Qualitative Probing of Deep Contextual Models: We need your help!

Kyle Richardson

Allen Institute for Artificial Intelligence (AI2)

December 2020

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

Building NLU Models: Standard Picture

(Neural) Model Architecture

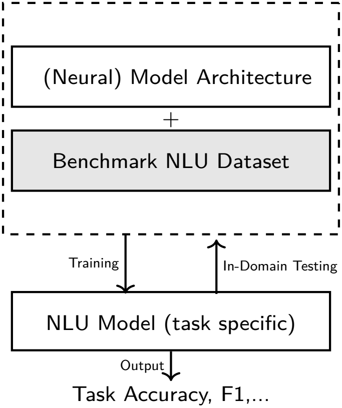
Building NLU Models: Standard Picture

(Neural) Model Architecture

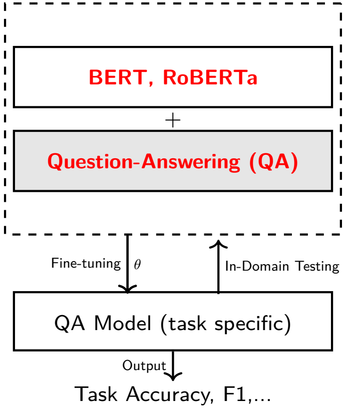
+

Benchmark NLU Dataset

Building NLU Models: Standard Picture

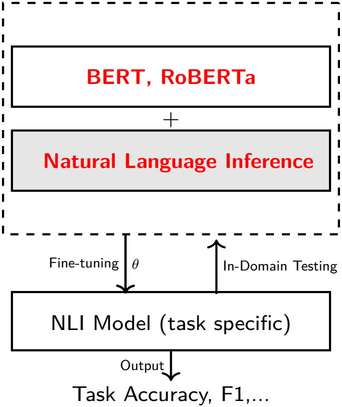


Building NLU Models: Standard Picture



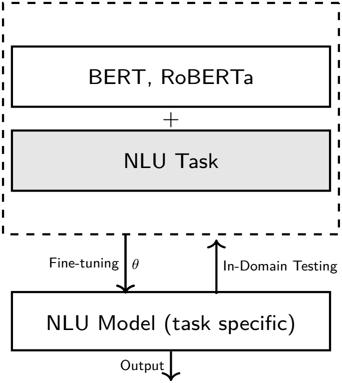
Multiple-Choice QA **(ARC Benchmark)**	
Question	*Which property of a mineral can be determined just by looking at it?*
Answers	(A) *luster* (correct answer) (B) *mass* (distractor 1) (C) *weight* (distractor 2) (D) *hardness* (distractor 3)

Building NLU Models: Standard Picture



Natural Language Inference (SNLI benchmark)	
Sen1	*A soccer game with multiple males playing.*
Sen2	*Some men are playing a sport.*
Label	`Yes/Entailment`

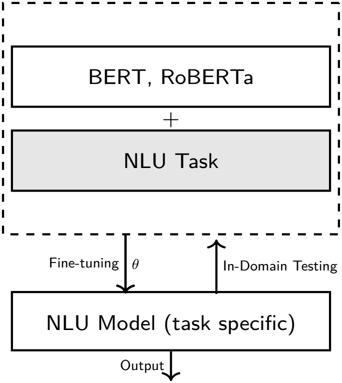
Qualitative Analysis of Models



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

- **Why**? Models sometimes do the right things for the wrong reasons; exploit biases (Gururangan et al., 2018); **model/bug repair**.

Qualitative Analysis of Models

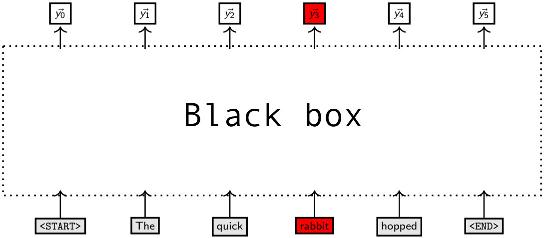


Desiderata
Does my model know about
Taxonomic relations, definitions, synonymy,
robust to perturbations/consistent, ...?

- **Bigger Issue** (not often discussed) Unclear how linguists, logicians, people working on classical AI fit into this picture; facilitate collab.

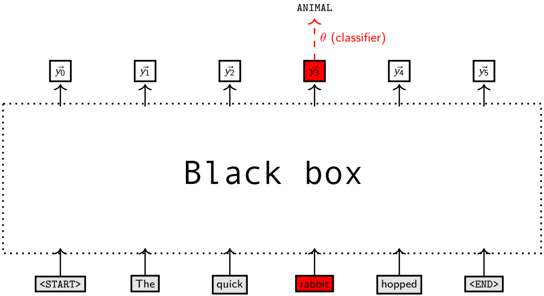
Contextual Models: A Cursory Overview

- **Role**: assign continuous (non-symbolic) vector representations $y \in \mathbb{R}$ to inputs based on their meaning in each instance; deep neural networks.



Contextual Models: A Cursory Overview

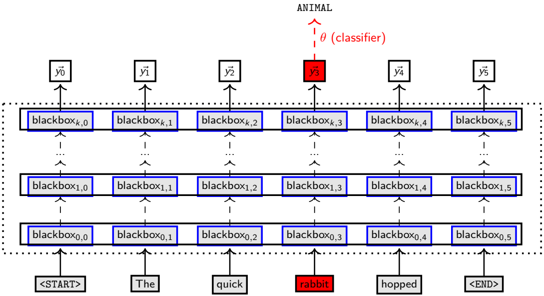
- **Role**: assign continuous (non-symbolic) vector representations $y \in \mathbb{R}$ to inputs based on their meaning in each instance; deep neural networks.



- **Words as Vectors**: Give models considerable power; word/concept similarity reduces to vector similarity, e.g., $\textsc{similarity}(\overrightarrow{\text{rabbit}}, \overrightarrow{\textit{bunny}})$.

Contextual Models: A Cursory Overview

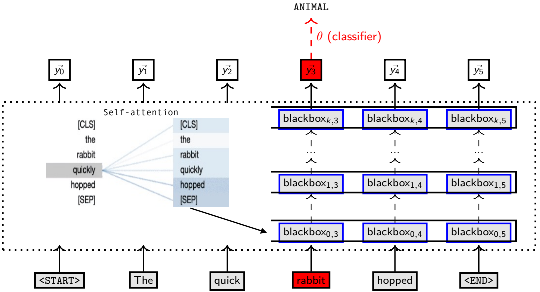
- **Role**: assign continuous (non-symbolic) vector representations $y \in \mathbb{R}$ to inputs based on their meaning in each instance; deep neural networks.



- **Development 1**: New architecture, **Transformers** (Vaswani et al., 2017), dispense with recurrent sequential structures, self-attention.

Contextual Models: A Cursory Overview

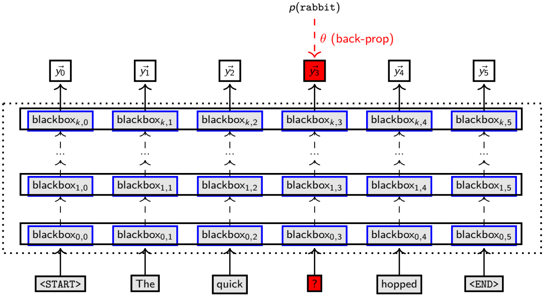
- **Role**: assign continuous (non-symbolic) vector representations $y \in \mathbb{R}$ to inputs based on their meaning in each instance; deep neural networks.



- **Development 1**: New architecture, **Transformers** (Vaswani et al., 2017), dispense with recurrent sequential structures, self-attention.

Contextual Models: A Cursory Overview

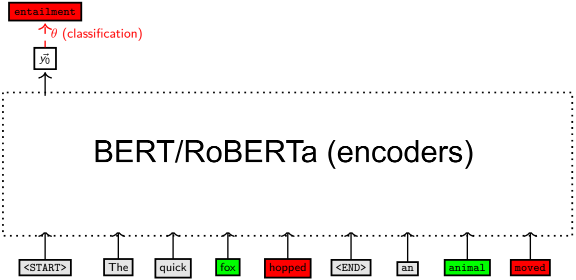
- **Role**: assign continuous (non-symbolic) vector representations $y \in \mathbb{R}$ to inputs based on their meaning in each instance; deep neural networks.



- **Development 2**: **Model pre-training**: Have the model read the internet (terabytes of data) and learn by solving word completion (*cloze*) tasks.

Contextual Models: A Cursory Overview

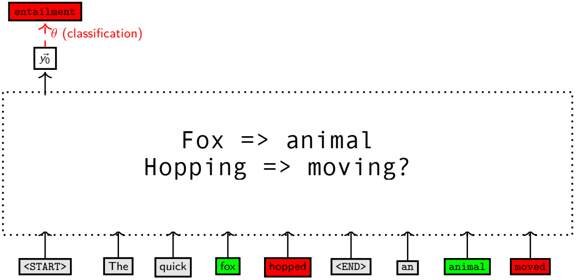
- **Role**: assign continuous (non-symbolic) vector representations $y \in \mathbb{R}$ to inputs based on their meaning in each instance; deep neural networks.



- **Fine-tuning**: Training models on smaller customized tasks, exploiting pre-trained knowledge. Pioneered in Devlin et al. (2018).

Contextual Models: A Cursory Overview

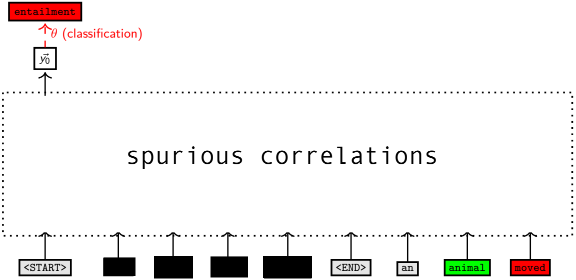
- **Role**: assign continuous (non-symbolic) vector representations $y \in \mathbb{R}$ to inputs based on their meaning in each instance; deep neural networks.



- Are these models actually knowledgeable, or just exploiting tricks and systematic biases in data? Gururangan et al. (2018)

Contextual Models: A Cursory Overview

- **Role**: assign continuous (non-symbolic) vector representations $y \in \mathbb{R}$ to inputs based on their meaning in each instance; deep neural networks.



- Are these models actually knowledgeable, or just exploiting tricks and systematic biases in data? Gururangan et al. (2018)

Requirements for Demonstrating knowledge

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

Requirements for Demonstrating knowledge

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though 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...)***	

Requirements for Demonstrating knowledge

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though 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***)	

Requirements for Demonstrating knowledge

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though 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***)	

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

Requirements for Demonstrating Knowledge

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though 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***)	

To demonstrate competence a **model should:**

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

Requirements for Demonstrating Knowledge

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though 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* $\underbrace{\textit{is a type of}}_{\textbf{ISA reasoning}}$ *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 *many concepts;*
2. be robust to *perturbations*
3. *and varying levels of reasoning complexity*.

Requirements for Demonstrating Knowledge

- QA in the **Science domain**, well studied qualitatively (Clark et al., 2018; Boratko et al., 2018), though 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 $\underbrace{\text{is a } \cancel{\text{type}} \text{ form of}}_{\textbf{ISA reasoning}}$ ~~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 *many concepts;*
2. be robust to *perturbations*
3. *and varying levels of reasoning complexity* .

Requirements for Demonstrate Knowledge

- When demonstrating knowledge, want to consider the extreme cases with considerable complexity; can result in pedantic English.

sen1	sen2	Label
Mitchell is as tall as Fred, Fred is as tall as Karl, Karl is as tall as Jon, Jon is as tall as Darryl, Darryl is as tall as Theodore, Theodore is as tall as Calvin, Calvin is as tall as Eddie , Eddie is as tall as Philip , Philip is taller than Travis	Calvin is taller than Travis .	Entailment
A bat with a strong odor did not hit several dogs	A bat with a strong smell did not hit many poodles	Entailment

Requirements for Demonstrate Knowledge

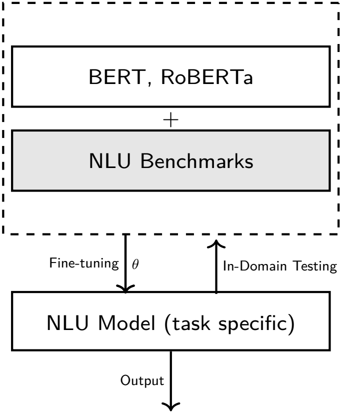
- When demonstrating knowledge, want to consider the extreme cases with considerable complexity; can result in pedantic English.

sen1	sen2	Label
Mitchell is as tall as Fred, Fred is as tall as Karl, Karl is as tall as Jon, Jon is as tall as Darryl, Darryl is as tall as Theodore, Theodore is as tall as Calvin, Calvin is as tall as Eddie , Eddie is as tall as Philip , Philip is taller than Travis	Calvin is taller than Travis .	Entailment
A bat with a strong odor did not hit several dogs	A bat with a strong smell did not hit many poodles	Entailment

- **Other robustness measures**: out of domain testing, lexical diversity (Rozen et al., 2019).

Diagnostic Tasks for NLU

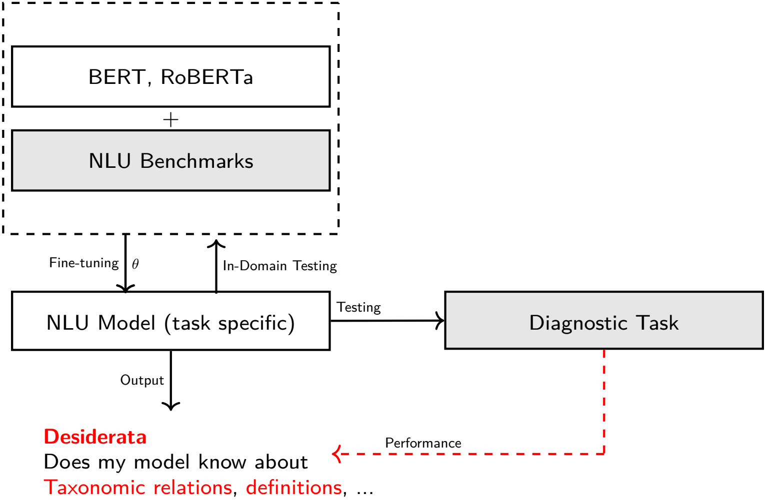
- unit testing (Ribeiro et al., 2020), challenge tasks/stress tests (**task specific**) (Naik et al., 2018; Glockner et al., 2018), *inter alia*.



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

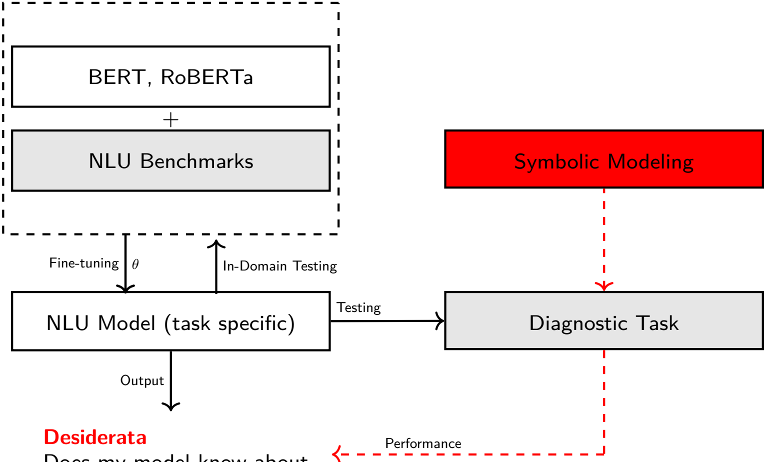
Diagnostic Tasks for NLU

- unit testing (Ribeiro et al., 2020), challenge tasks/stress tests (**task specific**) (Naik et al., 2018; Glockner et al., 2018), *inter alia*.



Diagnostic Tasks for NLU

- unit testing (Ribeiro et al., 2020), challenge tasks/stress tests (**task specific**) (Naik et al., 2018; Glockner et al., 2018), *inter alia*.



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

<Building Diagnostic Tasks>

(3 Example Studies)

Diagnostic Tasks via Expert Knowledge (Richardson and Sabharwal, 2020)[TACL]

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

Diagnostic Tasks via Expert Knowledge (Richardson and Sabharwal, 2020)[TACL]

- 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..

Diagnostic Tasks via Expert Knowledge (Richardson and Sabharwal, 2020)[TACL]

- 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`	...
...	...	...	...

Diagnostic Tasks via Expert Knowledge (Richardson and Sabharwal, 2020)[TACL]

- 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**

Diagnostic Tasks via Expert Knowledge (Richardson and Sabharwal, 2020)[TACL]

- 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 comfort-ably*	DEF	*The baby nestled her head..*
elude.v	escape.v	ISA	*The thief eluded po-lice...*
trouser.n	consumer good.n	ISA	*The man bought trousers..*
poet.n	writer.n	ISA	...
...	...	...	...

templates

Probing Questions

Question	Answer	Test
Given *'The baby nes-tled her head'*, **nes-tled** is defined as	*position comfort-ably*	**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

Diagnostic Tasks via Expert Knowledge (Richardson and Sabharwal, 2020)[TACL]

- 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.** ✓)

Diagnostic Tasks via Expert Knowledge (Richardson and Sabharwal, 2020)[TACL]

- 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 QA Diagnostics

- **Resources**: WordNet, GCIDE dictionary; **5 individual tasks**: Definitions, Synonymy, Hypernymy (ISA), and Hyponymy (ISA), WordSense.
  - WordNet tasks involve $\sim 30k$ 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) ...

Semantic Fragments for NLI Richardson et al. (2020)[AAAI]

- **Semantic Fragment**: subset of language *equipped with semantics which translate into some formal system* ... (Pratt-Hartmann, 2004)

Semantic Fragments for NLI Richardson et al. (2020)[AAAI]

- **Semantic Fragment**: subset of language *equipped with semantics which translate into some formal system* ... (Pratt-Hartmann, 2004)

Formal Specification of Facts about Quantifiers (van Benthem (1986))

<u>`all`</u> X Y	$\models$	<u>`all`</u> X' Y', s.t. $X' \leq X, Y \leq Y'$
<u>`some`</u> X Y	$\models$	<u>`some`</u> X' Y', s.t. $X \leq X', \ldots$
<u>`exactly`</u> N X ..	$\models$	....

Semantic Fragments for NLI Richardson et al. (2020)[AAAI]

- **Semantic Fragment**: subset of language *equipped with semantics which translate into some formal system ...* (Pratt-Hartmann, 2004)

Formal Specification of Facts about Quantifiers (van Benthem (1986))

$\underline{\texttt{all}}\ X\ Y$	$\vDash$	$\underline{\texttt{all}}\ X'\ Y'$, s.t. $X' \leq X, Y \leq Y'$
$\underline{\texttt{some}}\ X\ Y$	$\vDash$	$\underline{\texttt{some}}\ X'\ Y'$, s.t. $X \leq X', \ldots$
$\underline{\texttt{exactly}}\ N\ X\ ..$	$\vDash$	....

symbolic model+generator+lexicon

Example Semantic Fragment

All dogs ran $\vDash$ *All small dogs ran, All furry dogs barked* $\underset{\Uparrow}{\vDash}$ *All animals barked, Some dog ran* $\vDash$ *Some animal moved,...*

Semantic Fragments for NLI Richardson et al. (2020)[AAAI]

- **Semantic Fragment**: subset of language *equipped with semantics which translate into some formal system ...* (Pratt-Hartmann, 2004)

Formal Specification of Facts about Quantifiers (van Benthem (1986))

`all` $X\ Y$	$\models$	`all` $X'\ Y'$, s.t. $X' \leq X, Y \leq Y'$
`some` $X\ Y$	$\models$	`some` $X'\ Y'$, s.t. $X \leq X', \ldots$
`exactly` $N\ X$ ..	$\models$	....

symbolic model+generator+lexicon

Example Semantic Fragment

All dogs ran $\models$ *All small dogs ran, All furry dogs barked* $\models$ *All animals barked, Some dog ran* $\models$ *Some animal moved,...*

NLI format, standard splitting,..

Diagnostic Task

Semantic Fragments for NLI Richardson et al. (2020)[AAAI]

- **Semantic Fragment**: subset of language *equipped with semantics which translate into some formal system ...* (Pratt-Hartmann, 2004)

Formal Specification of Facts about Quantifiers (van Benthem (1986))

$\underline{\texttt{all}}\ X\ Y \models \underline{\texttt{all}}\ X'\ Y'$, s.t. $X' \leq X, Y \leq Y'$

$\underline{\texttt{some}}\ X\ Y \models \underline{\texttt{some}}\ X'\ Y'$, s.t. $X \leq X', \ldots$

$\underline{\texttt{exactly}}\ N\ X\ .. \models \ldots.$

symbolic model+generator+lexicon

Example Semantic Fragment

All dogs ran $\models$ *All small dogs ran, All furry dogs barked* $\models$ *All animals barked, Some dog ran* $\models$ *Some animal moved,...*

NLI format, standard splitting,..

Diagnostic Task

Non-standard in NLP: Using symbolic models (vs. humans) to elicit data; standard tool in linguistics (Montague (1973)).

Semantic Fragments for NLI Richardson et al. (2020)[AAAI]

- **7 Tasks:** elementary logic (e.g., boolean algebra, quantification, conditionals) and monotonicity reasoning

Fragments	**Example** (`premise`,`label`,`hypothesis`)
`Negation`	*Laurie has only visited Nephi, Marion has only visited Calistoga.* `CONTRADICTION` *Laurie didn't visit Nephi*
`Boolean`	*Travis, Arthur, Henry and Dan have only visited Georgia* `ENTAILMENT` *Dan didn't visit Rwanda*
`Quantifier`	*Everyone has visited every place* `NEUTRAL` *Virgil didn't visit Barry*
`Counting`	*Nellie has visited Carrie, Billie, John, Mike, Thomas, Mark, .., and Arthur.* `ENTAILMENT` *Nellie has visited more than 10 people.*
`Conditionals`	*Francisco has visited Potsdam and if Francisco has visited Potsdam then Tyrone has visited Pampa* `ENTAILMENT` *Tyrone has visited Pampa.*
`Comparatives`	*John is taller than Gordon and Erik..., and Mitchell is as tall as John* `NEUTRAL` *Erik is taller than Gordon.*
`Monotonicity`	*All black mammals saw exactly 5 stallions who danced* `ENTAILMENT` *A brown or black poodle saw exactly 5 stallions who danced*

Semantic Fragments for NLI Richardson et al. (2020)[AAAI]

- **7 Tasks:** elementary logic (e.g., boolean algebra, quantification, conditionals) and monotonicity reasoning

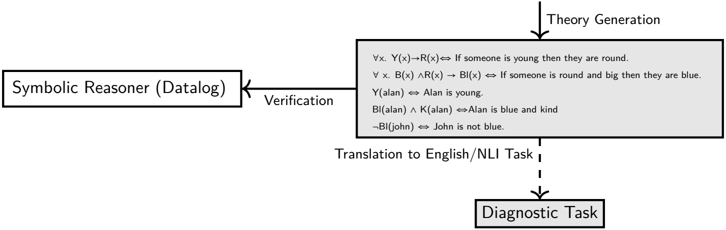
Fragments	**Example** (`premise`,`label`,`hypothesis`)
`Negation`	*Laurie has only visited Nephi, Marion has only visited Calistoga.* `CONTRADICTION` *Laurie didn't visit Nephi*
`Boolean`	*Travis, Arthur, Henry and Dan have only visited Georgia* `ENTAILMENT` *Dan didn't visit Rwanda*
`Quantifier`	*Everyone has visited every place* `NEUTRAL` *Virgil didn't visit Barry*
`Counting`	*Nellie has visited Carrie, Billie, John, Mike, Thomas, Mark, .., and Arthur.* `ENTAILMENT` *Nellie has visited more than 10 people.*
`Conditionals`	*Francisco has visited Potsdam and if Francisco has visited Potsdam then Tyrone has visited Pampa* `ENTAILMENT` *Tyrone has visited Pampa.*
`Comparatives`	*John is taller than Gordon and Erik..., and Mitchell is as tall as John* `NEUTRAL` *Erik is taller than Gordon.*
`Monotonicity`	*All black mammals saw exactly 5 stallions who danced* `ENTAILMENT` *A brown or black poodle saw exactly 5 stallions who danced*

Done in collaboration with logicians and linguists (Indiana University); generated using simple templates , formal grammars.

Rule Taker: Training Models to do Formalized Reasoning.

(Clark et al., 2020)[IJCAI]

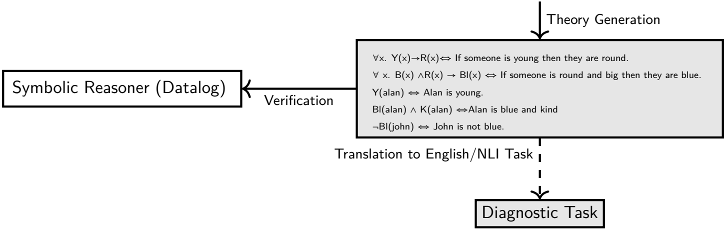
- **Idea**: Synthesize deductively valid entailment data (theories and queries) with the help of symbolic theorem prover; render as English.



Rule Taker: Training Models to do Formalized Reasoning.

(Clark et al., 2020)[IJCAI]

- **Idea**: Synthesize deductively valid entailment data (theories and queries) with the help of symbolic theorem prover; render as English.



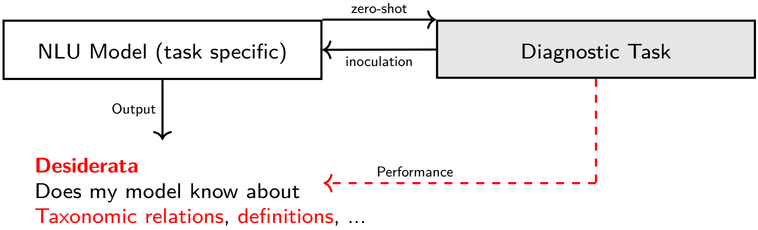
Bigger idea: demonstrating model correctness can be achieved by testing on data that is *correct by construction*.

- **Components**: reasoning depth, vocabulary overlap/mismatch.

Probing Methodology and Experiments (QA+NLI Tasks)

- Trained single models on standard benchmarks; **Ask the following empirical questions**:

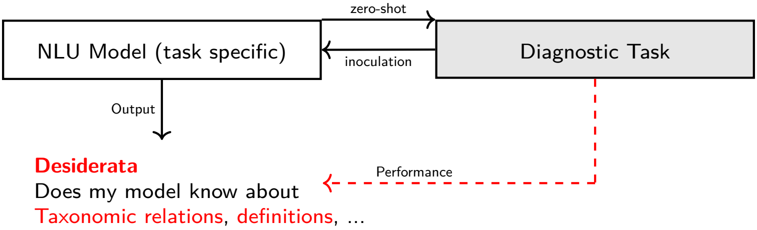
  1. How well do benchmark models perform on each *individual* probing on diagnostic task without specialized training (**zero-shot**)?
  2. How well models perform after a small amount of additional training on probes (**inoculation** (Liu et al., 2019))?



Probing Methodology and Experiments (QA+NLI Tasks)

- Trained single models on standard 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))?



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

What happens when we do unit testing?

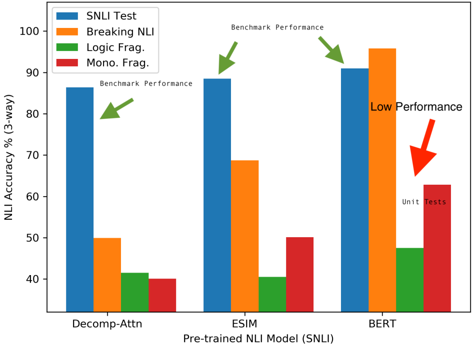
(NLI $\rightarrow$ QA)

Result 1: It's easy to find bugs: zero-shot (NLI)

- **Approach**: do testing on models trained on benchmark tasks, look at difference in performance; the difference is usually large.

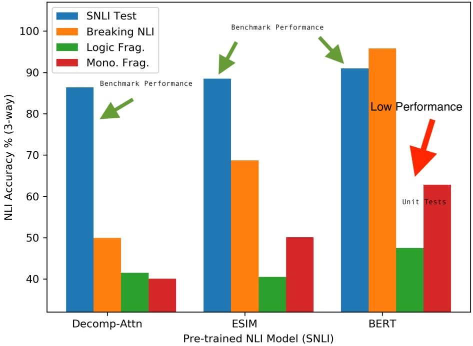
Result 1: It's easy to find bugs: zero-shot (NLI)

- **Approach**: do testing on models trained on benchmark tasks, look at difference in performance; the difference is usually large.



Result 1: It's easy to find bugs: zero-shot (NLI)

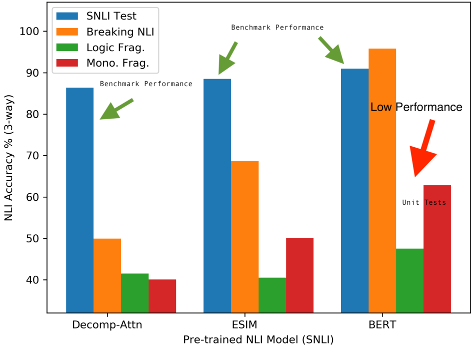
- **Approach**: do testing on models trained on benchmark tasks, look at difference in performance; the difference is usually large.



NLI models trained on standard benchmarks are still lacking in basic linguistic and reasoning abilities.

Result 1: It's easy to find bugs: zero-shot (NLI)

- **Approach**: do testing on models trained on benchmark tasks, look at difference in performance; the difference is usually large.



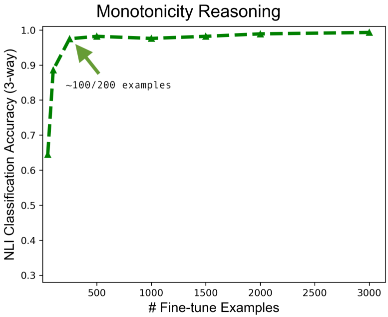
Caveats: Do models really not posses the target knowledge, or lack knowledge of format?

Result 2: Patching bugs can be easy (best NLI models)

- **Model Inoculation** (Richardson et al., 2020)[AAAI]: Continue training models on small amounts of diagnostic; aim to (quickly) fix model.

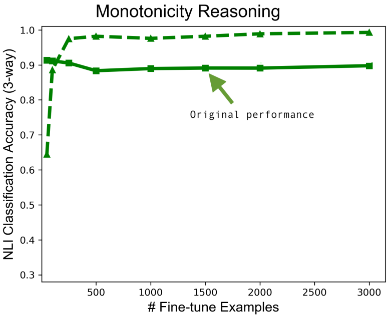
Result 2: Patching bugs can be easy (best NLI models)

- **Model Inoculation** (Richardson et al., 2020)[AAAI]: Continue training models on small amounts of diagnostic; aim to (quickly) fix model.



Result 2: Patching bugs can be easy (best NLI models)

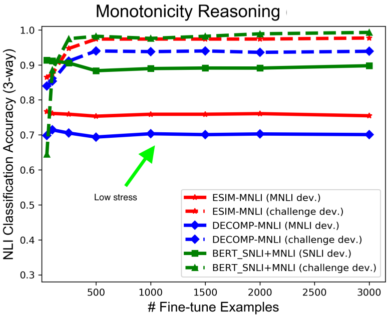
- **Model Inoculation** (Richardson et al., 2020)[AAAI]: Continue training models on small amounts of diagnostic; aim to (quickly) fix model.



Assumption: Ability of model to quickly learn new tasks with minimal effect on original task indicates competence.

Result 2: Patching bugs can be easy (best NLI models)

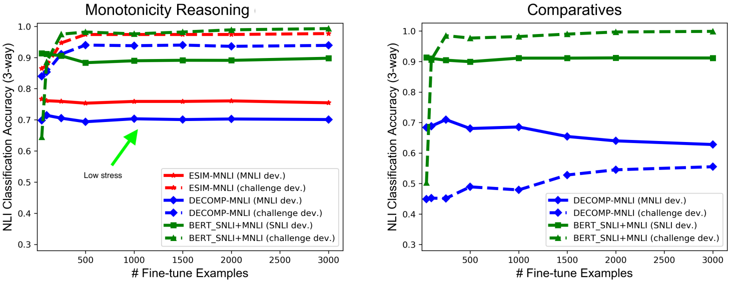
- **Model Inoculation** (Richardson et al., 2020)[AAAI]: Continue training models on small amounts of diagnostic; aim to (quickly) fix model.



Assumption: Ability of model to quickly learn new tasks with minimal effect on original task indicates competence.

Result 2: Patching bugs can be easy (best NLI models)

- **Model Inoculation** (Richardson et al., 2020)[AAAI]: Continue training models on small amounts of diagnostic; aim to (quickly) fix model.



Assumption: Ability of model to quickly learn new tasks with minimal effect on original task indicates competence.

Result 2: Patching bugs can be easy (best NLI models)

- **Model Inoculation** (Richardson et al., 2020)[AAAI]: Continue training models on small amounts of diagnostic; aim to (quickly) fix model.

Category ↓, Model →	**BERT Transformer**	**ESIM**	**DecAttn**
`Conditionals`	☺	☺	☺
`Counting`	☺	☺	☹
`Quantifiers`	☺	☹	☹
`Negation`	☺	☹	☹
`Boolean Coordination`	☺	☹	☹
`Comparatives`	☺	☹	☹

☹ = (bad/mediocre performance + forgetting on **test**), ☺ = (high performance + minimal forgetting on **test**)

Result 3. Transformers do have lexical knowledge (QA)

- **Zero-shot**, models do well on *some* categories of knowledge ; far outpace baselines trained on diagnostics.

Result 3. Transformers do have lexical knowledge (QA)

- **Zero-shot**, models do well on *some* categories of knowledge; far outpace baselines trained on diagnostics.

Diagnostic performance (QA Accuracy %; random ~ 20%)

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%

Result 3. Transformers do have lexical knowledge (QA)

- **Zero-shot**, models do well on *some* categories of knowledge ; far outpace baselines trained on diagnostics.

Diagnostic performance (QA Accuracy %; random ~ 20%)

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).

Result 3. Transformers do have lexical knowledge (QA)

- **Inoculation setting**: Models quickly start reaching near human performance .

Diagnostic performance (QA Accuracy %; random ~ 20%)

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%

Result 3. Transformers do have lexical knowledge (QA)

- **Inoculation setting**: Models quickly start reaching near human performance .

Diagnostic performance (QA Accuracy %; random ~ 20%)

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.

Result 3. Looking a bit deeper... (QA)

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

Result 3. Looking a bit deeper... (QA)

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

LSTM 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

Distractor Types

Result 3. Looking a bit deeper... (QA)

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

ISA reasoning 3 steps
moderate distractors

LSTM 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

Distractor Types

Result 3. Looking a bit deeper... (QA)

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

ISA reasoning 3 steps,
moderate distractors

LSTM 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	$\boxed{0.29}$	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

Distractor Types

(QA task) + 100 ex.

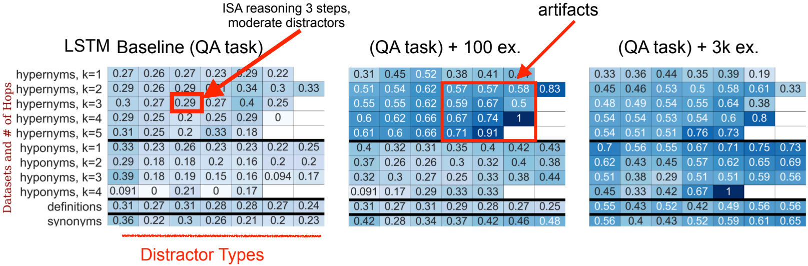
Datasets and # of Hops							
hypernyms, k=1	0.31	0.45	0.52	0.38	0.41	0.44	
hypernyms, k=2	0.51	0.54	0.62	0.57	0.57	0.58	0.83
hypernyms, k=3	0.55	0.55	0.62	0.59	0.67	0.5	
hypernyms, k=4	0.6	0.62	0.66	0.67	0.74	1	
hypernyms, k=5	0.61	0.6	0.66	0.71	0.91		
hyponyms, k=1	0.4	0.32	0.31	0.35	0.4	0.42	0.43
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

Result 3. Looking a bit deeper... (QA)

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



LSTM 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	[illegible]	
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

Result 3. Looking a bit deeper... (QA)

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

ISA reasoning 3 steps, moderate distractors

LSTM 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

Result 3. Looking a bit deeper... (QA)

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

ISA reasoning 3 steps, moderate distractors

Datasets and # of Hops

LSTM 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

Datasets and # of Hops

SOTA Transformer (QA task)

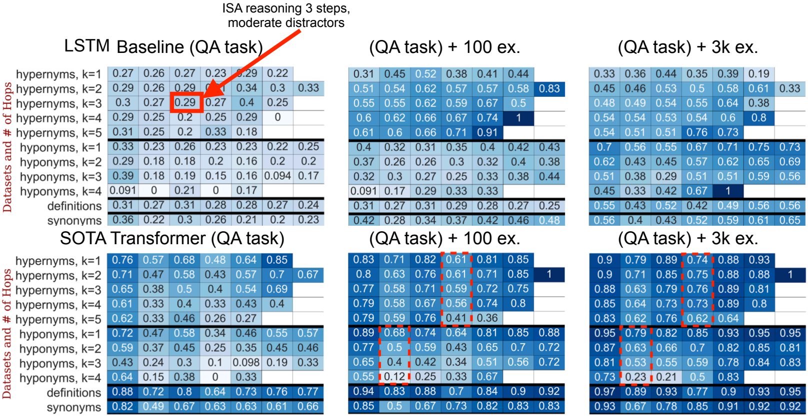
hypernyms, k=1	0.76	0.57	0.68	0.48	0.64	0.85	
hypernyms, k=2	0.71	0.47	0.58	0.43	0.57	0.7	0.67
hypernyms, k=3	0.65	0.38	0.5	0.4	0.54	0.69	
hypernyms, k=4	0.61	0.33	0.4	0.33	0.43	0.4	
hypernyms, k=5	0.62	0.33	0.46	0.26	0.27		
hyponyms, k=1	0.72	0.47	0.58	0.34	0.46	0.55	0.57
hyponyms, k=2	0.59	0.37	0.45	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

Result 3. Looking a bit deeper... (QA)

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



Can nudge models to bring out knowledge with small set of examples, cheap way to **inject knowledge** into transformers, **can be used as KBs**.

Result 3. Looking a bit deeper... (QA)

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

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

(QA task) + 100 ex.

0.83	0.71	0.82	0.61	0.81	0.85	
0.8	0.63	0.76	0.61	0.71	0.85	1
0.77	0.58	0.71	0.59	0.72	0.75	
0.79	0.58	0.67	0.56	0.74	0.8	
0.79	0.59	0.76	0.41	0.36		
0.89	0.68	0.74	0.64	0.81	0.85	0.88
0.77	0.5	0.59	0.43	0.65	0.7	0.72
0.65	0.4	0.42	0.34	0.51	0.56	0.72
0.55	0.12	0.25	0.33	0.67		
0.94	0.83	[illegible]	0.7	0.84	0.9	0.92
0.85	0.5	0.67	[illegible]	0.82	0.83	0.83

easy cases

(QA task) + 3k ex.

0.9	0.79	0.89	0.74	0.88	0.93	
0.9	0.71	0.85	0.75	0.88	0.88	1
0.88	0.63	0.79	0.76	0.89	0.81	
0.85	0.64	0.73	0.73	0.89	0.8	
0.83	0.62	0.76	0.62	0.64		
0.95	0.79	0.82	0.85	0.93	0.95	0.95
0.87	0.63	0.66	0.7	0.82	0.85	0.81
0.81	0.53	0.55	0.59	0.78	0.84	0.83
[illegible]	0.23	0.21	0.5	0.83		
[illegible]	0.89	0.93	0.77	0.9	0.93	0.95
0.93	0.67	0.78	0.85	0.91	0.92	0.92

hard cases

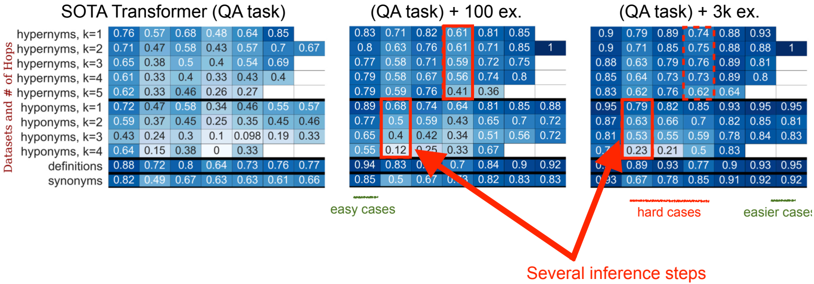
easier cases

Several inference steps

Model does show sensitivity to reasoning complexity, weak areas.

Result 3. Looking a bit deeper... (QA)

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

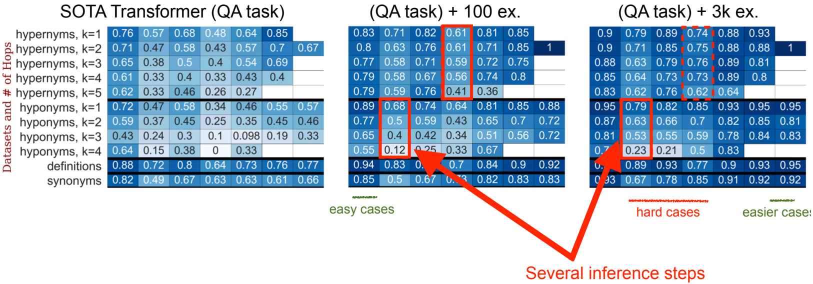


Model does show sensitivity to reasoning complexity; weak areas.

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

Result 3. Looking a bit deeper... (QA)

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



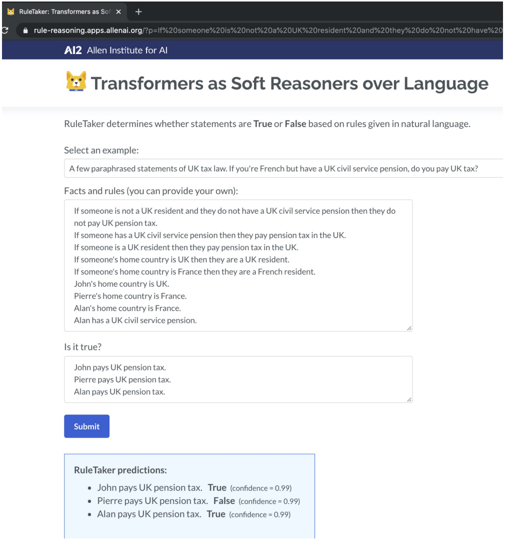
Model does show sensitivity to reasoning complexity; weak areas.

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

Probing is difficult! Definitive proof of model knowledge is difficult.

Transformers for Formalized (Deductive) Reasoning

- Achieve high accuracy on reasoning tasks (99% accuracy); Ability to generalize to human authored theories , reasoning depths.



</Results>

Conclusions

- Probing using symbolic models, tasks with 1. wide range of concepts 2. systematic perturbations ; 3. variable complexity.
  - Useful for better understanding models, supplement to existing NLU research; **few-shot learning** for model fixing.
  - Inherently collaborative enterprise, need help!

- several diagnostic tasks for QA, NLI; **extending to other tasks,** behavioral testing + **interventions** (manipulations of network states) (Geiger et al., 2020)[BlackBoxNLP].

Conclusions

- Probing using symbolic models, tasks with 1. wide range of concepts 2. systematic perturbations ; 3. variable complexity.
  - Useful for better understanding models, supplement to existing NLU research; **few-shot learning** for model fixing.
  - Inherently collaborative enterprise, need help!

- several diagnostic tasks for QA, NLI; **extending to other tasks,** behavioral testing + **interventions** (manipulations of network states) (Geiger et al., 2020)[BlackBoxNLP].

  Tooling: allowing non-experts to author their own datasets, *democratize* the dataset construction process.

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*.

Clark, P., Tafjord, O., and Richardson, K. (2020). Transformers as soft reasoners over language. *Proceedings of IJCAI*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Geiger, A., Richardson, K., and Potts, C. (2020). Modular representation underlies systematic generalization in neural natural language inference models. *arXiv preprint arXiv:2004.14623*.

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

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. In *Proceedings of NAACL: HLT*.

References II

Khot, T., Khashabi, D., Richardson, K., Clark, P., and Sabharwal, A. (2020). Text Modular Networks: Learning to Decompose Tasks in the Language of Existing Models. *arXiv preprint arXiv:2009.00751*.

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*.

Montague, R. (1973). The proper treatment of quantification in ordinary english. In *Approaches to natural language*, pages 221–242. Springer.

Naik, A., Ravichander, A., Sadeh, N., Rose, C., and Neubig, G. (2018). Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*.

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*.

Pratt-Hartmann, I. (2004). Fragments of language. *Journal of Logic, Language and Information*, 13(2):207–223.

Pratt-Hartmann, I. (2015). Semantic complexity in natural language. *The Handbook of Contemporary Semantic Theory*, page 429.

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

References III

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

Richardson, K. and Sabharwal, A. (2020). What Does My QA Model Know? Devising Controlled Probes using Expert Knowledge. *to appear in TACL*.

Rozen, O., Shwartz, V., Aharoni, R., and Dagan, I. (2019). Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets. *arXiv preprint arXiv:1910.09302*.

van Benthem, J. (1986). *Essays in Logical Semantics*, volume 29 of *Studies in Linguistics and Philosophy*. D. Reidel Publishing Co., Dordrecht.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.