What Does My QA Model Know? Devising Controlled Probes using Expert Knowledge

Kyle Richardson, Ashish Sabharwal

Allen Institute for Artificial Intelligence (AI2), Seattle WA.

EMNLP 2020 (TACL track)

Probing Natural Language Understanding (NLU) Models

- **Probing**: understanding the strengths/weaknesses of models; **measuring model competence qualitatively**; **behavioral (input/output) testing**.

Probing Natural Language Understanding (NLU) Models

- **Probing**: understanding the strengths/weaknesses of models; **measuring model competence qualitatively**; **behavioral (input/output) testing**.

Building NLU Models: Standard Picture

$\boxed{\text{(Neural) Model Architecture}}$

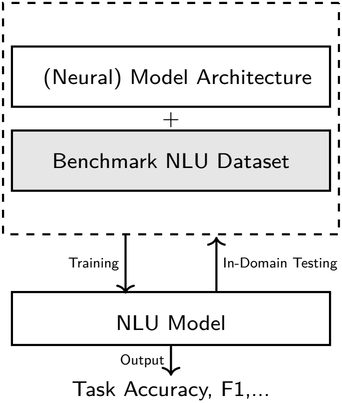
Building NLU Models: Standard Picture

(Neural) Model Architecture

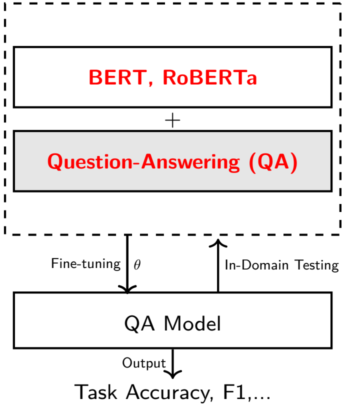
+

Benchmark NLU Dataset

Building NLU Models: Standard Picture

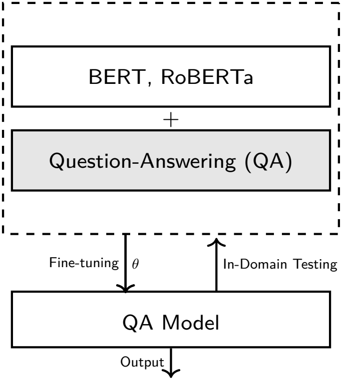


Building NLU Models: Standard Picture



Multiple-Choice QA **(ARC Benchmark)**								
Question	*Which property of a mineral can be determined just by looking at it?*							
Answers	(A)	$\underbrace{\textit{luster}}_{\text{correct answer}}$	(B)	$\underbrace{\textit{mass}}_{\text{distractor 1}}$	(C)	$\underbrace{\textit{weight}}_{\text{distractor 2}}$	(D)	$\underbrace{\textit{hardness}}_{\text{distractor 3}}$

Qualitative Analysis of Models



Desiderata
Does my model know about
Taxonomic relations, definitions, synonymy,
robust to perturbations/consistent, ...?

What Does My QA Model *Actually* Know?

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though analyses tend to be anecdotal and post-hoc.

What Does My QA Model *Actually* Know?

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though analyses tend to be anecdotal and post-hoc.

ARC Challenge (Clark et al., 2018)	
Question	*What is a worldwide increase in temperature $\underbrace{\textit{called}}_{\textbf{\textit{Definition}}}$?*
Answers	(A) greenhouse effect (B) global warming (C) ozone depletion (D) solar heating.
Knowledge: `DEF`(***global warming,worldwide increase in...***)	

What Does My QA Model *Actually* Know?

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though analyses tend to be anecdotal and post-hoc.

ARC Challenge (Clark et al., 2018)	
Question	*What is a worldwide increase in temperature $\underbrace{\text{called}}_{\textbf{Definition}}$?*
Answers	(A) greenhouse effect (B) global warming (C) ozone depletion (D) solar heating.
Knowledge: `DEF`(***global warming,worldwide increase in...***)	

OpenBookQA (Mihaylov et al., 2018)	
Question	*Which of the following $\underbrace{\text{is a type of}}_{\textbf{ISA reasoning}}$ learned behavior?*
Answers	(A) cooking (B) thinking (C) hearing (D) breathing
Knowledge: `ISA`(***cooking,learned behavior***)	

What Does My QA Model *Actually* Know?

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though analyses tend to be anecdotal and post-hoc.

ARC Challenge (Clark et al., 2018)	
Question	*What is a worldwide increase in temperature* $\underbrace{\textit{called}}_{\textbf{Definition}}$ *?*
Answers	(A) greenhouse effect (B) global warming (C) ozone depletion (D) solar heating.
Knowledge: DEF(***global warming,worldwide increase in...***)	

OpenBookQA (Mihaylov et al., 2018)	
Question	*Which of the following* $\underbrace{\textit{is a type of}}_{\textbf{ISA reasoning}}$ *learned behavior?*
Answers	(A) cooking (B) thinking (C) hearing (D) breathing
Knowledge: ISA(***cooking,learned behavior***)	

Do models truly possess the basic knowledge/reasoning skills we think they do? Hard to say without **specialized tests**.

What Does My QA Model *Actually* Know?

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though analyses tend to be anecdotal and post-hoc.

ARC Challenge (Clark et al., 2018)	
Question	*What is a worldwide increase in temperature* $\underbrace{\textit{called}}_{\textbf{\textit{Definition}}}$ *?*
Answers	(A) greenhouse effect (B) global warming (C) ozone depletion (D) solar heating.
Knowledge: `DEF`(***global warming,worldwide increase in...***)	

OpenBookQA (Mihaylov et al., 2018)	
Question	*Which of the following* $\underbrace{\textit{is a type of}}_{\textbf{\textit{ISA reasoning}}}$ *learned behavior?*
Answers	(A) cooking (B) thinking (C) hearing (D) breathing
Knowledge: `ISA`(***cooking,learned behavior***)	

To demonstrate competence a **model should:**

1. have knowledge across a *many concepts;*
2. be robust to *perturbations*
3. *and varying levels of reasoning complexity*.

What Does My QA Model *Actually* Know?

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though analyses tend to be anecdotal and post-hoc.

ARC Challenge (Clark et al., 2018)	
Question	*What is the thinning of Earth's upper atmosphere $\underbrace{\text{called}}_{\textbf{Definition}}$?*
Answers	(A) greenhouse effect (B) global warming (C) ozone depletion (D) solar heating.
Knowledge: DEF(***ozone depletion,thinning of the Earth's ...***)	

OpenBookQA (Mihaylov et al., 2018)	
Question	*Which of the following $\underbrace{\text{is a type of}}_{\textbf{ISA reasoning}}$ learned behavior?*
Answers	(A) cooking (B) thinking (C) hearing (D) breathing
Knowledge: ISA(***cooking,learned behavior***)	

To demonstrate competence a **model should:**

1. have knowledge across a *many concepts;*
2. be robust to *perturbations*
3. *and varying levels of reasoning complexity*.

What Does My QA Model *Actually* Know?

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though analyses tend to be anecdotal and post-hoc.

ARC Challenge (Clark et al., 2018)	
Question	*What is the thinning of Earth's upper atmosphere* $\underbrace{\textit{called}}_{\textbf{Definition}}$ *?*
Answers	(A) greenhouse effect (B) global warming (C) ozone depletion (D) solar heating.
Knowledge: DEF(***ozone depletion,thinning of the Earth's ...***)	

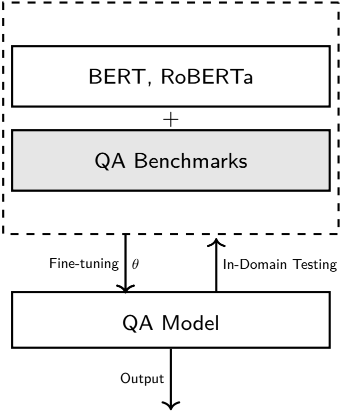
OpenBookQA (Mihaylov et al., 2018)	
Question	*Which of the following* $\underbrace{\textit{is a } \cancel{\textit{type}} \textit{ form of}}_{\textbf{ISA reasoning}}$ *~~learned~~ behavior?*
Answers	(A) cooking (B) thinking (C) hearing (D) ~~breathing~~ eating
Knowledge: ISA(***cooking,~~learned~~ behavior***)	

To demonstrate competence a **model should:**

1. have knowledge across a *many concepts;*
2. be robust to *perturbations*
3. *and varying levels of reasoning complexity*.

Diagnostic Tasks for NLU

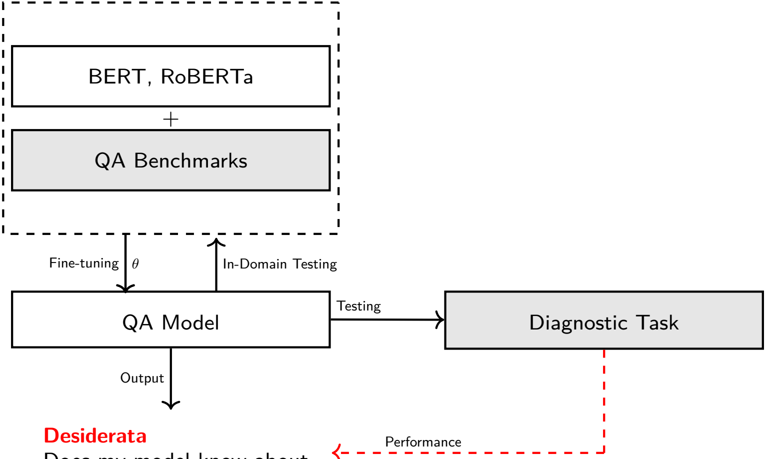
- Unit testing (Ribeiro et al., 2020), LMs as KBs (Petroni et al., 2019), challenge tasks (Glockner et al., 2018; Richardson et al., 2020); *inter alia*.



Desiderata
Does my model know about
Taxonomic relations, definitions, ...

Diagnostic Tasks for NLU

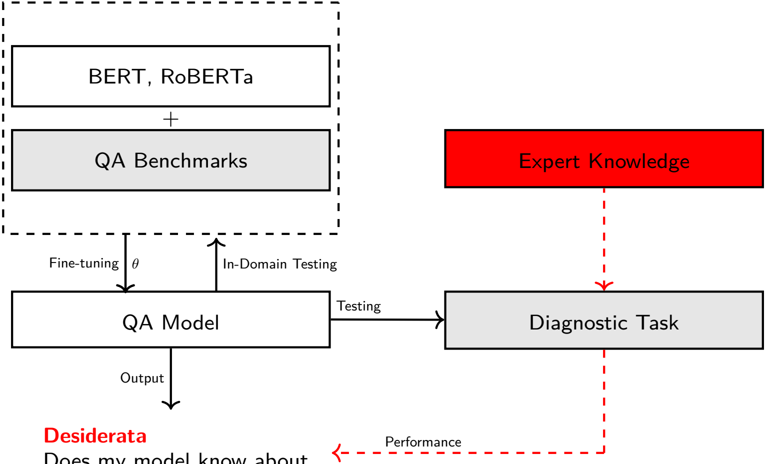
- Unit testing (Ribeiro et al., 2020), LMs as KBs (Petroni et al., 2019), challenge tasks (Glockner et al., 2018; Richardson et al., 2020); *inter alia*.



Desiderata
Does my model know about
Taxonomic relations, definitions, ...

Diagnostic Tasks for NLU

- Unit testing (Ribeiro et al., 2020), LMs as KBs (Petroni et al., 2019), challenge tasks (Glockner et al., 2018; Richardson et al., 2020); *inter alia*.



Desiderata
Does my model know about
Taxonomic relations, definitions, ...

<Building Diagnostic Tasks>

Building Diagnostic Tasks using Expert Knowledge

- A model should 1. have knowledge across many concepts ; 2. robust to perturbations ; 3. varying complexity .

Building Diagnostic Tasks using Expert Knowledge

- A ~~model~~ **dataset** should 1. ~~have~~ **test** knowledge across many concepts ; 2. ~~robust to~~ **have** perturbations ; 3. ~~varying~~ **controlled** complexity .

  Assumption: we can demonstrate that models exhibit these properties by testing them on data that has these properties..

Building Diagnostic Tasks using Expert Knowledge

- A ~~model~~ **dataset** should 1. ~~have~~ **test** knowledge across many concepts ; 2. ~~robust to~~ **have** perturbations ; 3. ~~varying~~ **controlled** complexity .

Assumption: we can demonstrate that models exhibit these properties by testing them on data that has these properties..

Expert Knowledge (KBs, lexical ontologies)

Arg1	Arg2	REL	EX
`nestle.v`	*position comfortably*	`DEF`	*The baby nestled her head..*
`elude.v`	`escape.v`	`ISA`	*The thief eluded police...*
`trouser.n`	`consumer good.n`	`ISA`	*The man bought trousers..*
`poet.n`	`writer.n`	`ISA`	...
...	...	...	...

Building Diagnostic Tasks using Expert Knowledge

- A ~~model~~ **dataset** should 1. ~~have~~ **test** knowledge across many concepts ; 2. ~~robust to~~ **have** perturbations ; 3. ~~varying~~ **controlled** complexity .

Assumption: we can demonstrate that models exhibit these properties by testing them on data that has these properties..

Expert Knowledge (KBs, lexical ontologies)

Arg1	Arg2	REL	EX
`nestle.v`	*position comfortably*	`DEF`	*The baby nestled her head..*
`elude.v`	`escape.v`	`ISA`	*The thief eluded police...*
`trouser.n`	`consumer good.n`	`ISA`	*The man bought trousers..*
`poet.n`	`writer.n`	`ISA`	...
...	...	...	...

templates →

Probing Questions

Question	Answer	Test
Given *'The baby nestled her head'*, **nestled** is defined as	*position comfortably*	**def**
In *'we had to spell our name for the police'*, **spell** is a type of	recite event	**isa**
In the context, *'the poet published his new poem'*, **poet** is best defined as	a writer of poems....	**def**

Building Diagnostic Tasks using Expert Knowledge

- A ~~model~~ **dataset** should 1. ~~have~~ **test** knowledge across many concepts ; 2. ~~robust to~~ **have** perturbations ; 3. ~~varying~~ **controlled** complexity .

Assumption: we can demonstrate that models exhibit these properties by testing them on data that has these properties..

Expert Knowledge (KBs, lexical ontologies)

Arg1	Arg2	REL	EX
nestle.v	*position comfortably*	DEF	*The baby nestled her head..*
elude.v	escape.v	ISA	*The thief eluded police...*
trouser.n	consumer good.n	ISA	*The man bought trousers..*
poet.n	writer.n	ISA	...
...	...	...	...

templates

Probing Questions

Question	Answer	Test
Given *'The baby nestled her head'*, **nestled** is defined as	*position comfortably*	**def**
In *'we had to spell our name for the police'*, **spell** is a type of	recite event	**isa**
In the context, *'the poet published his new poem'*, **poet** is best defined as	a writer of poems....	**def**

distractor assignment/taxonomic constraints

Diagnostic Task

Building Diagnostic Tasks using Expert Knowledge

- A ~~model~~ **dataset** should 1. ~~have~~ **test** knowledge across many concepts ; 2. ~~robust to~~ **have** perturbations ; 3. ~~varying~~ **controlled** complexity .

Assumption: we can demonstrate that models exhibit these properties by testing them on data that has these properties..

Expert Knowledge (KBs, lexical ontologies)

Arg1	**Arg2**	**REL**	**EX**
`nestle.v`	*position comfortably*	DEF	*The baby nestled her head..*
`elude.v`	`escape.v`	ISA	*The thief eluded police...*
`trouser.n`	`consumer good.n`	ISA	*The man bought trousers..*
`poet.n`	`writer.n`	ISA	...
...	...	...	...

templates

Probing Questions

Question	**Answer**	**Test**
Given *'The baby nestled her head'*, **nestled** is defined as	*position comfortably*	**def**
In *'we had to spell our name for the police'*, **spell** is a type of	recite event	**isa**
In the context, *'the poet published his new poem'*, **poet** is best defined as	a writer of poems....	**def**

distractor assignment/taxonomic constraints

Diagnostic Task

Meta-level QA: Asking questions about abstract knowledge; many concepts (**1.** ✓); controlled templates/distractor complexity (**2.**✓ **3.** ✓)

Building Diagnostic Tasks using Expert Knowledge

- A ~~model~~ **dataset** should 1. ~~have~~ **test** knowledge across many concepts ; 2. ~~robust to~~ **have** perturbations ; 3. ~~varying~~ **controlled** complexity .

Assumption: we can demonstrate that models exhibit these properties by testing them on data that has these properties..

Expert Knowledge (KBs, lexical ontologies)

Arg1	Arg2	REL	EX
`nestle.v`	*position comfortably*	DEF	*The baby nestled her head..*
`elude.v`	`escape.v`	ISA	*The thief eluded police...*
`trouser.n`	`consumer good.n`	ISA	*The man bought trousers..*
`poet.n`	`writer.n`	ISA	...
...	...	...	...

templates

Probing Questions

Question	Answer	Test
Given *'The baby nestled her head'*, **nestled** is defined as	*position comfortably*	**def**
In *'we had to spell our name for the police'*, **spell** is a type of	recite event	**isa**
In the context, *'the poet published his new poem'*, **poet** is best defined as	a writer of poems....	**def**

distractor assignment/taxonomic constraints

Diagnostic Task

Trade-offs: KBs tends to be noisy; dealt by synthesizing large amount of data, contextualizing questions, gold test annotation (where needed).

Example Diagnostics

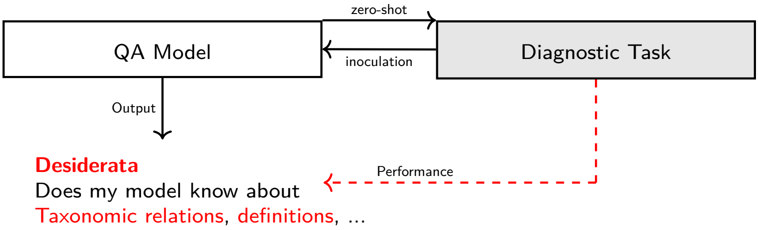
- **Resources**: WordNet, GCIDE dictionary; **5 individual tasks**: Definitions, Synonymy, Hypernymy (ISA), and Hyponymy (ISA), WordSense.
  - WordNet tasks involve ~ $30k$ atomic concepts, exhaustive combinations of distractors.

Probe	**Example**
Definitions + Word Sense	*In the sentence* *The baby nestled her head* *, the word* *nestled* *is best defined as (A)* $\underbrace{\textit{position comfortably}}_{\text{correct answer}}$ *(B)* $\underbrace{\textit{put in a certain place}}_{\text{hard/close distractor}}$ *(C)* $\underbrace{\textit{a type of fish}}_{\text{easy/random distractor}}$ ...
Hypernymy (ISA)	In The thief eluded the police , the word of concept eluded is best described as (A) ... (B) $\underbrace{\text{an escape event, defined as ...}}_{\text{correct answer}}$ (C) ...
Hyponymy (ISA)	Given the context They awaited her arrival , which of the following is a specific type of arrival (A) $\underbrace{\text{driving a car}}_{\text{related concept}}$ (B) $\underbrace{\text{crash landing, defined as ......}}_{\text{correct answer}}$
Synonymy	Which set of words best corresponds to the definition of a grammatical category in inflected languages... (A) $\underbrace{\text{gender}}_{\text{correct answer}}$ (B) ...

Probing Methodology and Experiments

- Trained single models (BERT, RoBERTa) on aggregated science QA dataset (4 benchmarks); **Ask the following empirical questions**:

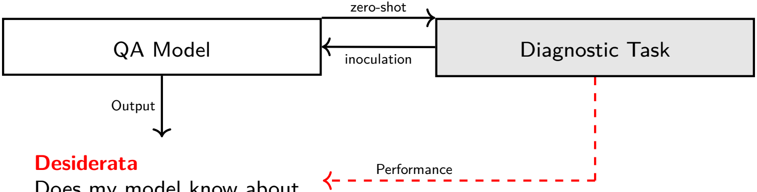
  1. How well do benchmark models perform on each *individual* probing on diagnostic task without specialized training (**zero-shot**)?
  2. How well models perform after a small amount of additional training on probes (**inoculation** (Liu et al., 2019))?



Probing Methodology and Experiments

- Trained single models (BERT, RoBERTa) on aggregated science QA dataset (4 benchmarks); **Ask the following empirical questions**:

  1. How well do benchmark models perform on each *individual* probing on diagnostic task without specialized training (**zero-shot**)?
  2. How well models perform after a small amount of additional training on probes (**inoculation** (Liu et al., 2019))?



Desiderata
Does my model know about
Taxonomic relations, definitions, ...

Controls: Probes should be demonstrably difficult (**strong baselines**); Re-training must preserve performance (minimal **inoculation loss**).

<General Findings>

1. Zero-shot performance (*Challenge Task* setting)

- Without specialized training, models do well on *some* categories of knowledge; sometimes far outpace baselines trained on diagnostics.

1. Zero-shot performance (*Challenge Task* setting)

- Without specialized training, models do well on *some* categories of knowledge; sometimes far outpace baselines trained on diagnostics.

Diagnostic performance (QA Accuracy %; random ~ 30%)

Model	Definitions	Synonymy	Hypernymy	Hyponymy	WordSense
trained LSTM + GloVe	51.8%	55.3%	47.0%	64.2%	53.5%
BERT (zero-shot)	55.7%	60.9%	51.0%	27.0%	42.9%
RoBERTa (zero-shot)	77.1 %	64.2%	71.0%	58.0%	55.1%
Human	91.2%	87.4%	96%	95.5%	95.6%

1. Zero-shot performance (*Challenge Task* setting)

- Without specialized training, models do well on *some* categories of knowledge; sometimes far outpace baselines trained on diagnostics.

Diagnostic performance (QA Accuracy %; random ~ 30%)

Model	Definitions	Synonymy	Hypernymy	Hyponymy	WordSense
trained LSTM + GloVe	51.8%	55.3%	47.0%	64.2%	53.5%
BERT (zero-shot)	55.7%	60.9%	51.0%	27.0%	42.9%
RoBERTa (zero-shot)	77.1 %	64.2%	71.0%	58.0%	55.1%
Human	91.2%	87.4%	96%	95.5%	95.6%

Caveats: Reflect true model knowledge or (non-)familiarity with format? Lower-bound estimate (Petroni et al., 2019).

2. Continue training (*inoculation setting*)

- Bring out knowledge by continue training with a small *dosage* (Liu et al., 2019) of diagnostic data, **inoculate** against dataset.

Diagnostic performance (QA Accuracy %; random ~ 30%)

Model	Definitions	Synonymy	Hypernymy	Hyponymy	WordSense
BERT (inoculation)	84.1%	79.7%	82.7%	88.0%	79.1%
RoBERTa (inoculation)	89.3 %	81.3%	87.0%	89.4%	85.4%
Human	91.2%	87.4%	96%	95.5%	95.6%

2. Continue training (*inoculation setting*)

- Bring out knowledge by continue training with a small *dosage* (Liu et al., 2019) of diagnostic data, **inoculate** against dataset.

Diagnostic performance (QA Accuracy %; random ~ 30%)

Model	Definitions	Synonymy	Hypernymy	Hyponymy	WordSense
BERT (inoculation)	84.1%	79.7%	82.7%	88.0%	79.1%
RoBERTa (inoculation)	89.3 %	81.3%	87.0%	89.4%	85.4%
Human	91.2%	87.4%	96%	95.5%	95.6%

Giving the model the chance to learn **target format** is important, gives better picture of competence; minimal loss on original task.

2. Continue training (*inoculation setting*): nuances

- The controlled nature of the probes allows for a more granular examination of performance.

2. Continue training (*inoculation setting*): nuances

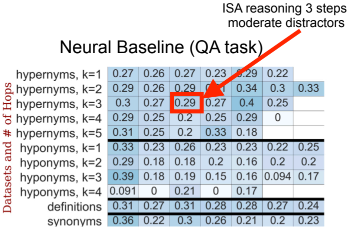
- The controlled nature of the probes allows for a more granular examination of performance.

Neural Baseline (QA task)

Datasets and # of Hops							
hypernyms, k=1	0.27	0.26	0.27	0.23	0.29	0.22	
hypernyms, k=2	0.29	0.26	0.29	0.31	0.34	0.3	0.33
hypernyms, k=3	0.3	0.27	0.29	0.27	0.4	0.25	
hypernyms, k=4	0.29	0.25	0.2	0.25	0.29	0	
hypernyms, k=5	0.31	0.25	0.2	0.33	0.18		
hyponyms, k=1	0.33	0.23	0.26	0.23	0.23	0.22	0.25
hyponyms, k=2	0.29	0.18	0.18	0.2	0.16	0.2	0.2
hyponyms, k=3	0.39	0.18	0.19	0.15	0.16	0.094	0.17
hyponyms, k=4	0.091	0	0.21	0	0.17		
definitions	0.31	0.27	0.31	0.28	0.28	0.27	0.24
synonyms	0.36	0.22	0.3	0.26	0.21	0.2	0.23

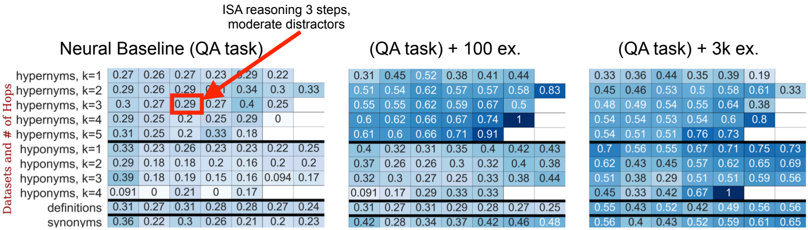
2. Continue training (*inoculation setting*): nuances

- The controlled nature of the probes allows for a more granular examination of performance.



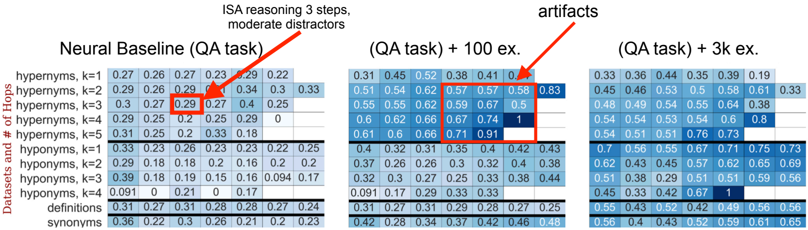
2. Continue training (*inoculation setting*): nuances

- The controlled nature of the probes allows for a more granular examination of performance.



2. Continue training (*inoculation setting*): nuances

- The controlled nature of the probes allows for a more granular examination of performance.



2. Continue training (*inoculation setting*): nuances

- The controlled nature of the probes allows for a more granular examination of performance.

ISA reasoning 3 steps, moderate distractors

Neural Baseline (QA task)

Datasets and # of Hops							
hypernyms, k=1	0.27	0.26	0.27	0.23	0.29	0.22	
hypernyms, k=2	0.29	0.26	0.29	[illegible]	0.34	0.3	0.33
hypernyms, k=3	0.3	0.27	0.29	0.27	0.4	0.25	
hypernyms, k=4	0.29	0.25	0.2	0.25	0.29	0	
hypernyms, k=5	0.31	0.25	0.2	0.33	0.18		
hyponyms, k=1	0.33	0.23	0.26	0.23	0.23	0.22	0.25
hyponyms, k=2	0.29	0.18	0.18	0.2	0.16	0.2	0.2
hyponyms, k=3	0.39	0.18	0.19	0.15	0.16	0.094	0.17
hyponyms, k=4	0.091	0	0.21	0	0.17		
definitions	0.31	0.27	0.31	0.28	0.28	0.27	0.24
synonyms	0.36	0.22	0.3	0.26	0.21	0.2	0.23

(QA task) + 100 ex.

Datasets and # of Hops							
hypernyms, k=1	0.31	0.45	0.52	0.38	0.41	0.44	
hypernyms, k=2	0.51	0.54	0.62	0.57	0.57	0.58	0.83
hypernyms, k=3	0.55	0.55	0.62	0.59	0.67	0.5	
hypernyms, k=4	0.6	0.62	0.66	0.67	0.74	1	
hypernyms, k=5	0.61	0.6	0.66	0.71	0.91		
hyponyms, k=1	0.4	0.32	0.31	0.35	0.4	0.42	0.43
hyponyms, k=2	0.37	0.26	0.26	0.3	0.32	0.4	0.38
hyponyms, k=3	0.32	0.3	0.27	0.25	0.33	0.38	0.44
hyponyms, k=4	0.091	0.17	0.29	0.33	0.33		
definitions	0.31	0.27	0.31	0.29	0.28	0.27	0.25
synonyms	0.42	0.28	0.34	0.37	0.42	0.46	0.48

(QA task) + 3k ex.

Datasets and # of Hops							
hypernyms, k=1	0.33	0.36	0.44	0.35	0.39	0.19	
hypernyms, k=2	0.45	0.46	0.53	0.5	0.58	0.61	0.33
hypernyms, k=3	0.48	0.49	0.54	0.55	0.64	0.38	
hypernyms, k=4	0.54	0.54	0.53	0.54	0.6	0.8	
hypernyms, k=5	0.54	0.51	0.51	0.76	0.73		
hyponyms, k=1	0.7	0.56	0.55	0.67	0.71	0.75	0.73
hyponyms, k=2	0.62	0.43	0.45	0.57	0.62	0.65	0.69
hyponyms, k=3	0.51	0.38	0.29	0.51	0.51	0.59	0.56
hyponyms, k=4	0.45	0.33	0.42	0.67	1		
definitions	0.55	0.43	0.52	0.42	0.49	0.56	0.56
synonyms	0.56	0.4	0.43	0.52	0.59	0.61	0.65

SOTA Transformer (QA task)

Datasets and # of Hops							
hypernyms, k=1	0.76	0.57	0.68	0.48	0.64	0.85	
hypernyms, k=2	0.71	0.47	0.58	0.43	0.57	0.7	0.67
hypernyms, k=3	0.65	0.38	0.5	0.4	0.54	0.69	
hypernyms, k=4	0.61	0.33	0.4	0.33	0.43	0.4	
hypernyms, k=5	0.62	0.33	0.46	0.26	0.27		
hyponyms, k=1	0.72	0.47	0.58	0.34	0.46	0.55	0.57
hyponyms, k=2	0.59	0.37	0.45	0.25	0.35	0.45	0.46
hyponyms, k=3	0.43	0.24	0.3	0.1	0.098	0.19	0.33
hyponyms, k=4	0.64	0.15	0.38	0	0.33		
definitions	0.88	0.72	0.8	0.64	0.73	0.76	0.77
synonyms	0.82	0.49	0.67	0.63	0.63	0.61	0.66

2. Continue training (*inoculation setting*): nuances

- The controlled nature of the probes allows for a more granular examination of performance.

ISA reasoning 3 steps, moderate distractors

Datasets and # of Hops

Neural Baseline (QA task)

hypernyms, k=1	0.27	0.26	0.27	0.23	0.29	0.22	
hypernyms, k=2	0.29	0.26	0.29	[illegible]	0.34	0.3	0.33
hypernyms, k=3	0.3	0.27	$\boxed{0.29}$	0.27	0.4	0.25	
hypernyms, k=4	0.29	0.25	0.2	0.25	0.29	0	
hypernyms, k=5	0.31	0.25	0.2	0.33	0.18		
hyponyms, k=1	0.33	0.23	0.26	0.23	0.23	0.22	0.25
hyponyms, k=2	0.29	0.18	0.18	0.2	0.16	0.2	0.2
hyponyms, k=3	0.39	0.18	0.19	0.15	0.16	0.094	0.17
hyponyms, k=4	0.091	0	0.21	0	0.17		
definitions	0.31	0.27	0.31	0.28	0.28	0.27	0.24
synonyms	0.36	0.22	0.3	0.26	0.21	0.2	0.23

(QA task) + 100 ex.

hypernyms, k=1	0.31	0.45	0.52	0.38	0.41	0.44	
hypernyms, k=2	0.51	0.54	0.62	0.57	0.57	0.58	0.83
hypernyms, k=3	0.55	0.55	0.62	0.59	0.67	0.5	
hypernyms, k=4	0.6	0.62	0.66	0.67	0.74	1	
hypernyms, k=5	0.61	0.6	0.66	0.71	0.91		
hyponyms, k=1	0.4	0.32	0.31	0.35	0.4	0.42	0.43
hyponyms, k=2	0.37	0.26	0.26	0.3	0.32	0.4	0.38
hyponyms, k=3	0.32	0.3	0.27	0.25	0.33	0.38	0.44
hyponyms, k=4	0.091	0.17	0.29	0.33	0.33		
definitions	0.31	0.27	0.31	0.29	0.28	0.27	0.25
synonyms	0.42	0.28	0.34	0.37	0.42	0.46	0.48

(QA task) + 3k ex.

hypernyms, k=1	0.33	0.36	0.44	0.35	0.39	0.19	
hypernyms, k=2	0.45	0.46	0.53	0.5	0.58	0.61	0.33
hypernyms, k=3	0.48	0.49	0.54	0.55	0.64	0.38	
hypernyms, k=4	0.54	0.54	0.53	0.54	0.6	0.8	
hypernyms, k=5	0.54	0.51	0.51	0.76	0.73		
hyponyms, k=1	0.7	0.56	0.55	0.67	0.71	0.75	0.73
hyponyms, k=2	0.62	0.43	0.45	0.57	0.62	0.65	0.69
hyponyms, k=3	0.51	0.38	0.29	0.51	0.51	0.59	0.56
hyponyms, k=4	0.45	0.33	0.42	0.67	1		
definitions	0.55	0.43	0.52	0.42	0.49	0.56	0.56
synonyms	0.56	0.4	0.43	0.52	0.59	0.61	0.65

SOTA Transformer (QA task)

hypernyms, k=1	0.76	0.57	0.68	0.48	0.64	0.85	
hypernyms, k=2	0.71	0.47	0.58	0.43	0.57	0.7	0.67
hypernyms, k=3	0.65	0.38	0.5	0.4	0.54	0.69	
hypernyms, k=4	0.61	0.33	0.4	0.33	0.43	0.4	
hypernyms, k=5	0.62	0.33	0.46	0.26	0.27		
hyponyms, k=1	0.72	0.47	0.58	0.34	0.46	0.55	0.57
hyponyms, k=2	0.59	0.37	0.45	0.25	0.35	0.45	0.46
hyponyms, k=3	0.43	0.24	0.3	0.1	0.098	0.19	0.33
hyponyms, k=4	0.64	0.15	0.38	0	0.33		
definitions	0.88	0.72	0.8	0.64	0.73	0.76	0.77
synonyms	0.82	0.49	0.67	0.63	0.63	0.61	0.66

(QA task) + 100 ex.

hypernyms, k=1	0.83	0.71	0.82	0.61	0.81	0.85	
hypernyms, k=2	0.8	0.63	0.76	0.61	0.71	0.85	1
hypernyms, k=3	0.77	0.58	0.71	0.59	0.72	0.75	
hypernyms, k=4	0.79	0.58	0.67	0.56	0.74	0.8	
hypernyms, k=5	0.79	0.59	0.76	0.41	0.36		
hyponyms, k=1	0.89	0.68	0.74	0.64	0.81	0.85	0.88
hyponyms, k=2	0.77	0.5	0.59	0.43	0.65	0.7	0.72
hyponyms, k=3	0.65	0.4	0.42	0.34	0.51	0.56	0.72
hyponyms, k=4	0.55	0.12	0.25	0.33	0.67		
definitions	0.94	0.83	0.88	0.7	0.84	0.9	0.92
synonyms	0.85	0.5	0.67	0.73	0.82	0.83	0.83

2. Continue training (*inoculation setting*): nuances

- The controlled nature of the probes allows for a more granular examination of performance.

ISA reasoning 3 steps, moderate distractors

Datasets and # of Hops

Neural Baseline (QA task)

hypernyms, k=1	0.27	0.26	0.27	0.23	0.29	0.22	
hypernyms, k=2	0.29	0.26	0.29	[illegible]	0.34	0.3	0.33
hypernyms, k=3	0.3	0.27	0.29	0.27	0.4	0.25	
hypernyms, k=4	0.29	0.25	0.2	0.25	0.29	0	
hypernyms, k=5	0.31	0.25	0.2	0.33	0.18		
hyponyms, k=1	0.33	0.23	0.26	0.23	0.23	0.22	0.25
hyponyms, k=2	0.29	0.18	0.18	0.2	0.16	0.2	0.2
hyponyms, k=3	0.39	0.18	0.19	0.15	0.16	0.094	0.17
hyponyms, k=4	0.091	0	0.21	0	0.17		
definitions	0.31	0.27	0.31	0.28	0.28	0.27	0.24
synonyms	0.36	0.22	0.3	0.26	0.21	0.2	0.23

(QA task) + 100 ex.

hypernyms, k=1	0.31	0.45	0.52	0.38	0.41	0.44	
hypernyms, k=2	0.51	0.54	0.62	0.57	0.57	0.58	0.83
hypernyms, k=3	0.55	0.55	0.62	0.59	0.67	0.5	
hypernyms, k=4	0.6	0.62	0.66	0.67	0.74	1	
hypernyms, k=5	0.61	0.6	0.66	0.71	0.91		
hyponyms, k=1	0.4	0.32	0.31	0.35	0.4	0.42	0.43
hyponyms, k=2	0.37	0.26	0.26	0.3	0.32	0.4	0.38
hyponyms, k=3	0.32	0.3	0.27	0.25	0.33	0.38	0.44
hyponyms, k=4	0.091	0.17	0.29	0.33	0.33		
definitions	0.31	0.27	0.31	0.29	0.28	0.27	0.25
synonyms	0.42	0.28	0.34	0.37	0.42	0.46	0.48

(QA task) + 3k ex.

hypernyms, k=1	0.33	0.36	0.44	0.35	0.39	0.19	
hypernyms, k=2	0.45	0.46	0.53	0.5	0.58	0.61	0.33
hypernyms, k=3	0.48	0.49	0.54	0.55	0.64	0.38	
hypernyms, k=4	0.54	0.54	0.53	0.54	0.6	0.8	
hypernyms, k=5	0.54	0.51	0.51	0.76	0.73		
hyponyms, k=1	0.7	0.56	0.55	0.67	0.71	0.75	0.73
hyponyms, k=2	0.62	0.43	0.45	0.57	0.62	0.65	0.69
hyponyms, k=3	0.51	0.38	0.29	0.51	0.51	0.59	0.56
hyponyms, k=4	0.45	0.33	0.42	0.67	1		
definitions	0.55	0.43	0.52	0.42	0.49	0.56	0.56
synonyms	0.56	0.4	0.43	0.52	0.59	0.61	0.65

SOTA Transformer (QA task)

hypernyms, k=1	0.76	0.57	0.68	0.48	0.64	0.85	
hypernyms, k=2	0.71	0.47	0.58	0.43	0.57	0.7	0.67
hypernyms, k=3	0.65	0.38	0.5	0.4	0.54	0.69	
hypernyms, k=4	0.61	0.33	0.4	0.33	0.43	0.4	
hypernyms, k=5	0.62	0.33	0.46	0.26	0.27		
hyponyms, k=1	0.72	0.47	0.58	0.34	0.46	0.55	0.57
hyponyms, k=2	0.59	0.37	0.45	0.25	0.35	0.45	0.46
hyponyms, k=3	0.43	0.24	0.3	0.1	0.098	0.19	0.33
hyponyms, k=4	0.64	0.15	0.38	0	0.33		
definitions	0.88	0.72	0.8	0.64	0.73	0.76	0.77
synonyms	0.82	0.49	0.67	0.63	0.63	0.61	0.66

(QA task) + 100 ex.

hypernyms, k=1	0.83	0.71	0.82	0.61	0.81	0.85	
hypernyms, k=2	0.8	0.63	0.76	0.61	0.71	0.85	1
hypernyms, k=3	0.77	0.58	0.71	0.59	0.72	0.75	
hypernyms, k=4	0.79	0.58	0.67	0.56	0.74	0.8	
hypernyms, k=5	0.79	0.59	0.76	0.41	0.36		
hyponyms, k=1	0.89	0.68	0.74	0.64	0.81	0.85	0.88
hyponyms, k=2	0.77	0.5	0.59	0.43	0.65	0.7	0.72
hyponyms, k=3	0.65	0.4	0.42	0.34	0.51	0.56	0.72
hyponyms, k=4	0.55	0.12	0.25	0.33	0.67		
definitions	0.94	0.83	0.88	0.7	0.84	0.9	0.92
synonyms	0.85	0.5	0.67	0.73	0.82	0.83	0.83

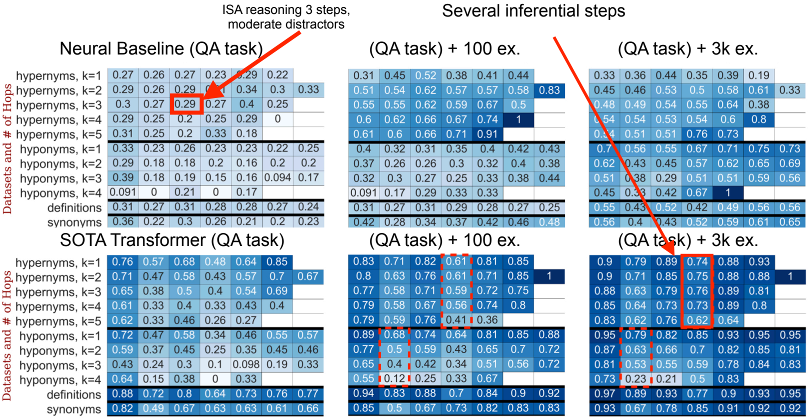
(QA task) + 3k ex.

hypernyms, k=1	0.9	0.79	0.89	0.74	0.88	0.93	
hypernyms, k=2	0.9	0.71	0.85	0.75	0.88	0.88	1
hypernyms, k=3	0.88	0.63	0.79	0.76	0.89	0.81	
hypernyms, k=4	0.85	0.64	0.73	0.73	0.89	0.8	
hypernyms, k=5	0.83	0.62	0.76	0.62	0.64		
hyponyms, k=1	0.95	0.79	0.82	0.85	0.93	0.95	0.95
hyponyms, k=2	0.87	0.63	0.66	0.7	0.82	0.85	0.81
hyponyms, k=3	0.81	0.53	0.55	0.59	0.78	0.84	0.83
hyponyms, k=4	0.73	0.23	0.21	0.5	0.83		
definitions	0.97	0.89	0.93	0.77	0.9	0.93	0.95
synonyms	0.93	0.67	0.78	0.85	0.91	0.92	0.92

Can nudge the models to bring out knowledge with small set of examples, cheap way to inject knowledge into transformers.

2. Continue training (*inoculation setting*): nuances

- The controlled nature of the probes allows for a more granular examination of performance.



Model does show sensitivity to reasoning complexity; is not always consistent across predictions. Hard to determine if model has knowledge.

</General Findings>

Conclusions

- Probing with expert knowledge: systematically constructed diagnostic tasks; supplement current QA research.
- Proposed 5 diagnostic tasks to look at performance of SOTA QA models for science; used lexical KBs (Wordnet) and other dictionaries.

Conclusions

- Probing with expert knowledge: systematically constructed diagnostic tasks; supplement current QA research.
- Proposed 5 diagnostic tasks to look at performance of SOTA QA models for science; used lexical KBs (Wordnet) and other dictionaries.
  - Models do exhibit impressive amounts of lexical and other structured knowledge.

Conclusions

- Probing with expert knowledge: systematically constructed diagnostic tasks; supplement current QA research.
- Proposed 5 diagnostic tasks to look at performance of SOTA QA models for science; used lexical KBs (Wordnet) and other dictionaries.
  - Models do exhibit impressive amounts of lexical and other structured knowledge.
  - Probing is difficult! Hard to achieve definitive proof of model knowledge (noisy knowledge, dataset biases).

Thank you.

References I

Boratko, M., Padigela, H., Mikkilineni, D., Yuvraj, P., Das, R., McCallum, A., Chang, M., Fokoue-Nkoutche, A., Kapanipathi, P., Mattei, N., et al. (2018). A systematic classification of knowledge, reasoning, and context within the arc dataset. *arXiv preprint arXiv:1806.00358*.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Glockner, M., Shwartz, V., and Goldberg, Y. (2018). Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. *arXiv preprint arXiv:1805.02266*.

Liu, N. F., Schwartz, R., and Smith, N. A. (2019). Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. *arXiv preprint arXiv:1904.02668*.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. (2018). Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. (2019). Language Models as Knowledge Bases? *arXiv preprint arXiv:1909.01066*.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of nlp models with checklist. *Proceedings of ACL*.

Richardson, K., Hu, H., Moss, L. S., and Sabharwal, A. (2020). Probing Natural Language Inference Models through Semantic Fragments. In *AAAI*, pages 8713–8721.