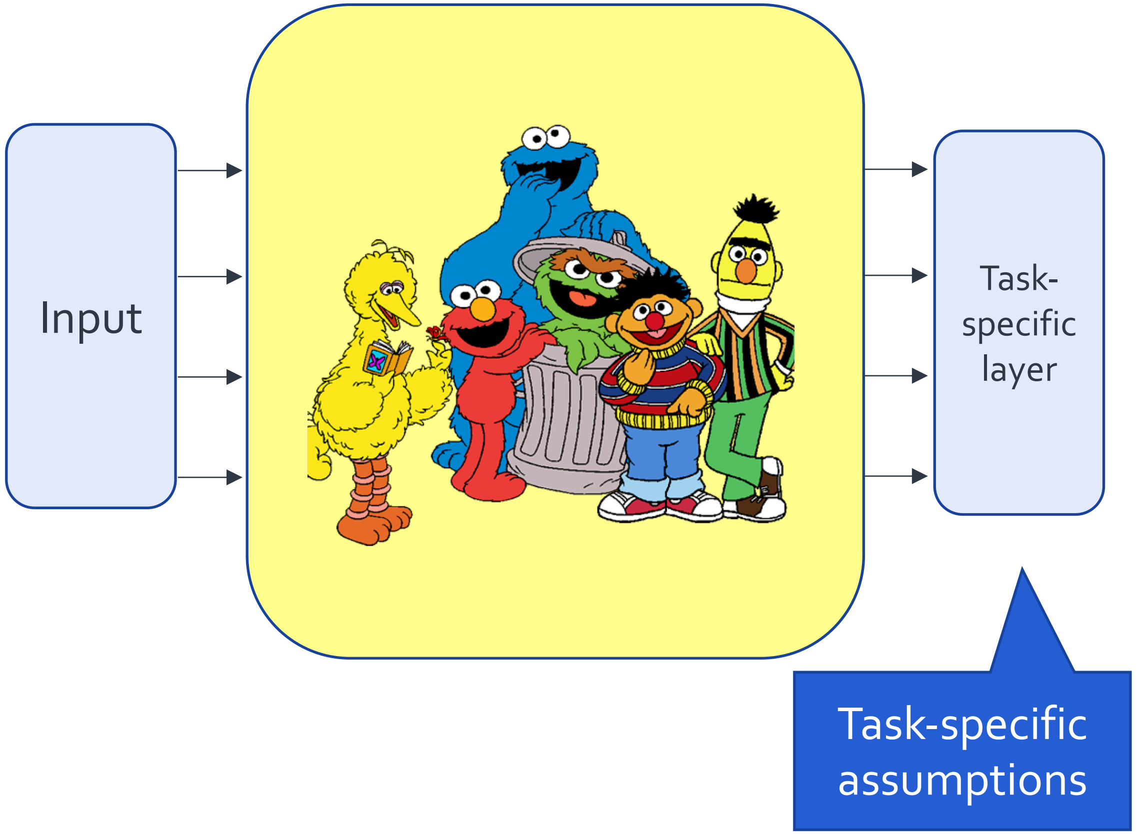


Consequences of format-specific design:

- Prevent generalization across formats
- Don't benefit from labeled data of other formats.

format	assumption
Yes/No QA	*binary output*
Multiple-choice QA	*exactly one of the candidate answers is correct.*
Extractive QA	*answer is a subset of a given paragraph*
Abstractive QA	*answer is a mixture of what is given and items not given.*

Our progress in QA: the bad



Consequences of format-specific design:

- Prevent generalization across formats
- Don't benefit from labeled data of other formats.

format	assumption
Yes/No QA	*binary output*
Multiple-choice QA	*exactly one of the candidate answers is correct.*
Extractive QA	*answer is a subset of a given paragraph*
Abstractive QA	*answer is a mixture of what is given and items not given.*

formats-specialized models

ExtractiveQA

MultipleChoiceQA

formats-specialized models

ExtractiveQA

Question: *"At what speed did the turbine operate?"*

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

"16,000 rpm"

MultipleChoiceQA

formats-specialized models

ExtractiveQA

Question: *"At what speed did the turbine operate?"*

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

"16,000 rpm"

MultipleChoiceQA

Question: *"What does photosynthesis produce that helps plants grow?"*

(A) water
(B) oxygen
(C) protein
(D) sugar

"sugar"

formats-specialized models

ExtractiveQA

Question: *"At what speed did the turbine operate?"*

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

"16,000 rpm"

MultipleChoiceQA

Question: *"What does photosynthesis produce that helps plants grow?"*

(A) water
(B) oxygen
(C) protein
(D) sugar

"sugar"

formats-specialized models

ExtractiveQA

Question: *"At what speed did the turbine operate?"*

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

"16,000 rpm"

MultipleChoiceQA

Question: *"What does photosynthesis produce that helps plants grow?"*

(A) water
(B) oxygen
(C) protein
(D) sugar

"sugar"

formats-specialized models

ExtractiveQA

Question: *"At what speed did the turbine operate?"*

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

"16,000 rpm"

MultipleChoiceQA

Question: *"What does photosynthesis produce that helps plants grow?"*

(A) water
(B) oxygen
(C) protein
(D) sugar

"sugar"

formats-specialized models

ExtractiveQA

Question: *"At what speed did the turbine operate?"*

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

"16,000 rpm"

MultipleChoiceQA

Question: *"What does photosynthesis produce that helps plants grow?"*

(A) water
(B) oxygen
(C) protein
(D) sugar

"sugar"

Beyond formats-specialized models

ExtractiveQA

Question: *"At what speed did the turbine operate?"*

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

"16,000 rpm"

MultipleChoiceQA

Question: *"What does photosynthesis produce that helps plants grow?"*

(A) water
(B) oxygen
(C) protein
(D) sugar

"sugar"

Beyond formats-specialized models

ExtractiveQA

Question: *"At what speed did the turbine operate?"*

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

MultipleChoiceQA

Question: *"What does photosynthesis produce that helps plants grow?"*

(A) water
(B) oxygen
(C) protein
(D) sugar

"16,000 rpm"

"sugar"

Beyond formats-specialized models

ExtractiveQA

Question: *"At what speed did the turbine operate?"*

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

"16,000 rpm"

MultipleChoiceQA

Question: *"What does photosynthesis produce that helps plants grow?"*

(A) water
(B) oxygen
(C) protein
(D) sugar

"sugar"

Beyond formats-specialized models

ExtractiveQA

Question: *"At what speed did the turbine operate?"*

(Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

"16,000 rpm"

MultipleChoiceQA

Question: *"What does photosynthesis produce that helps plants grow?"*

(A) water
(B) oxygen
(C) protein
(D) sugar

"sugar"

Talk Summary & Statement

Talk Summary & Statement

- Creating **format-specific QA** models distance us from broad QA.
- There is **overlap** between underlying reasoning abilities of formats.
  - One can benefit from **mixing** QA formats.
- UnifiedQA: a single QA system working across four common QA formats.
  - Fine-tuning models pre-trained on UnifiedQA yields SOTA results.

Talk Summary & Statement

- Creating **format-specific QA** models distance us from broad QA.
- There is **overlap** between underlying reasoning abilities of formats.
  - One can benefit from **mixing** QA formats.
- UnifiedQA: a single QA system working across four common QA formats.
  - Fine-tuning models pre-trained on UnifiedQA yields SOTA results.

Talk Summary & Statement

- Creating **format-specific QA** models distance us from broad QA.

- There is **overlap** between underlying reasoning abilities of formats.
  - One can benefit from **mixing** QA formats.

- UnifiedQA: a single QA system working across four common QA formats.
  - Fine-tuning models pre-trained on UnifiedQA yields SOTA results.

Earlier work on multi-task learning

Earlier work on multi-task learning

- In the same spirit as multi-task learning. [Caruana'97; McCann et al'18]

- The choice of tasks is also important.
  - Earlier works select too broad of tasks.
    - E.g., Raffel et al'19 diverse NLP tasks (machine translation, summarization, etc) and conclude that a single model for multiple NLP tasks underperform task-specific models.

- We narrow the scope of tasks to stay within the boundaries of QA.
  - No task/format specific encoding.

Earlier work on multi-task learning

- In the same spirit as multi-task learning. [Caruana'97; McCann et al'18]

Didn't work before; why would it work now? 🤔

- The choice of tasks is also important.
  - Earlier works select too broad of tasks.
    - E.g., Raffel et al'19 diverse NLP tasks (machine translation, summarization, etc) and conclude that a single model for multiple NLP tasks underperform task-specific models.
- We narrow the scope of tasks to stay within the boundaries of QA.
  - No task/format specific encoding.

Earlier work on multi-task learning

- In the same spirit as multi-task learning. [Caruana'97; McCann et al'18]

Didn't work before; why would it work now? 🤔

- The choice of tasks is also important.
  - Earlier works select too broad of tasks.
    - E.g., Raffel et al'19 diverse NLP tasks (machine translation, summarization, etc) and conclude that a single model for multiple NLP tasks underperform task-specific models.

- We narrow the scope of tasks to stay within the boundaries of QA.
  - No task/format specific encoding.

Earlier work on multi-task learning

- In the same spirit as multi-task learning. [Caruana'97; McCann et al'18]

Didn't work before; why would it work now? 🤔

- The choice of tasks is also important.
  - Earlier works select too broad of tasks.
    - E.g., Raffel et al'19 diverse NLP tasks (machine translation, summarization, etc) and conclude that a single model for multiple NLP tasks underperform task-specific models.
- We narrow the scope of tasks to stay within the boundaries of QA.
  - No task/format specific encoding.

Earlier work on multi-task learning

- In the same spirit as multi-task learning. [Caruana'97; McCann et al'18]

Didn't work before; why would it work now? 🤔

- The choice of tasks is also important.
  - Earlier works select too broad of tasks.
    - E.g., Raffel et al'19 diverse NLP tasks (machine translation, summarization, etc) and conclude that a single model for multiple NLP tasks underperform task-specific models.

- We narrow the scope of tasks to stay within the boundaries of QA.
  - No task/format specific encoding.

Earlier work on multi-task learning

- In the same spirit as multi-task learning. [Caruana'97; McCann et al'18]

Didn't work before; why would it work now? 🤔

- The choice of tasks is also important.
  - Earlier works select too broad of tasks.
    - E.g., Raffel et al'19 diverse NLP tasks (machine translation, summarization, etc) and conclude that a single model for multiple NLP tasks underperform task-specific models.
- We narrow the scope of tasks to stay within the boundaries of QA.
  - No task/format specific encoding.

Roadmap

1. Generalization across formats

2. UnifiedQA + Empirical Intuitions

3. Discussion and next steps

Roadmap

1. **Generalization across formats**
2. UnifiedQA + Empirical Intuitions
3. Discussion and next steps

UnifiedQA: a high-level definition

UnifiedQA: a high-level definition

1. It’s a single system that is supposed to work on a variety of **QA formats**.
2. The input should be *natural*.
   - Minimal-enough for a human solver to infer the format.

UnifiedQA: a high-level definition

1. It’s a single system that is supposed to work on a variety of **QA formats**.
2. The input should be *natural*.
   - Minimal-enough for a human solver to infer the format.

UnifiedQA: a high-level definition

1. It's a single system that is supposed to work on a variety of **QA formats**.
2. The input should be *natural*.
   - Minimal-enough for a human solver to infer the format.

```
"What causes sound?

(A) sunlight (B) vibrations (C) x-rays (D) pitch"
```

UnifiedQA: a high-level definition

1. It's a single system that is supposed to work on a variety of **QA formats**.
2. The input should be *natural*.
   - Minimal-enough for a human solver to infer the format.

```
"What causes sound?

(A) sunlight (B) vibrations (C) x-rays (D) pitch"
```

```
"vibrations"
```

UnifiedQA: a high-level definition

1. It’s a single system that is supposed to work on a variety of **QA formats**.
2. The input should be *natural*.
   - Minimal-enough for a human solver to infer the format.

```
“Is Jamaica considered part of the United States?

(Jamaica) Jamaica (/dʒəˈmeɪkə/ ( listen)) is an island
country situated in the Caribbean Sea. Spanning 10,990
square kilometres (4,240 sq mi) in area, it is the
third-largest island of the Greater Antilles and the
fourth-largest island country in the Caribbean.”
```

UnifiedQA: a high-level definition

1. It's a single system that is supposed to work on a variety of **QA formats**.
2. The input should be *natural*.
   - Minimal-enough for a human solver to infer the format.

```
"Is Jamaica considered part of the United States?

(Jamaica) Jamaica (/dʒəˈmeɪkə/ ( listen)) is an island
country situated in the Caribbean Sea. Spanning 10,990
square kilometres (4,240 sq mi) in area, it is the
third-largest island of the Greater Antilles and the
fourth-largest island country in the Caribbean."
```

"no"

UnifiedQA: a high-level definition

1. It's a single system that is supposed to work on a variety of **QA formats**.
2. The input should be *natural*.
   - Minimal-enough for a human solver to infer the format.

```
"What type of musical instruments did the Yuan bring to China?

(Yuan_dynasty) Western musical instruments were introduced to
enrich Chinese performing arts. From this period dates the
conversion to Islam, by Muslims of Central Asia, of growing
numbers of Chinese in the northwest and southwest. ..."
```

UnifiedQA: a high-level definition

1. It's a single system that is supposed to work on a variety of **QA formats**.
2. The input should be *natural*.
   - Minimal-enough for a human solver to infer the format.

```
"What type of musical instruments did the Yuan bring to China?

(Yuan_dynasty) Western musical instruments were introduced to
enrich Chinese performing arts. From this period dates the
conversion to Islam, by Muslims of Central Asia, of growing
numbers of Chinese in the northwest and southwest. ..."
```

```
"Western musical instruments"
```

UnifiedQA: a high-level definition

1. It's a single system that is supposed to work on a variety of **QA formats**.
2. The input should be *natural*.
   - Minimal-enough for a human solver to infer the format.

- *The question always comes first.*
- *Additional info are appended with "\n".*

```
"What type of musical instruments did the Yuan bring to China?

(Yuan_dynasty) Western musical instruments were introduced to
enrich Chinese performing arts. From this period dates the
conversion to Islam, by Muslims of Central Asia, of growing
numbers of Chinese in the northwest and southwest. ..."
```

```
"Western musical instruments"
```

UnifiedQA: towards an implementation

UnifiedQA: towards an implementation

- Use text-to-text architectures
  - T5 [Raffal et al, 2020], BART [Lewis et al, 2019], etc.

- Train simultaneously on all datasets jointly together.
  - Batches contains the same number of instances from each training set.

UnifiedQA: towards an implementation

- Use text-to-text architectures
  - T5 [Raffal et al, 2020], BART [Lewis et al, 2019], etc.

- Train simultaneously on all datasets jointly together.
  - Batches contains the same number of instances from each training set.

UnifiedQA: towards an implementation

- Use text-to-text architectures
  - T5 [Raffal et al, 2020], BART [Lewis et al, 2019], etc.

- Train simultaneously on all datasets jointly together.
  - Batches contains the same number of instances from each training set.

Mixing pairs of formats: experiment (1)

- Is there any value in out-of-format training?

Mixing pairs of formats: experiment (1)

- Is there any value in out-of-format training?

Mixing RACE (Multiple-Choice)

w/ datasets of different formats.

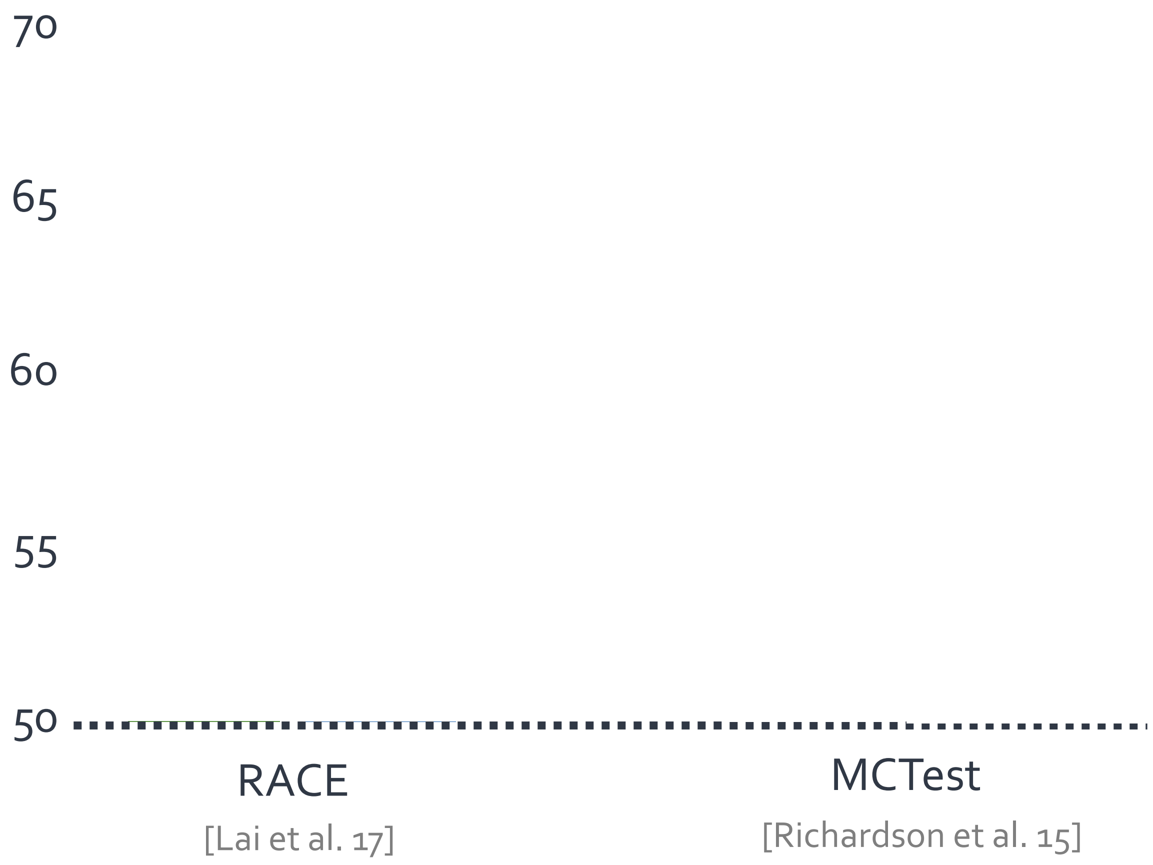
Mixing pairs of formats: experiment (1)

- Is there any value in out-of-format training?

Mixing RACE (Multiple-Choice)

w/ datasets of different formats.

Trained on RACE



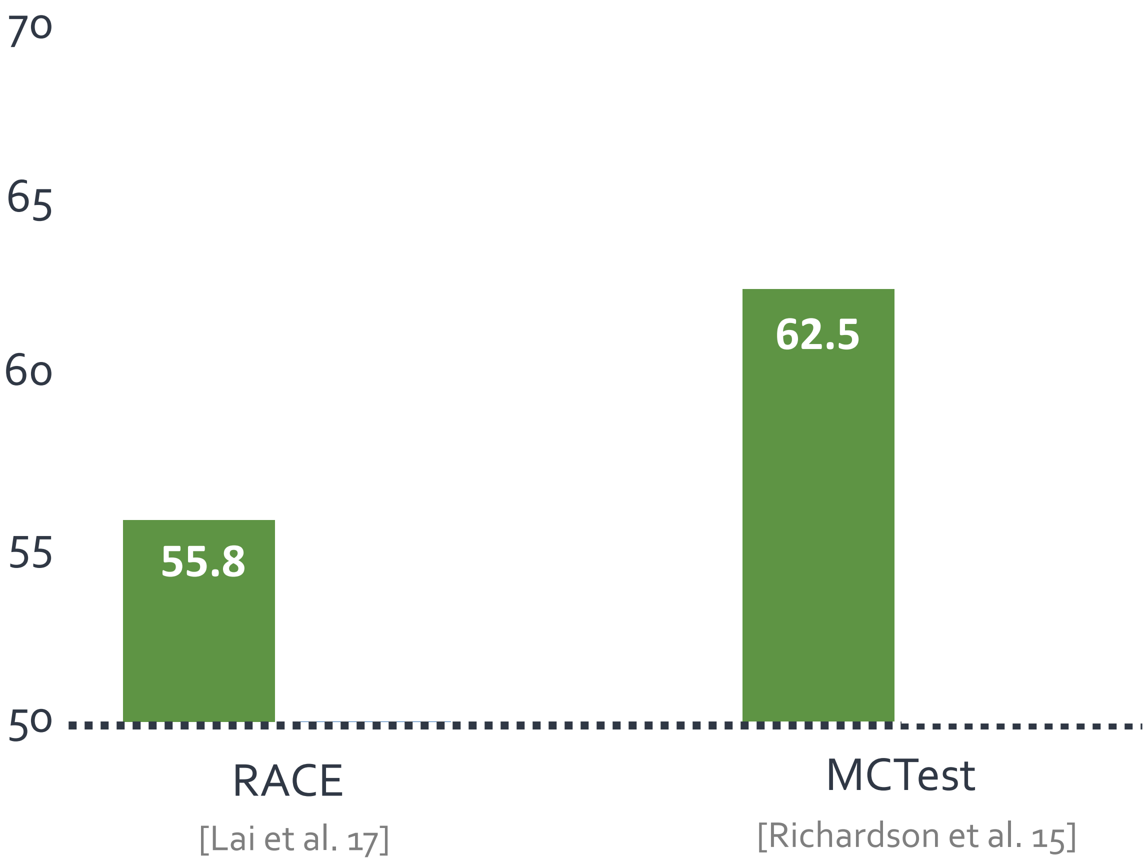
Mixing pairs of formats: experiment (1)

- Is there any value in out-of-format training?

Mixing RACE (Multiple-Choice)

w/ datasets of different formats.

Trained on RACE

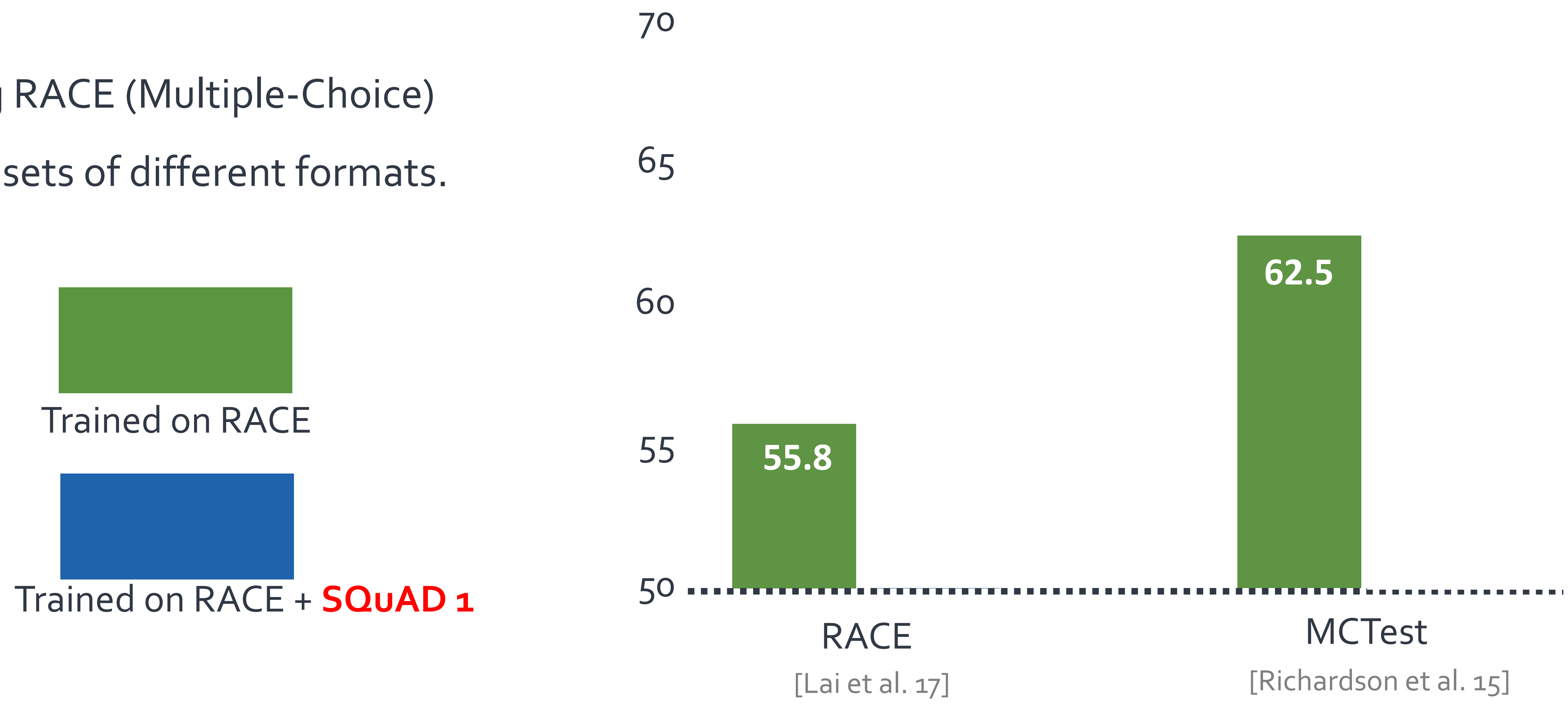


Mixing pairs of formats: experiment (1)

- Is there any value in out-of-format training?

Mixing RACE (Multiple-Choice)

w/ datasets of different formats.

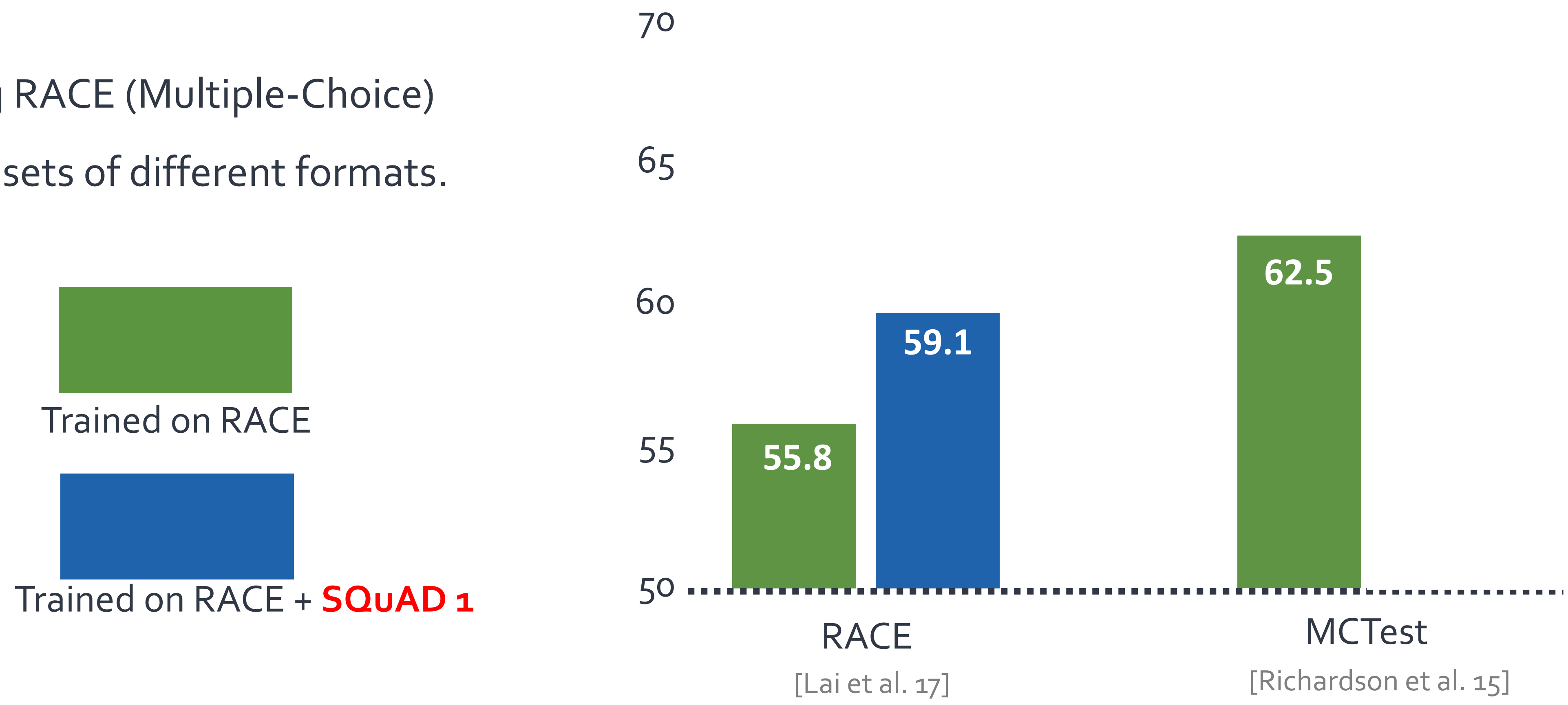


Mixing pairs of formats: experiment (1)

- Is there any value in out-of-format training?

Mixing RACE (Multiple-Choice)

w/ datasets of different formats.



Mixing pairs of formats: experiment (1)

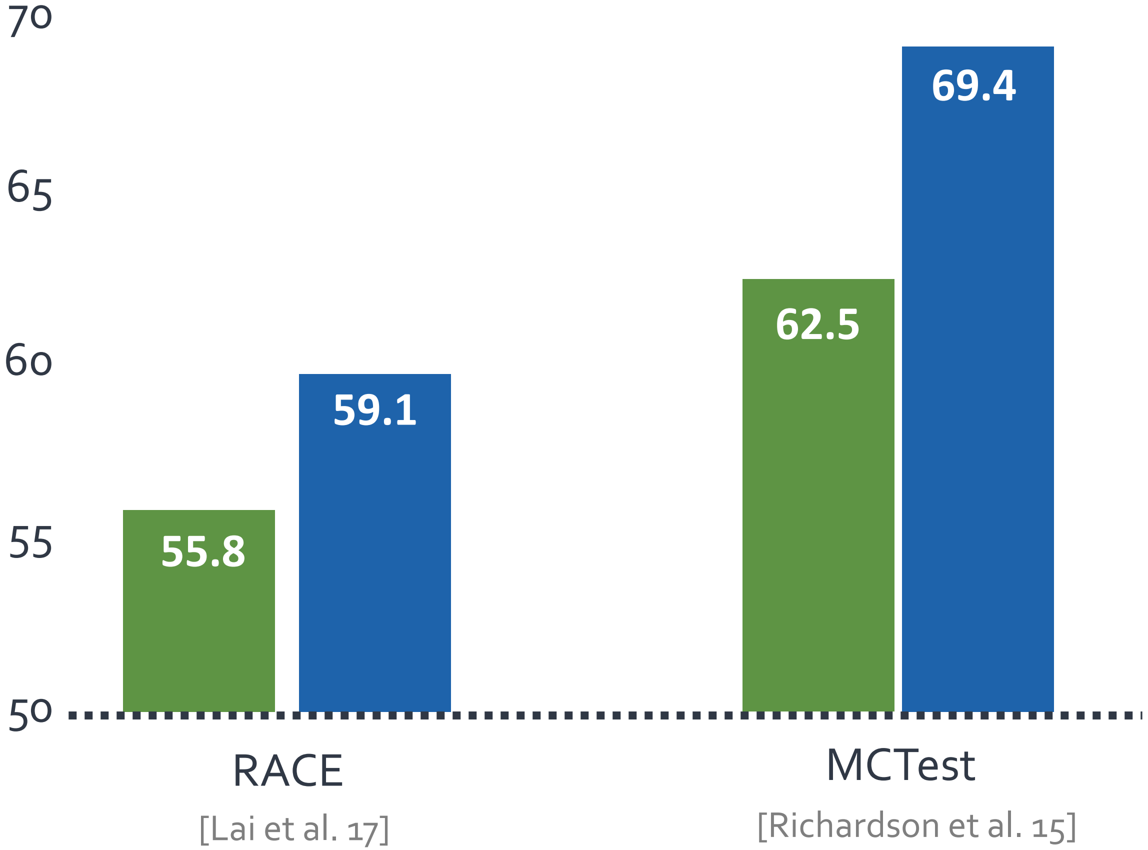
- Is there any value in out-of-format training?

Mixing RACE (Multiple-Choice)

w/ datasets of different formats.

Trained on RACE

Trained on RACE + **SQuAD 1**



Mixing pairs of formats: experiment (2)

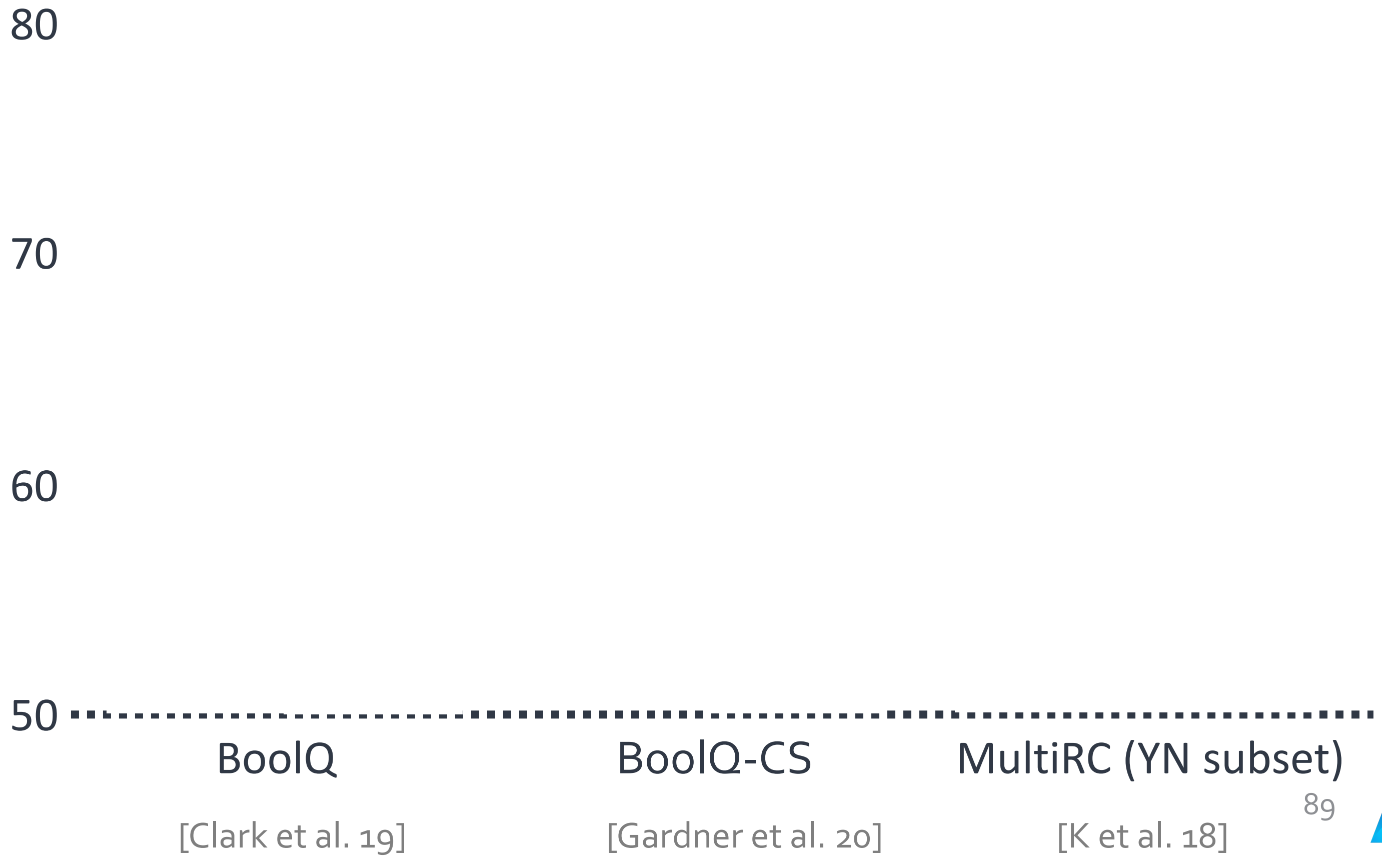
- Is there any value in out-of-format training?

Mixing BoolQ (YesNo)

w/ datasets of different formats.

Trained on BoolQ

Trained on BoolQ + X

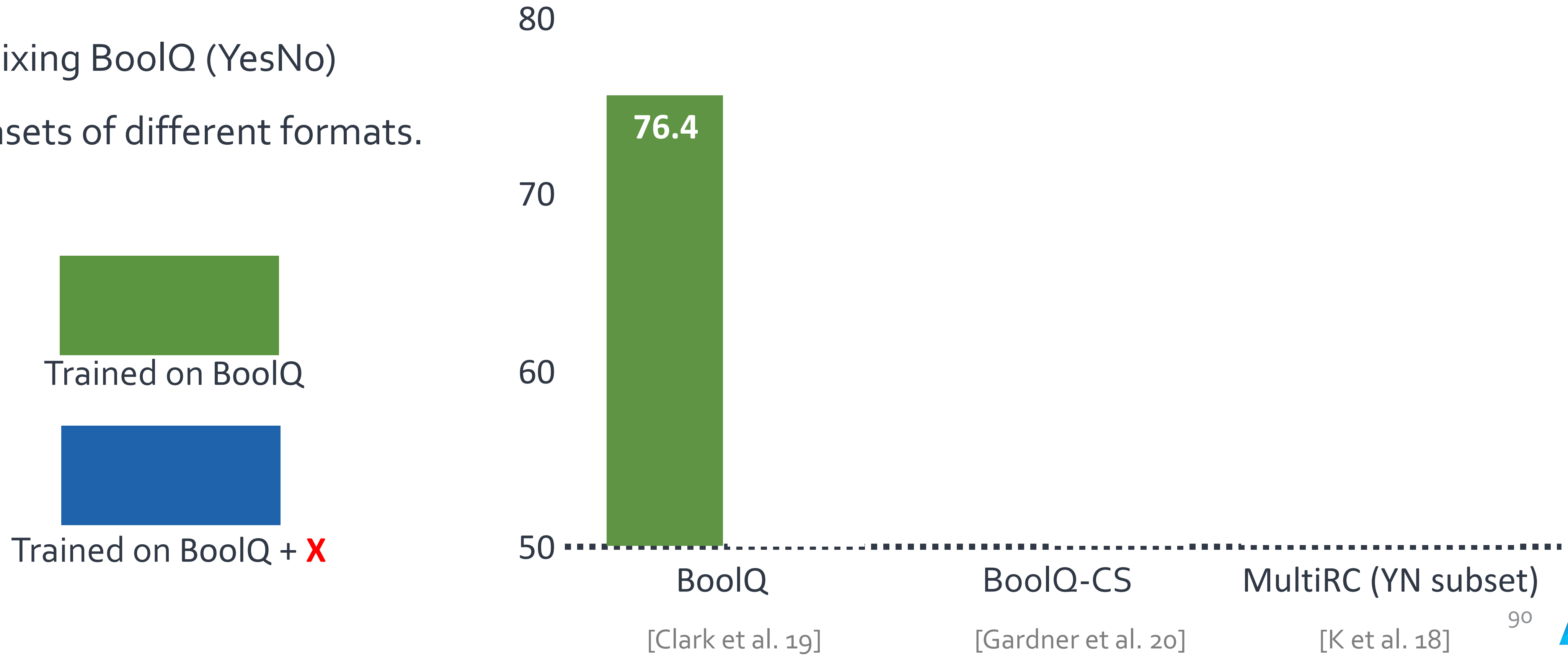


Mixing pairs of formats: experiment (2)

- Is there any value in out-of-format training?

Mixing BoolQ (YesNo)

w/ datasets of different formats.

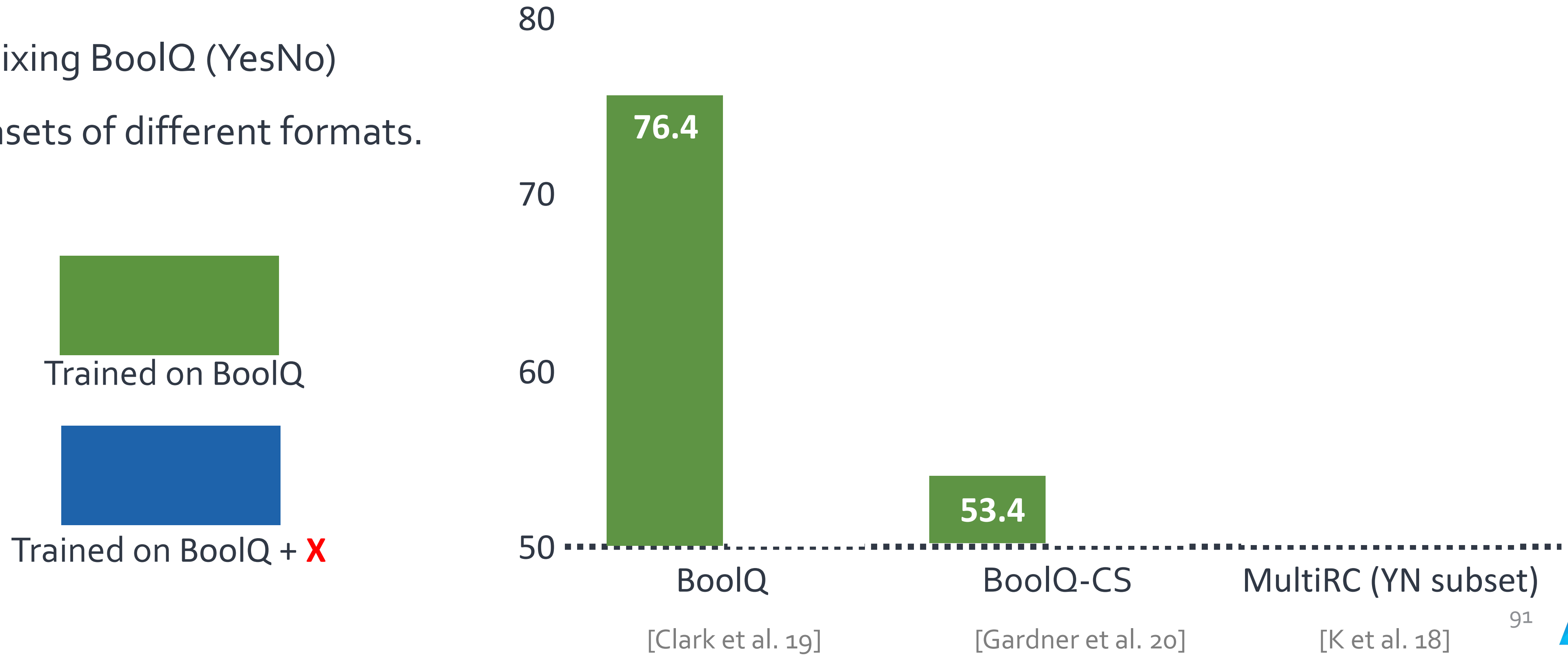


Mixing pairs of formats: experiment (2)

- Is there any value in out-of-format training?

Mixing BoolQ (YesNo)

w/ datasets of different formats.

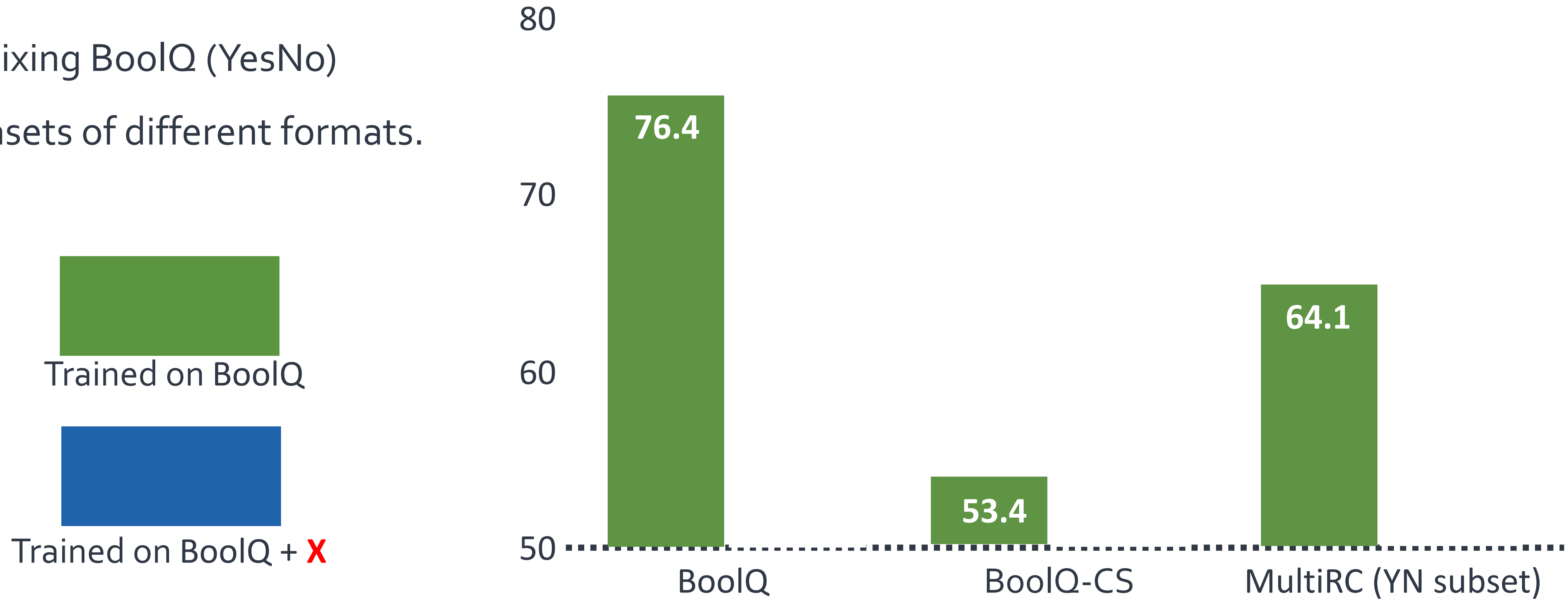


Mixing pairs of formats: experiment (2)

- Is there any value in out-of-format training?

Mixing BoolQ (YesNo)

w/ datasets of different formats.



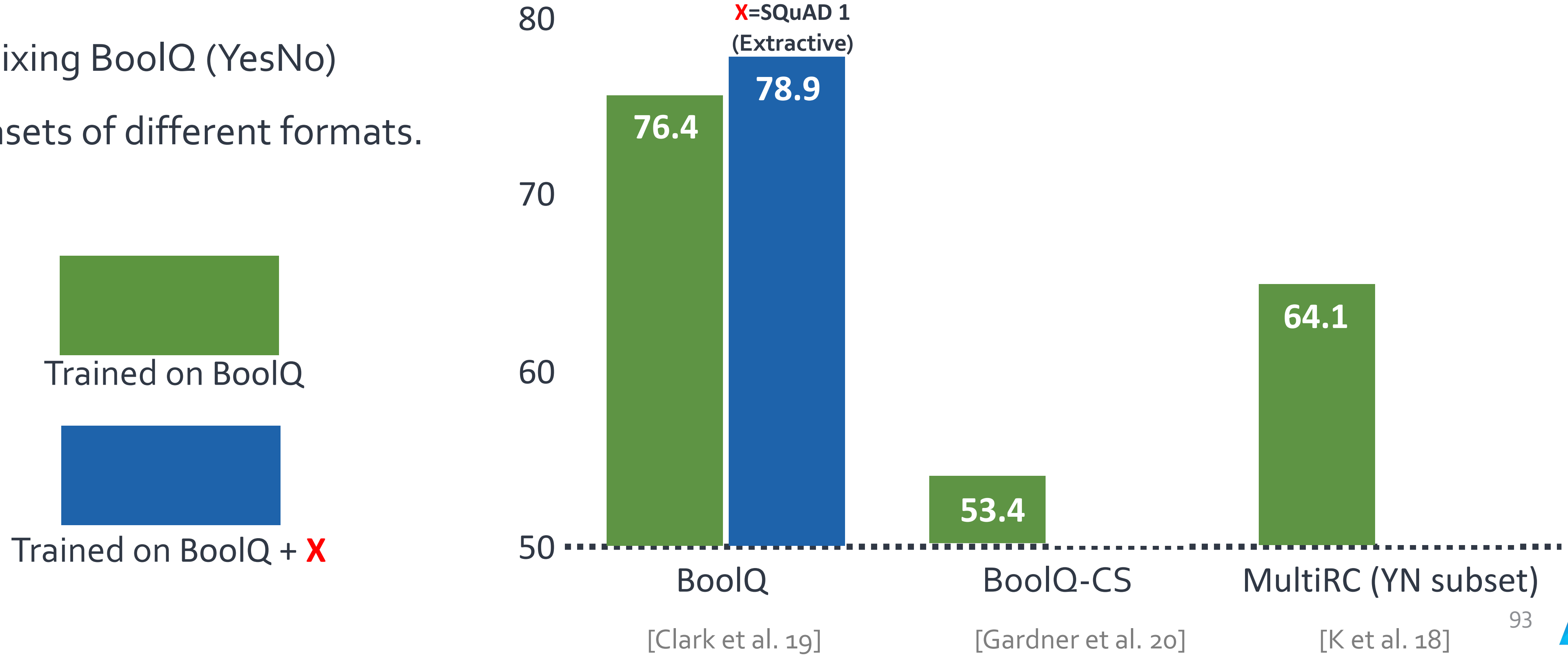
[Clark et al. 19] [Gardner et al. 20] [K et al. 18]

Mixing pairs of formats: experiment (2)

- Is there any value in out-of-format training?

Mixing BoolQ (YesNo)

w/ datasets of different formats.

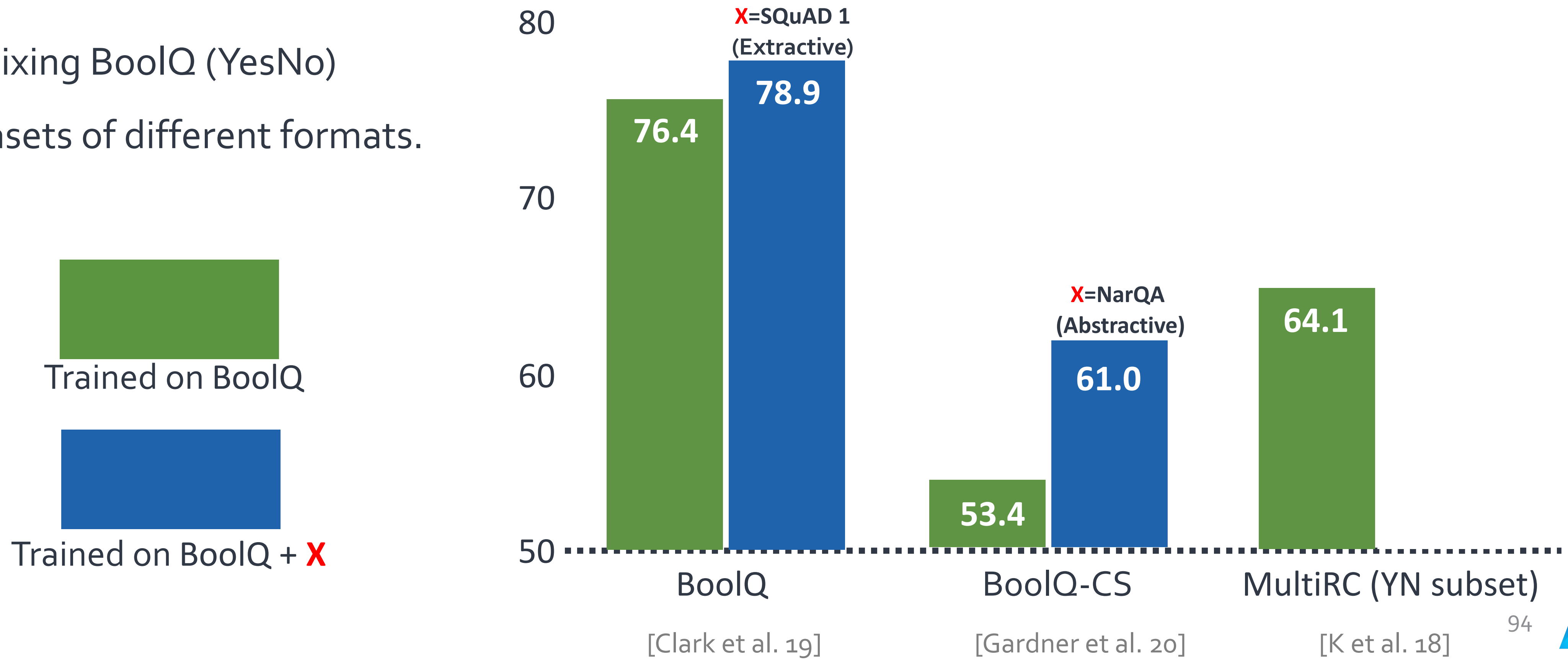


Mixing pairs of formats: experiment (2)

- Is there any value in out-of-format training?

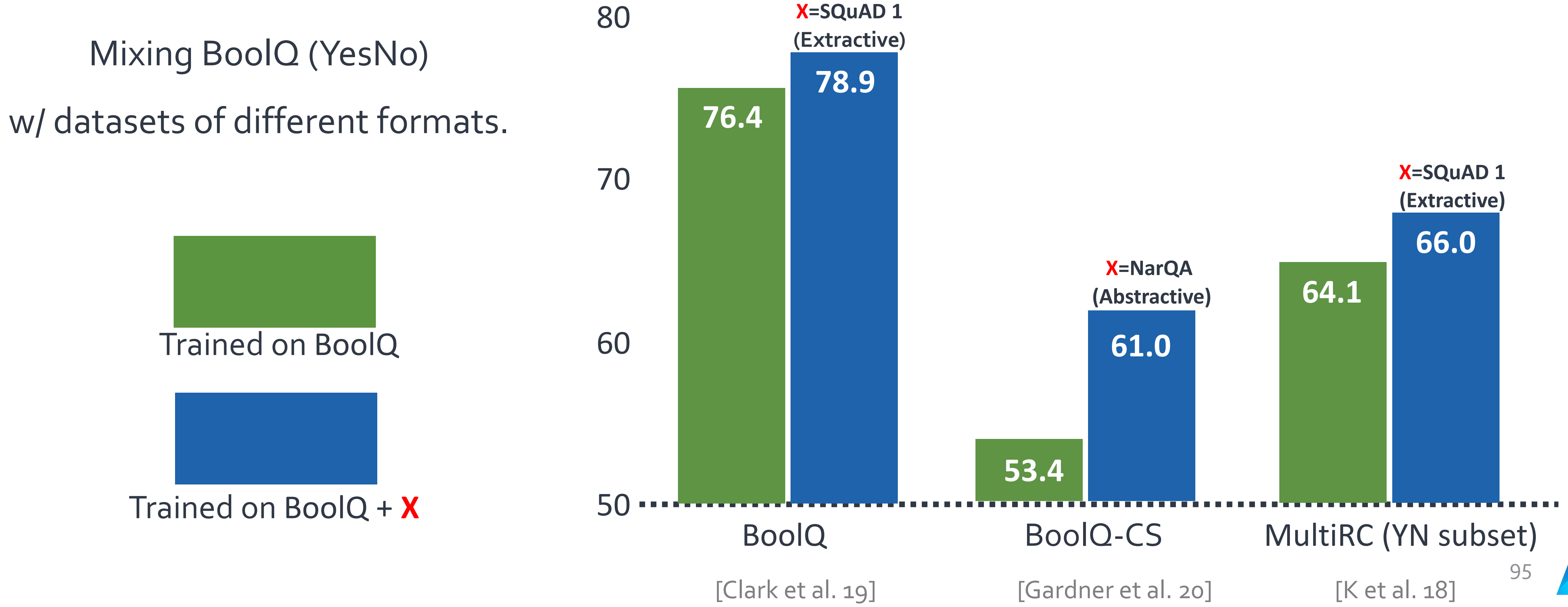
Mixing BoolQ (YesNo)

w/ datasets of different formats.



Mixing pairs of formats: experiment (2)

- Is there any value in out-of-format training?



Roadmap

1. **Generalization across formats**
2. UnifiedQA + Empirical Intuitions
3. Discussion and next steps

Roadmap

1. Generalization across formats
2. **UnifiedQA + Empirical Intuitions**
3. Discussion and next steps

UnifiedQA-v1

UnifiedQA-v1

- Trained on the union of different formats:
  - **Extractive:** SQuAD 1.1, SQuAD 2.0
  - **Abstractive:** NarrativeQA
  - **Multiple-choice:** RACE, ARC, OBQA, MCTest
  - **YesNo:** BoolQ

- Architectures:
  - T5 (11B, 3B, ...)
  - BART (large)

UnifiedQA-v1

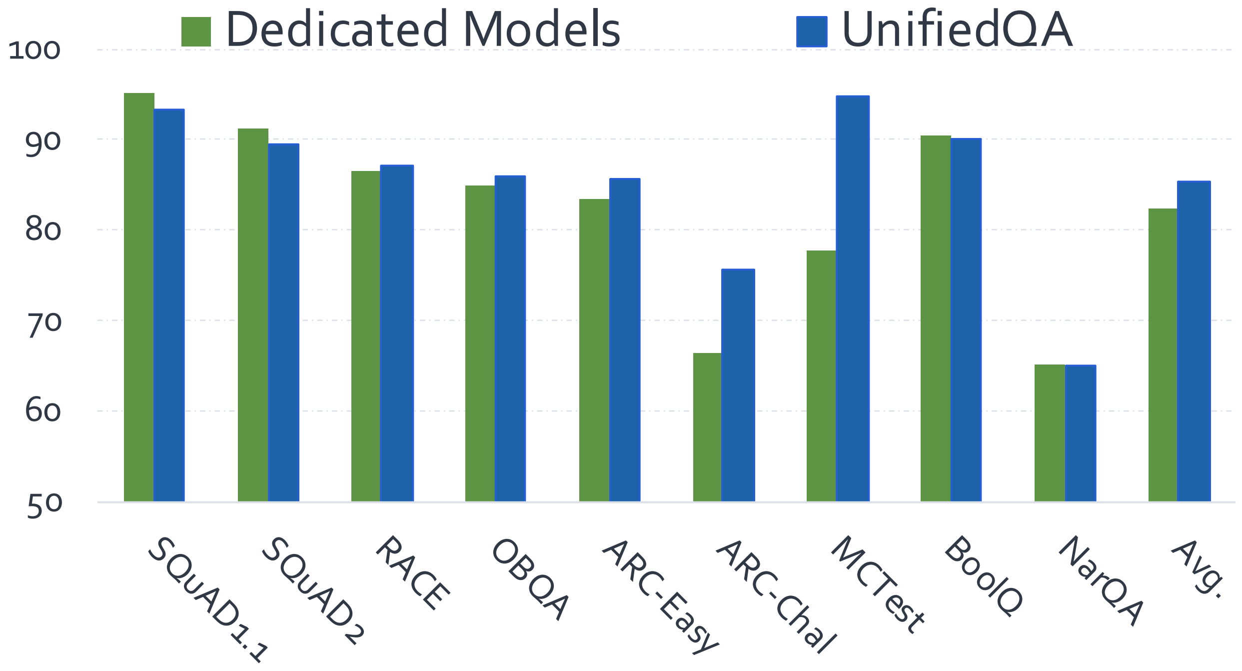
- Trained on the union of different formats:
  - **Extractive:** SQuAD 1.1, SQuAD 2.0
  - **Abstractive:** NarrativeQA
  - **Multiple-choice:** RACE, ARC, OBQA, MCTest
  - **YesNo:** BoolQ

- Architectures:
  - T5 (11B, 3B, …)
  - BART (large)

https://github.com/allenai/unifiedqa

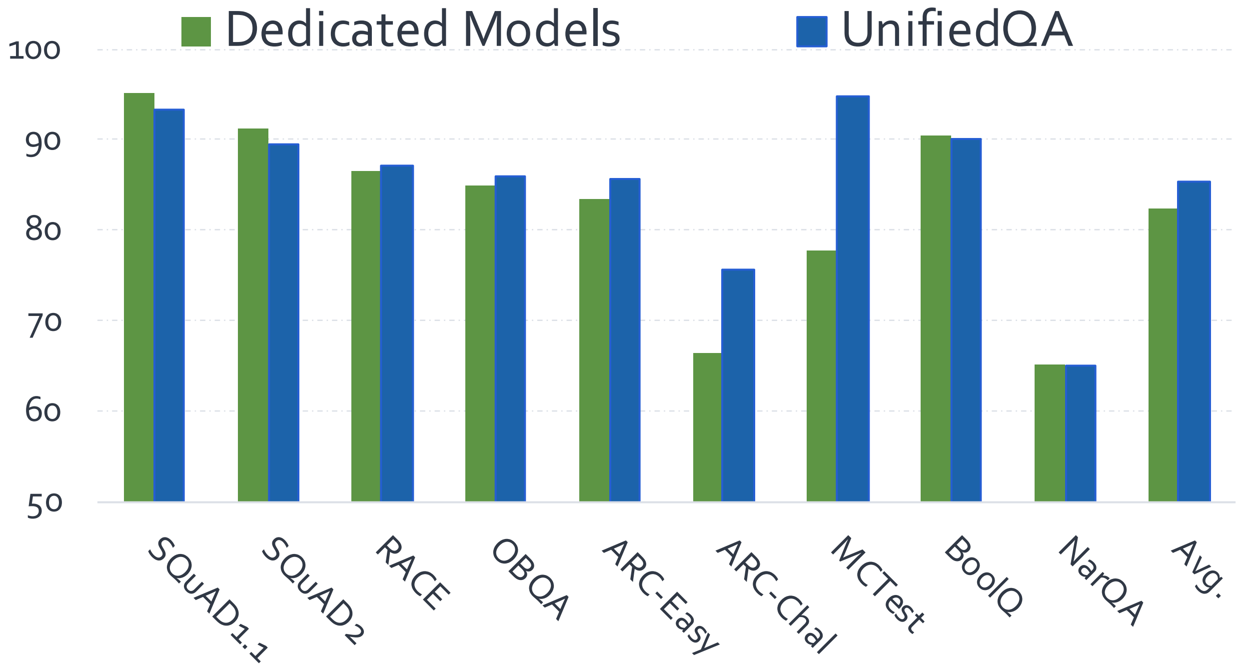
Intuition #1: Comparison w/ Dedicated Models

Intuition #1: Comparison w/ Dedicated Models



Intuition #1: Comparison w/ Dedicated Models

- Is UnifiedQA as good as systems dedicated to individual datasets?



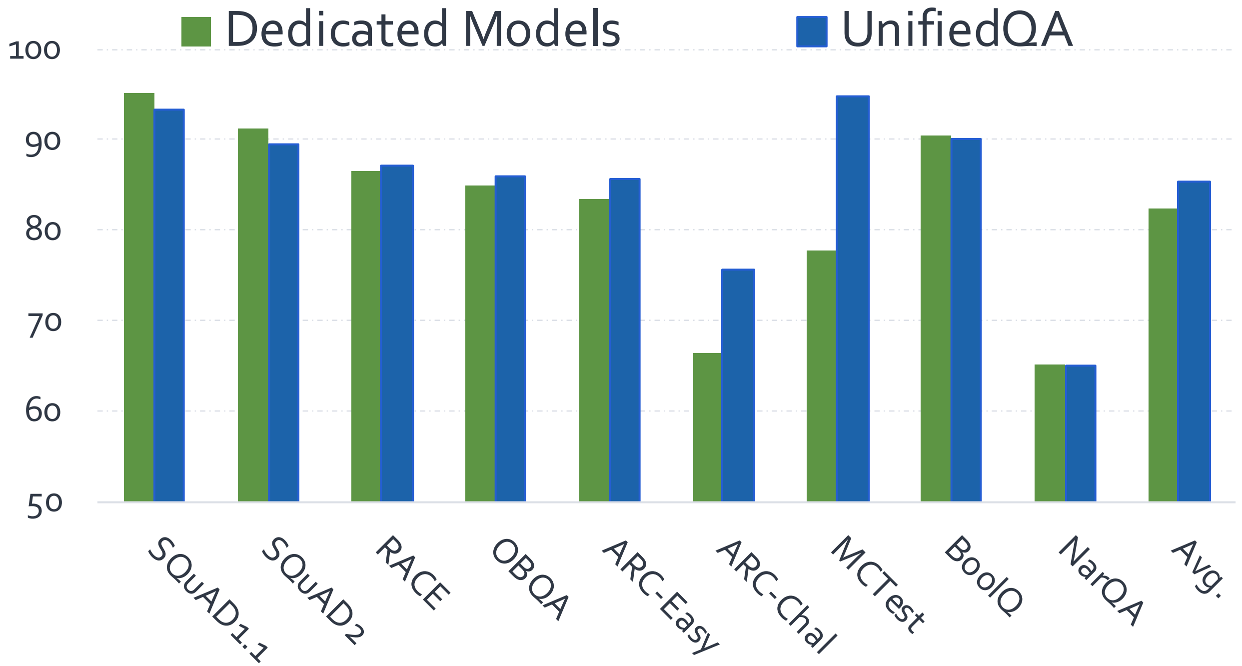
- UnifiedQA performs almost as good as individual T5 models targeted to each dataset.

Intuition #1: Comparison w/ Dedicated Models

- Is UnifiedQA as good as systems dedicated to individual datasets?

evaluation sets

	SQuAD2	RACE	BoolQ	NarQA
T5 (SQuAD 2)	**91**	33	12	51
T5 (RACE)	43	**87**	7	54
T5 (BoolQ)	4	22	**90**	0
T5 (NarQA)	45	48	47	**65**
UnifiedQA	90	**87**	**90**	**65**



- UnifiedQA performs almost as good as individual T5 models targeted to each dataset.

Intuition #2: Unseen Datasets

Intuition #2: Unseen Datasets

evaluation sets

models trained for individual formats

	NewsQA	Quoref	DROP	DROP-CS	QASC	CommonsenseQA	NP-BoolQ	BoolQ-CS	Avg
UnifiedQA [EX]	59	65	25	24	55	63	21	13	42
UnifiedQA [AB]	58	68	31	37	54	59	27	40	48
UnifiedQA [MC]	48	**68**	29	37	68	76	3	6	44
UnifiedQA [YN]	1	2	0	0	15	21	79	79	22
UnifiedQA	**59**	63	**33**	**40**	**68**	**76**	**81**	**80**	**62**

Intuition #2: Unseen Datasets

- Does UnifiedQA generalizes well to unseen datasets?

evaluation sets

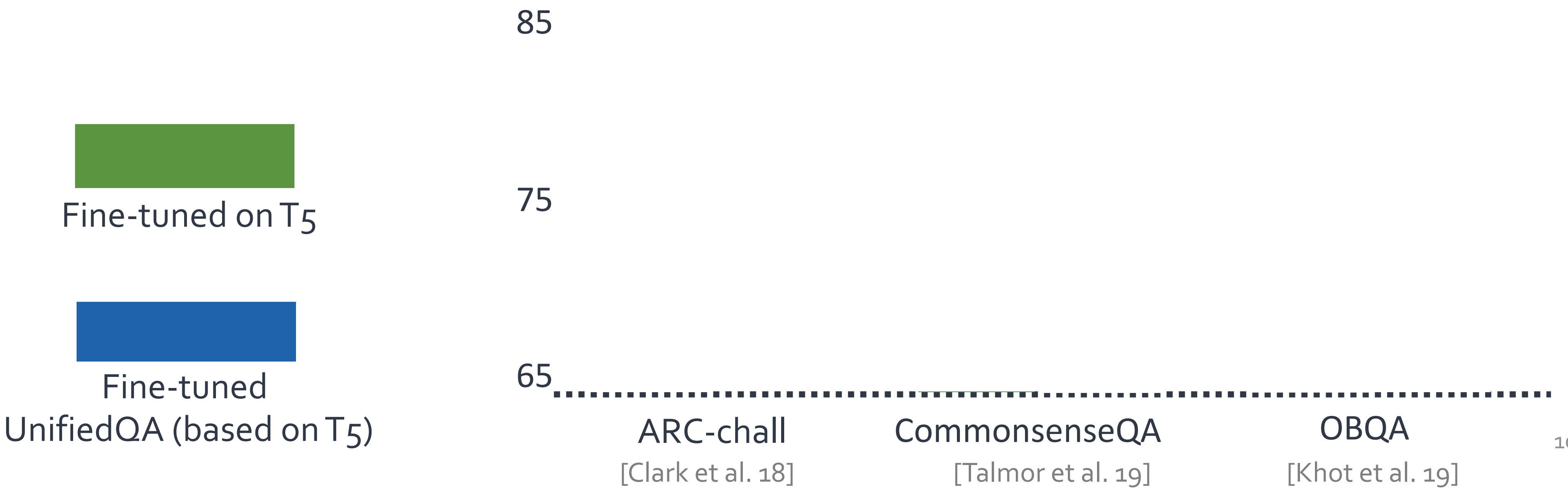
models trained for individual formats

	NewsQA	Quoref	DROP	DROP-CS	QASC	CommonsenseQA	NP-BoolQ	BoolQ-CS	Avg
UnifiedQA [EX]	59	65	25	24	55	63	21	13	42
UnifiedQA [AB]	58	68	31	37	54	59	27	40	48
UnifiedQA [MC]	48	**68**	29	37	68	76	3	6	44
UnifiedQA [YN]	1	2	0	0	15	21	79	79	22
UnifiedQA	**59**	63	**33**	**40**	**68**	**76**	**81**	**80**	**62**

- On average, UnifiedQA shows much stronger generalization across a wide range of datasets.

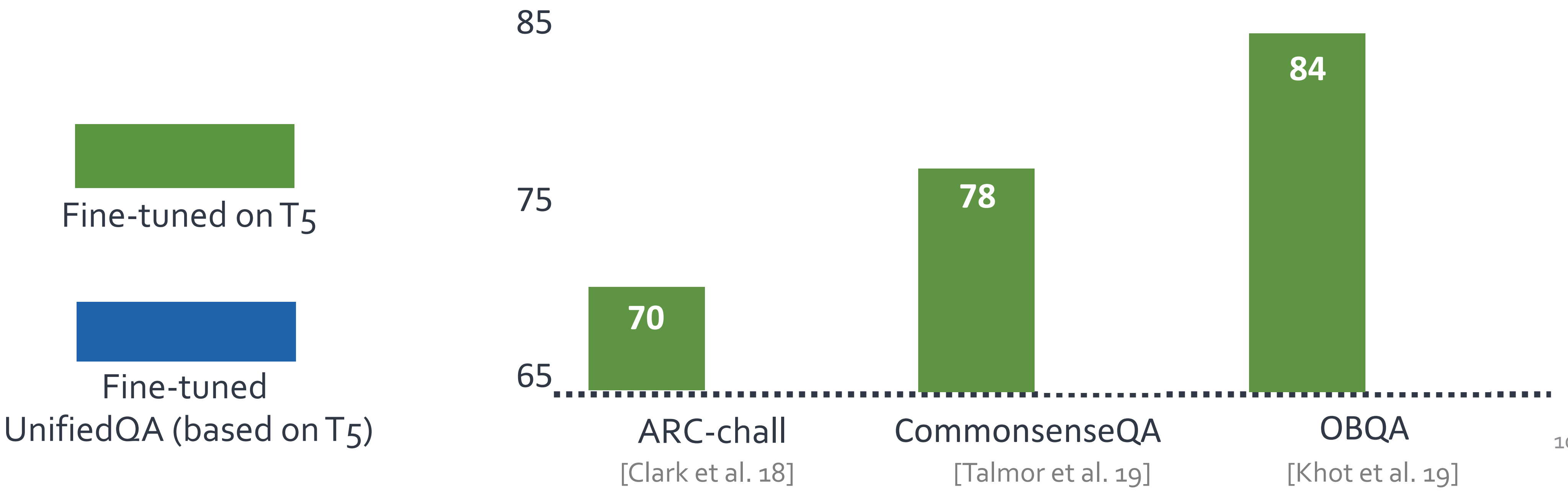
Fine-tuning on UnifiedQA

- Is there a value in using UnifiedQA as a starting point for fine-tuning?
  - Show SOTA on 10 datasets (OBQA, QASC, RACE, WinoGrande, PIQA, SIQA, ROPES)
  - Similar trends for BART



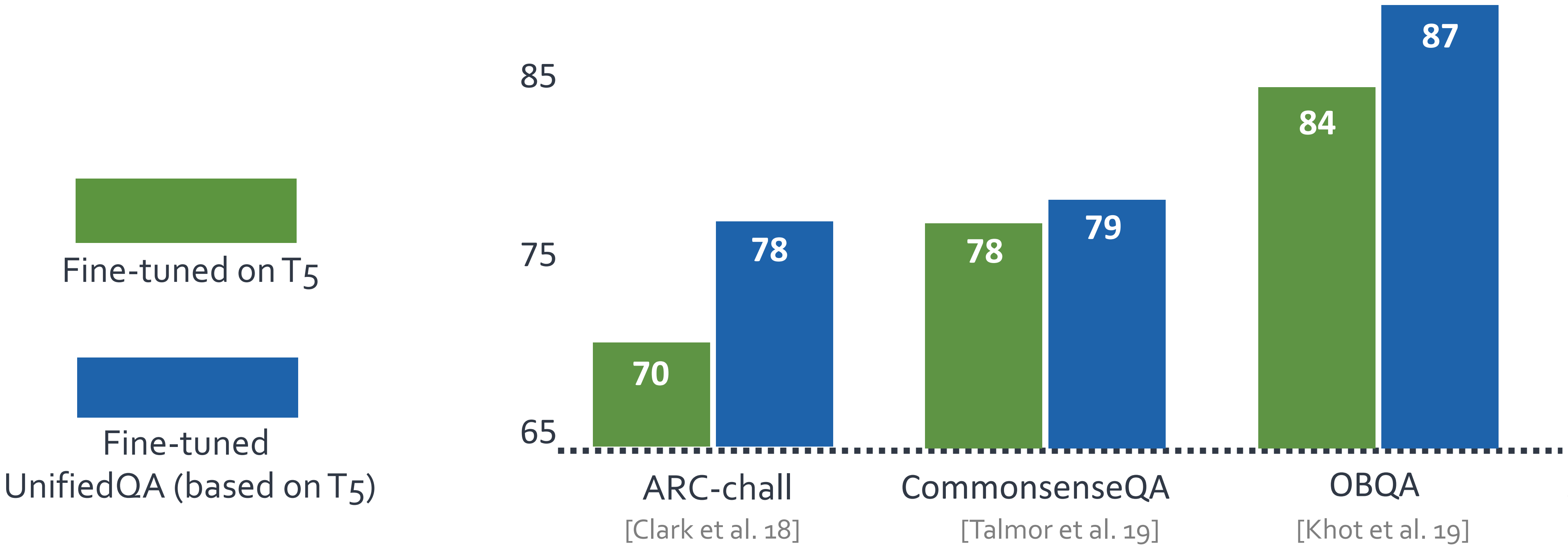
Fine-tuning on UnifiedQA

- Is there a value in using UnifiedQA as a starting point for fine-tuning?
  - Show SOTA on 10 datasets (OBQA, QASC, RACE, WinoGrande, PIQA, SIQA, ROPES)
  - Similar trends for BART



Fine-tuning on UnifiedQA

- Is there a value in using UnifiedQA as a starting point for fine-tuning?
  - Show SOTA on 10 datasets (OBQA, QASC, RACE, WinoGrande, PIQA, SIQA, ROPES)
  - Similar trends for BART



Demo

https://unifiedqa.apps.allenai.org

Roadmap

1. Generalization across formats
2. **UnifiedQA + Empirical Intuitions**
3. Discussion and next steps

Roadmap

1. Generalization across formats

2. UnifiedQA + Empirical Intuitions

3. **Discussion and next steps**

Methodological Issue: Data Leakage

Methodological Issue: Data Leakage

- *"have you done some studies on overlap across datasets?"*

  - **Easy answer:**
    - not much surface-form overlap between the datasets.

  - **Nuanced/ difficult answer:**
    - more data (especially during pre-training) increases the chances of (indirect) leakage.

Methodological Issue: Data Leakage

- *"have you done some studies on overlap across datasets?"*

  - **Easy answer:**
    - not much surface-form overlap between the datasets.

  - **Nuanced/ difficult answer:**
    - more data (especially during pre-training) increases the chances of (indirect) leakage.

Where do we go from here?

Where do we go from here?

- **More formats**
  - Can we incorporate other "natural" variations of QA in the study?
- **Smaller models:**
  - Can we build small and accurate models to make it more available?
- **Beyond QA/Text:**
  - Can you take these ideas and apply it to some other problems?

Where do we go from here?

- **More formats**
  - Can we incorporate other “natural” variations of QA in the study?
- **Smaller models:**
  - Can we build small and accurate models to make it more available?
- **Beyond QA/Text:**
  - Can you take these ideas and apply it to some other problems?

Where do we go from here?

- **More formats**
  - Can we incorporate other "natural" variations of QA in the study?
- **Smaller models:**
  - Can we build small and accurate models to make it more available?
- **Beyond QA/Text:**
  - Can you take these ideas and apply it to some other problems?

Take-home points

- The field relies excessively format-specific assumptions for system design.
  - Instead, we should move towards more general QA architectures.
- **Incentive:** there is value in mixing QA datasets of different formats.
- UnifiedQA, a single pre-trained QA system seeking to bring unification across common QA formats.

https://github.com/allenai/unifiedqa