# Probing Natural Language Inference Models through Semantic Fragments

**Kyle Richardson**[†] Hai Hu[‡] Lawrence S. Moss[‡]
Ashish Sabharwal[†]

[†] Allen Institute for Artificial Intelligence
{kyler,ashishs}@allenai.org

[‡] Indiana University
{lmoss,huhai}@indiana.edu

February 9, 2020

# Probing Natural Language Understanding (NLU) Models

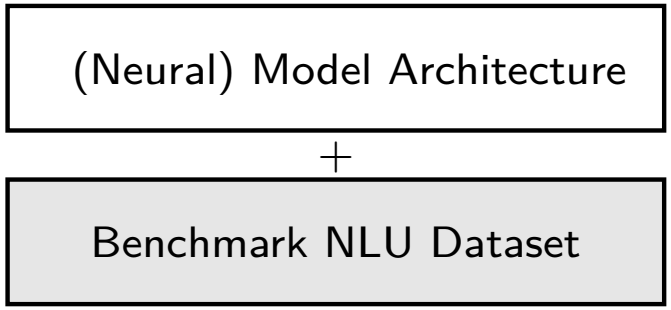
- **Probing**: trying to understand the strengths and weaknesses of models; measuring model performance and **competence** qualitatively.

# Building NLU Models: Standard Picture

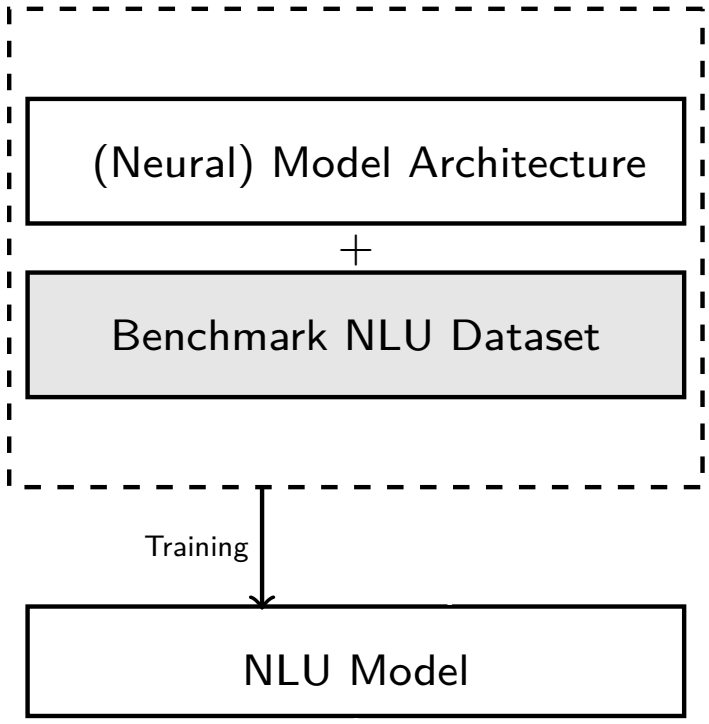
# Building NLU Models: Standard Picture

(Neural) Model Architecture

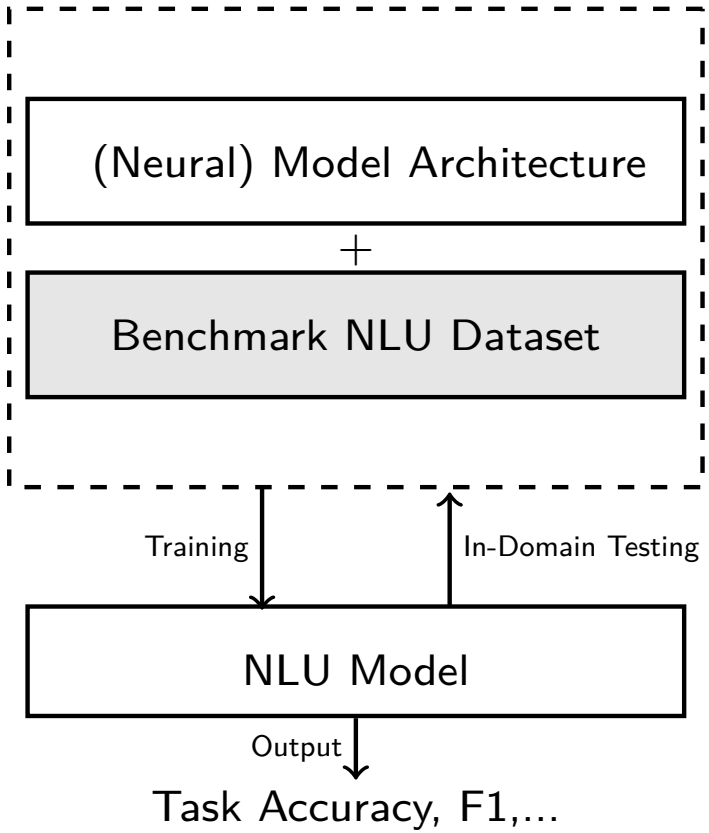
# Building NLU Models: Standard Picture



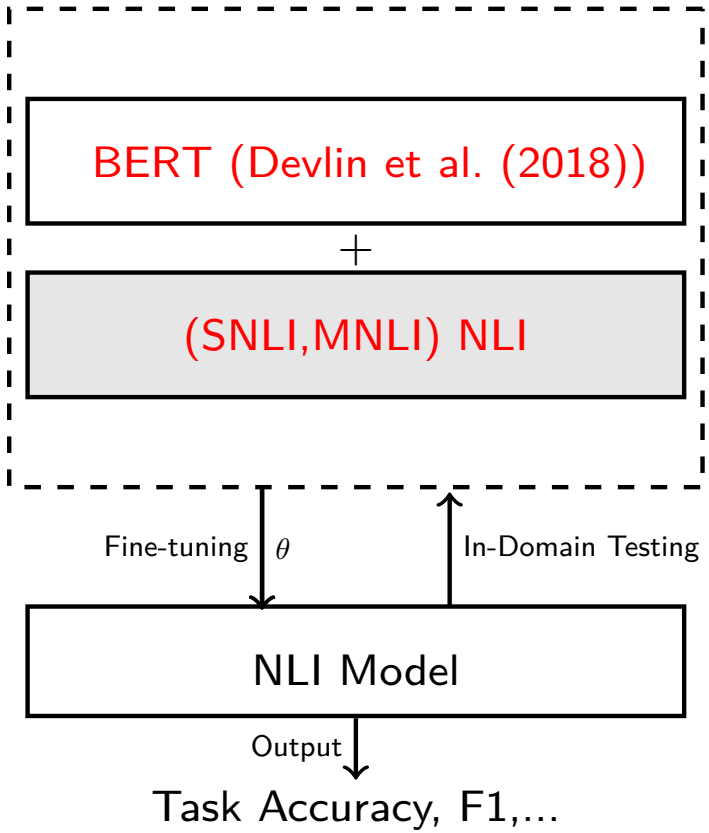
# Building NLU Models: Standard Picture



# Building NLU Models: Standard Picture

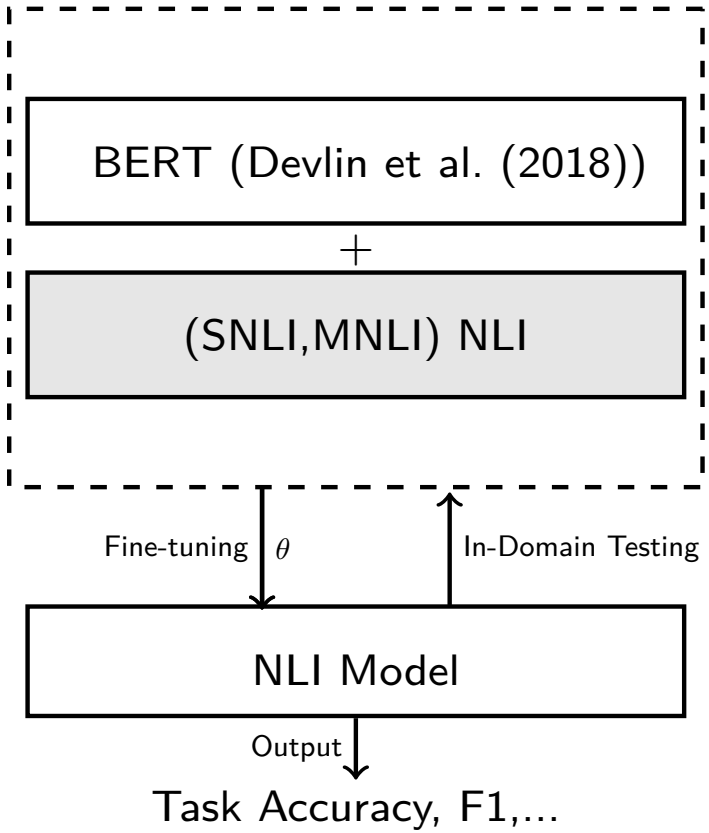


# Building NLU Models: Standard Picture



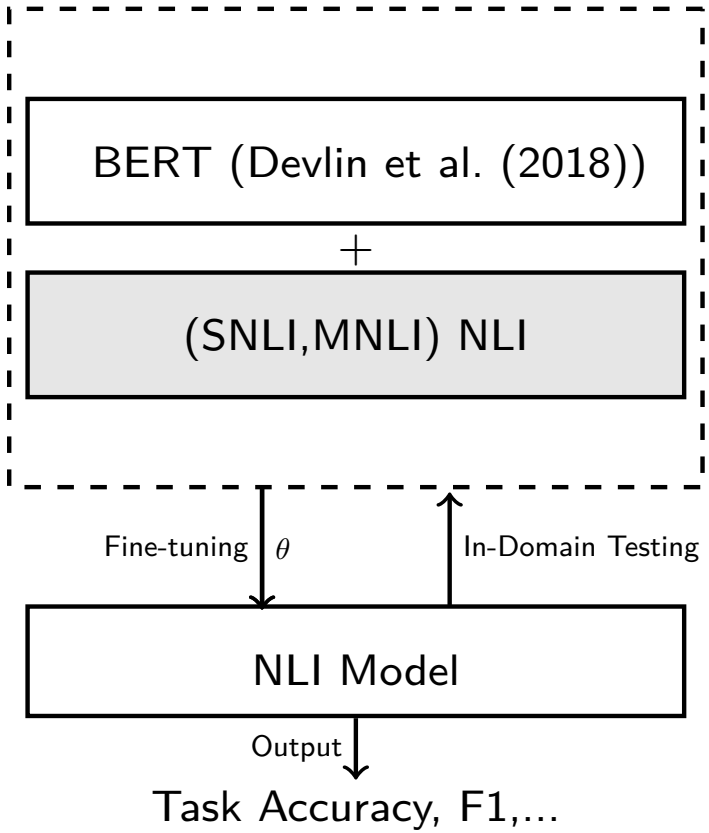

| **Stanford NLI Dataset (SNLI)** (Bowman et al. (2015)) | |
|---|---|
| **Premise** | *A soccer game with multiple males playing.* |
| **Hypothesis** | *Some men are playing a sport.* |
| **Label** | `Entailment` (= *meaning of* ***Hypothesis*** *is implied by* ***Premise***) |

# Building NLU Models: Standard Picture



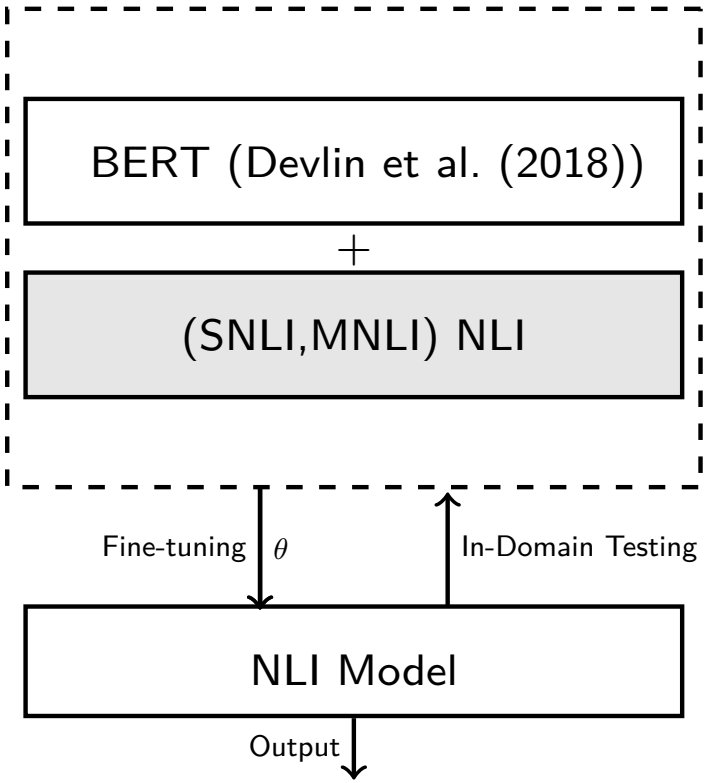

- **The Good:** Tremendous progress on benchmarks; **The Bad:** low interpretability; **The Ugly**:annotation artifacts (Gururangan et al. (2018)).

# Building NLU Models: Standard Picture




- **Bigger Issue:** (not often discussed) Unclear how linguists, logicians, people working in classical/symbolic AI fit into this picture.

# Probing NLU Models: What We Aim for Here




**Desiderata**

Does my model know about
quantification, negation, boolean algebra,
monotonicity, ....?

# This Talk

**Question**: Do NLI models that excel at standard tasks have the knowledge and reasoning abilities we expect them to? (can they be fixed?)

# This Talk

**Question**: Do NLI models that excel at standard tasks have the knowledge and reasoning abilities we expect them to? (can they be fixed?)

- **Contributions**:

# This Talk

**Question**: Do NLI models that excel at standard tasks have the knowledge and reasoning abilities we expect them to? (can they be fixed?)

- **Contributions**:
  - **New probing methodology** (3-step) centering around **semantic fragments**.

# This Talk

**Question**: Do NLI models that excel at standard tasks have the knowledge and reasoning abilities we expect them to? (can they be fixed?)

- **Contributions**:
  - **New probing methodology** (3-step) centering around **semantic fragments**.
  - **8 new diagnostic datasets** that probe *competence* of elementary logic and monotonicity reasoning.

# This Talk

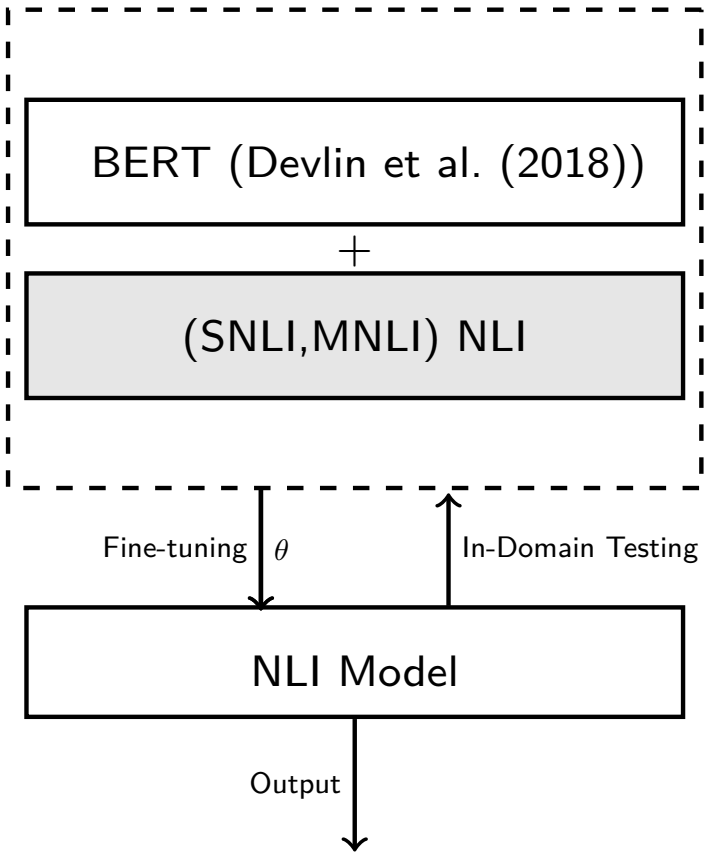
**Question**: Do NLI models that excel at standard tasks have the knowledge and reasoning abilities we expect them to? (can they be fixed?)

- **Contributions**:
  - **New probing methodology** (3-step) centering around **semantic fragments**.
  - **8 new diagnostic datasets** that probe *competence* of elementary logic and monotonicity reasoning.

Probing and measuring qualitative performance is difficult!

<Diagnostic Tasks/Semantic Fragments>

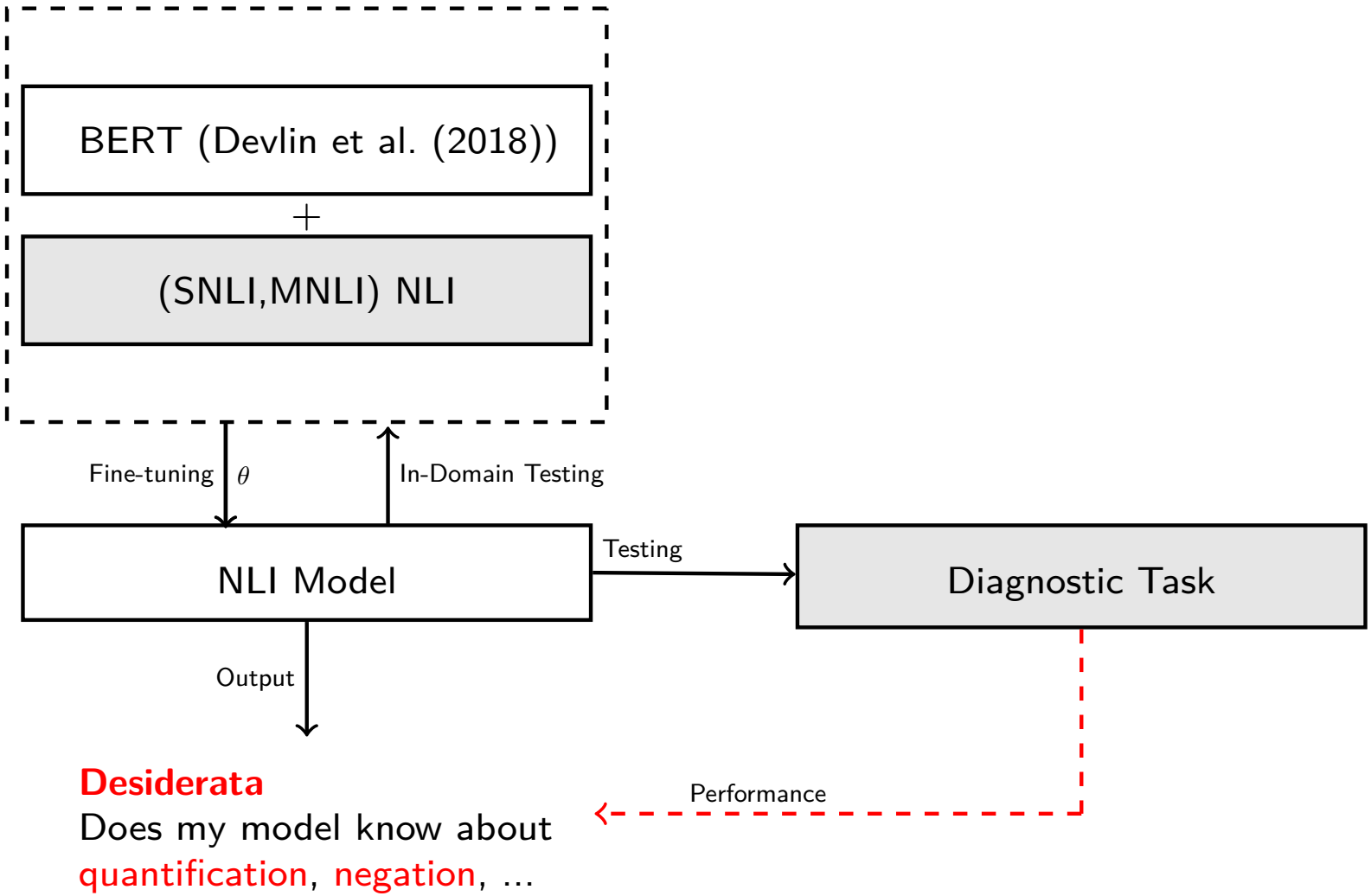
# Diagnostic Tasks for NLU




**Desiderata**

Does my model know about quantification, negation, ...

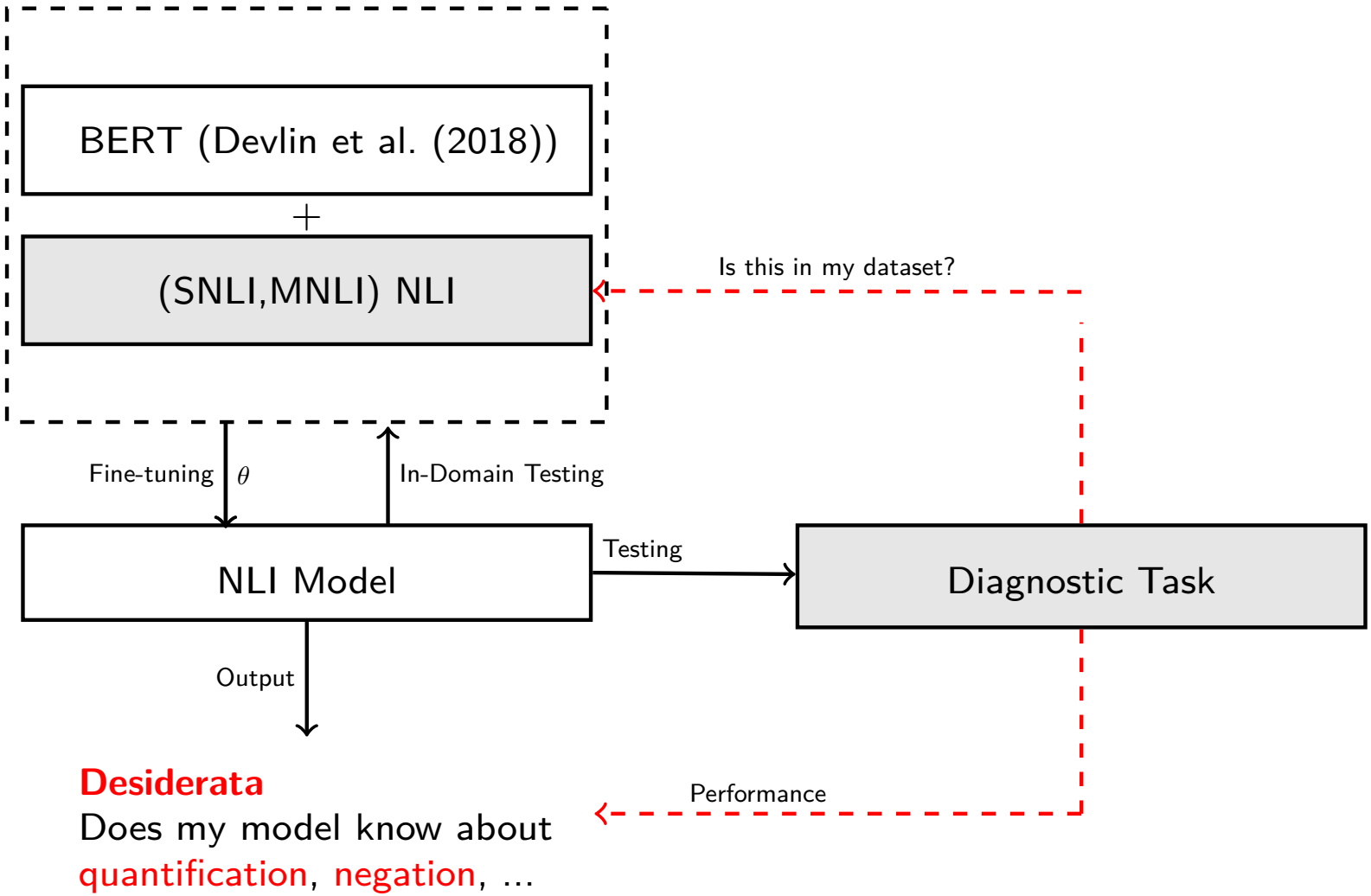
# Diagnostic Tasks for NLU




**Desiderata**

Does my model know about quantification, negation, ...

# Diagnostic Tasks for NLU

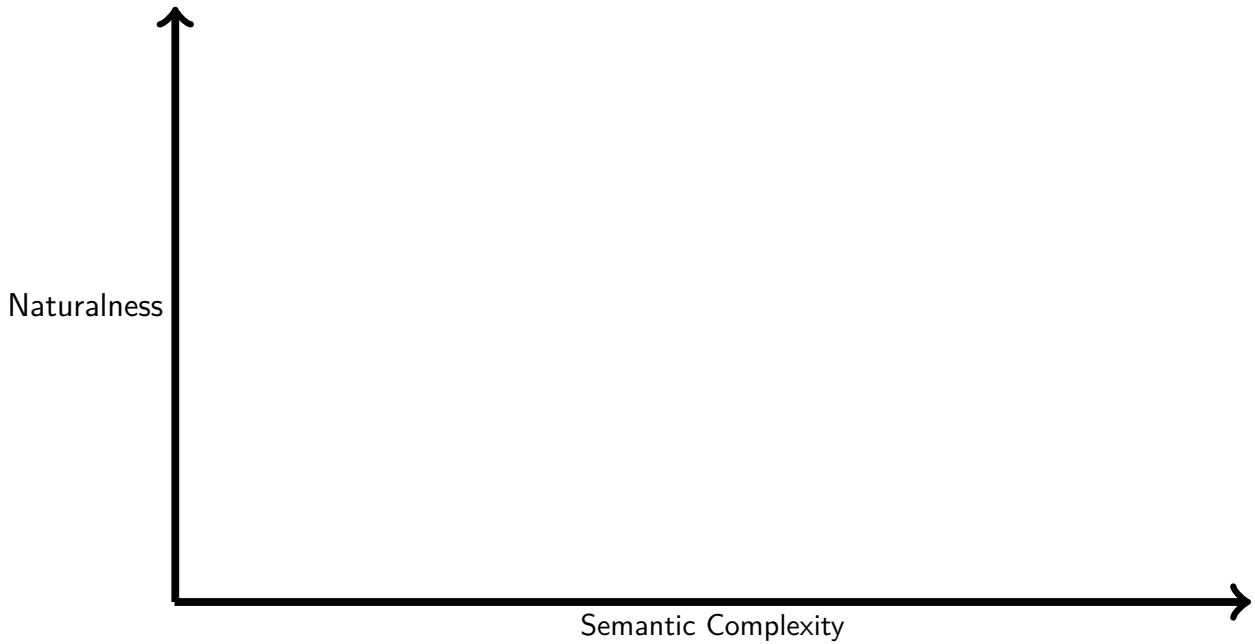



**Desiderata**

Does my model know about quantification, negation, ...

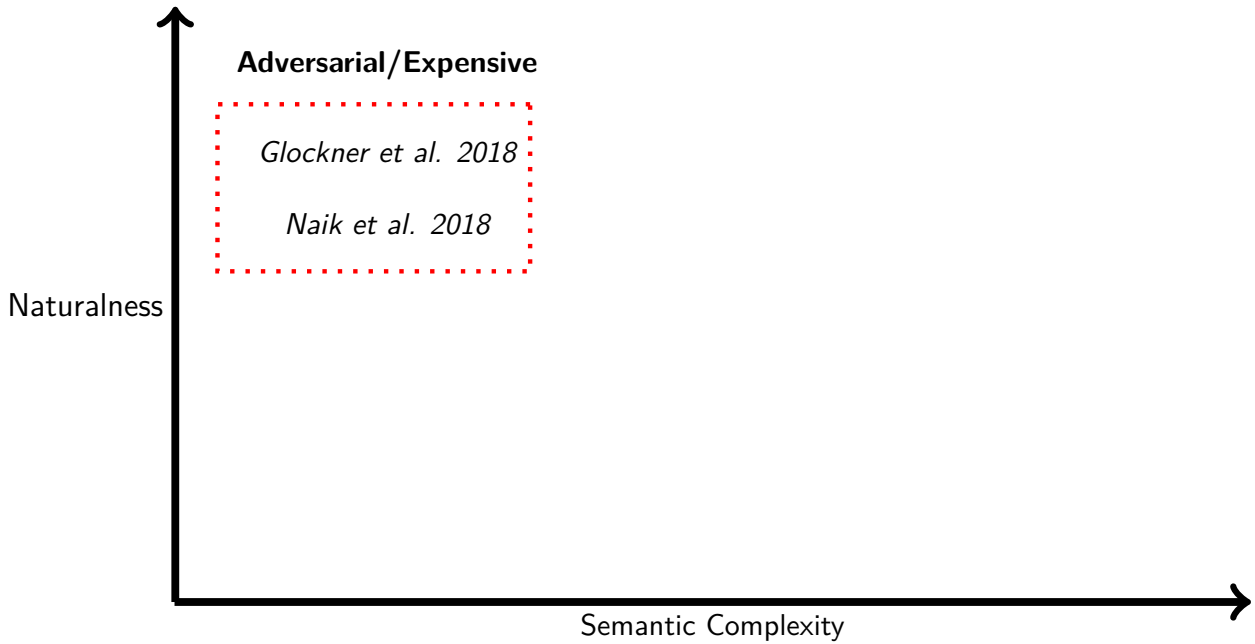
# Diagnostic Tasks in NLU

- Trade-off between **naturalness** and **semantic complexity**.



# Diagnostic Tasks in NLU

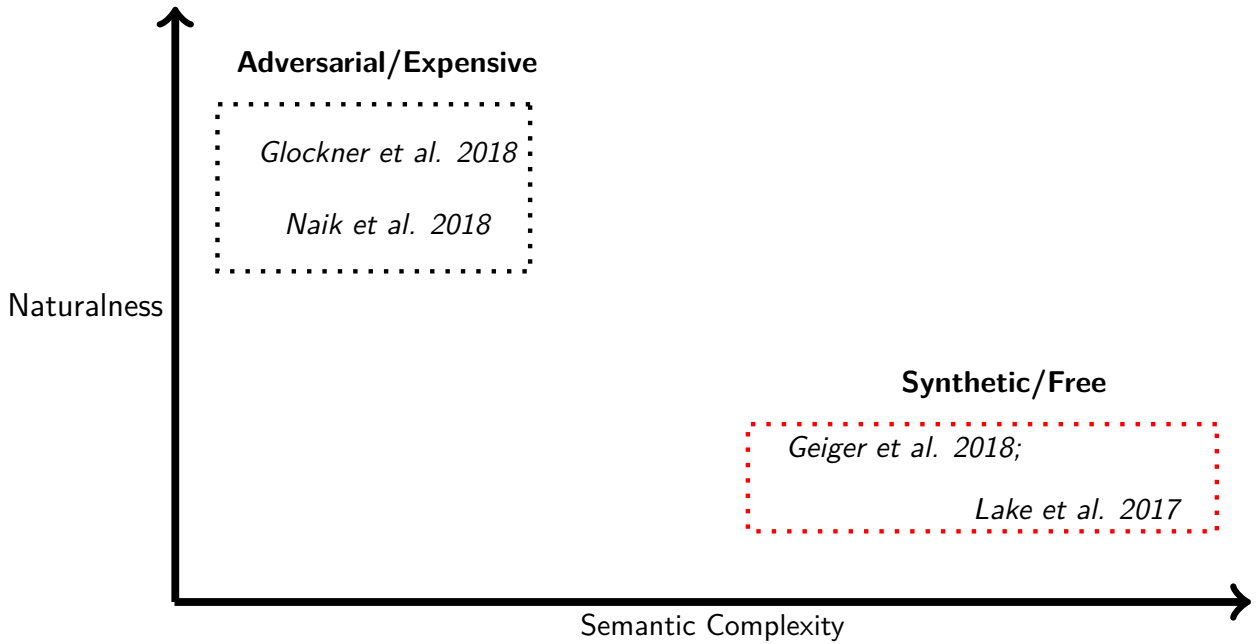
- Trade-off between **naturalness** and **semantic complexity**.




| **Breaking NLI Challenge** Glockner et al. (2018) | |
|---|---|
| **Premise** | *A soccer game with multiple males playing.* |
| **Hypothesis** | *Some ~~men~~ **women** are playing ~~a sport~~ **basketball**.* |

# Diagnostic Tasks in NLU

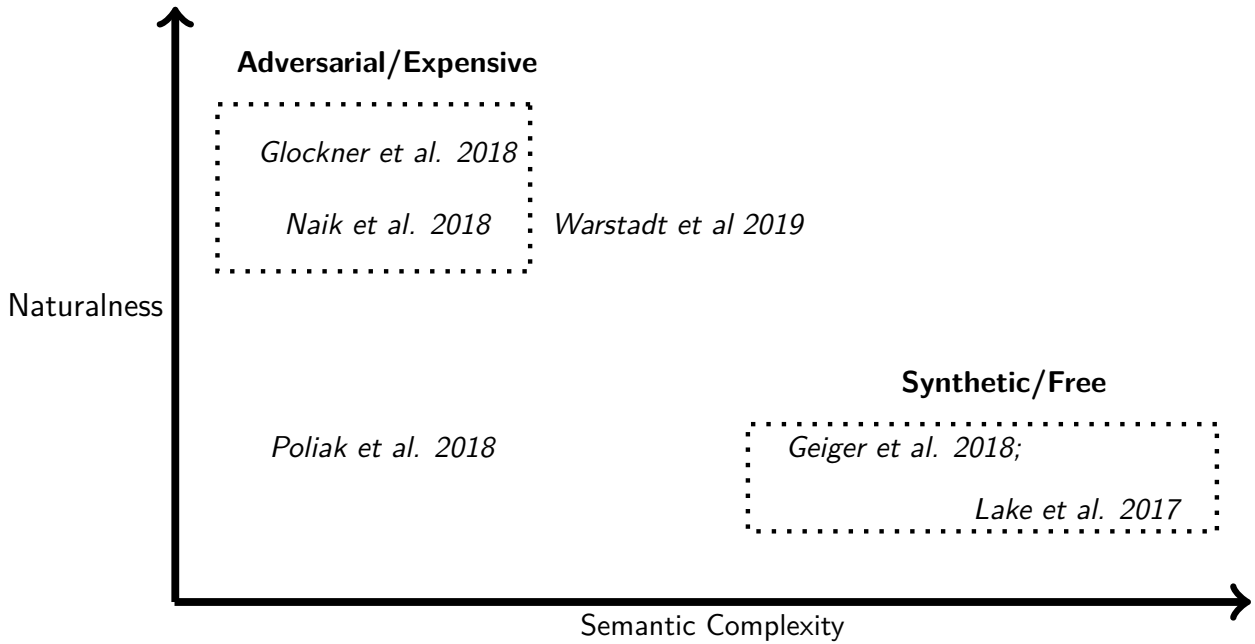
- Trade-off between **naturalness** and **semantic complexity**.




| **Stress Testing Multiple-Quantifier Sentences** Geiger et al. (2018) | |
|---|---|
| **Premise** | *emptystring servant does not carefully butters every pink baseball.* |
| **Hypothesis** | *every slimy servant does not carefully butters some pink baseball.* |

# Diagnostic Tasks in NLU

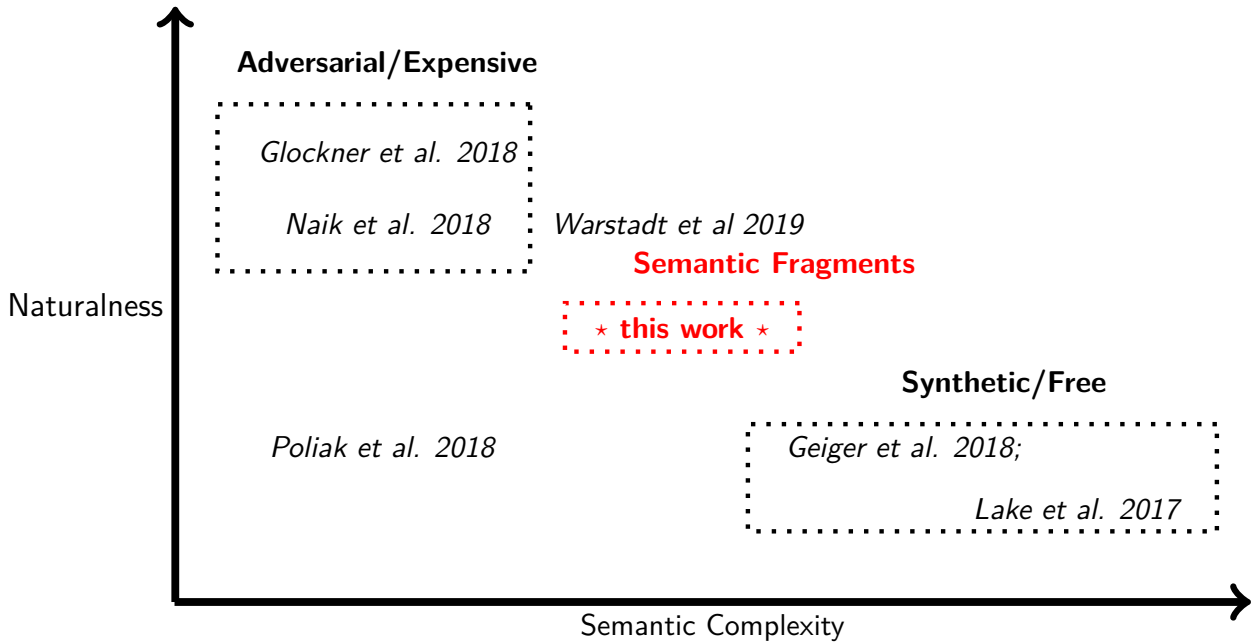
- Trade-off between **naturalness** and **semantic complexity**.




**Diverse NLI** Poliak et al. (2018); **Inference is Everything** White et al. (2017); Warstadt et al. (2019); McCoy et al. (2019), **GLUE Diagnostic** Wang et al. (2018)

# Diagnostic Tasks in NLU

- Trade-off between **naturalness** and **semantic complexity**.




**Semantic Fragments**: synthetic **linguistic fragments** that aim to capture **naturalistic** subsets of English. **Allow for systematic control.**

# Semantic Fragments: A Linguistic Approach to Probing

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

# Semantic Fragments: A Linguistic Approach to Probing

- **Semantic Fragment**: subset of language *equipped with semantics which translate sentences 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 $N$ $X$ .. | $\models$ | .... |

# Semantic Fragments: A Linguistic Approach to Probing

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

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

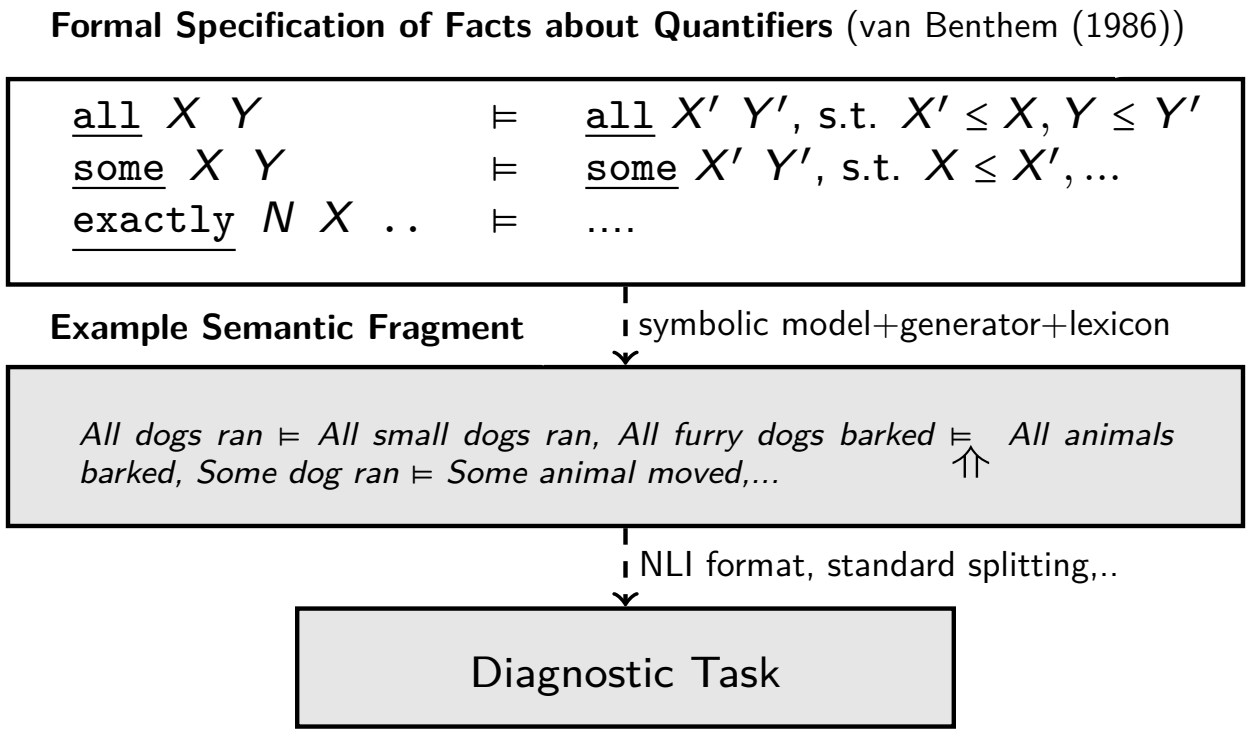
| | | |
|---|---|---|
| all $X$ $Y$ | $\models$ | all $X'$ $Y'$, s.t. $X' \leq X, Y \leq Y'$ |
| some $X$ $Y$ | $\models$ | some $X'$ $Y'$, s.t. $X \leq X', \ldots$ |
| exactly $N$ $X$ .. | $\models$ | .... |

**Example Semantic Fragment** symbolic model+generator+lexicon

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

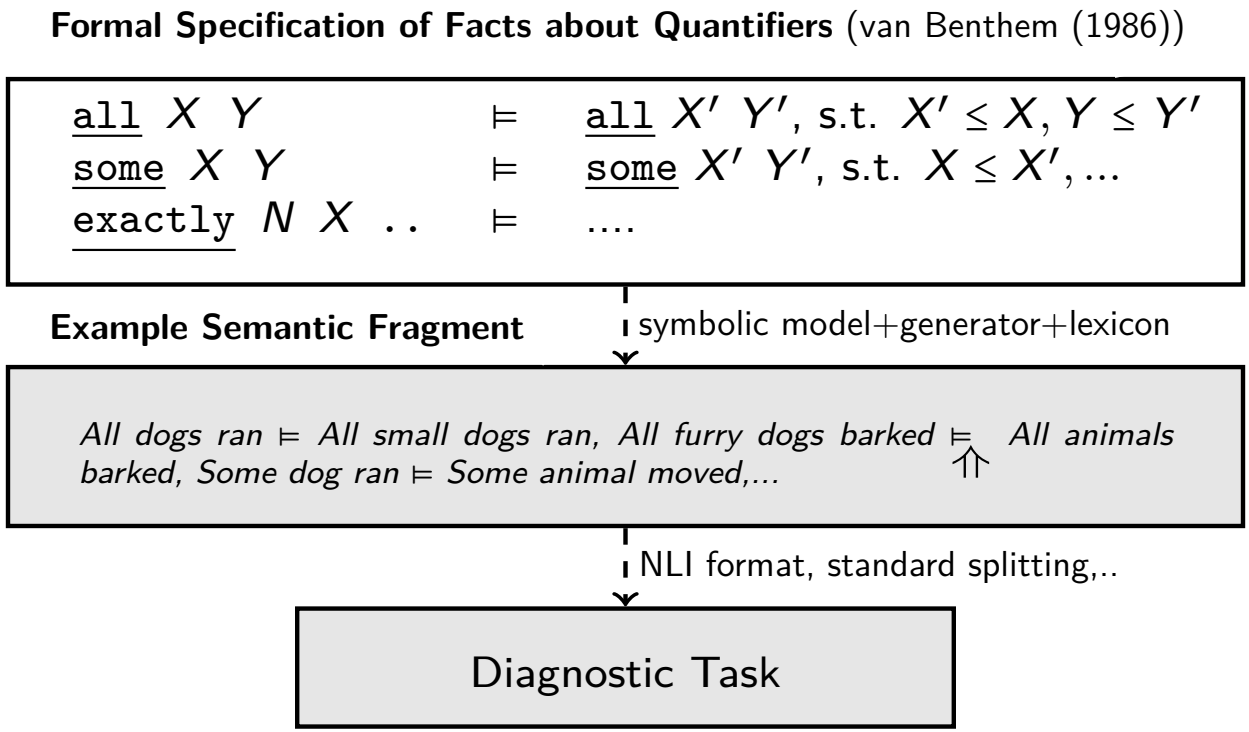
# Semantic Fragments: A Linguistic Approach to Probing

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



# Semantic Fragments: A Linguistic Approach to Probing

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




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

# 8 Example Fragments

- **Logic** (6 datasets): Test elementary logic, counting and aggregation; re-purposed from Salvatore et al. (2019); built using templates.
- **Monotonicity** (2 datasets; hard/easy): built using automatic polarity projection Hu and Moss (2018) and formal grammars.

| **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* |

# 8 Example Fragments

- Straddle the boundaries of ordinary and pedantic English; can involve long and complicated examples (e.g., transitive reasoning):

| **Premise** | **Hypothesis** | `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` |

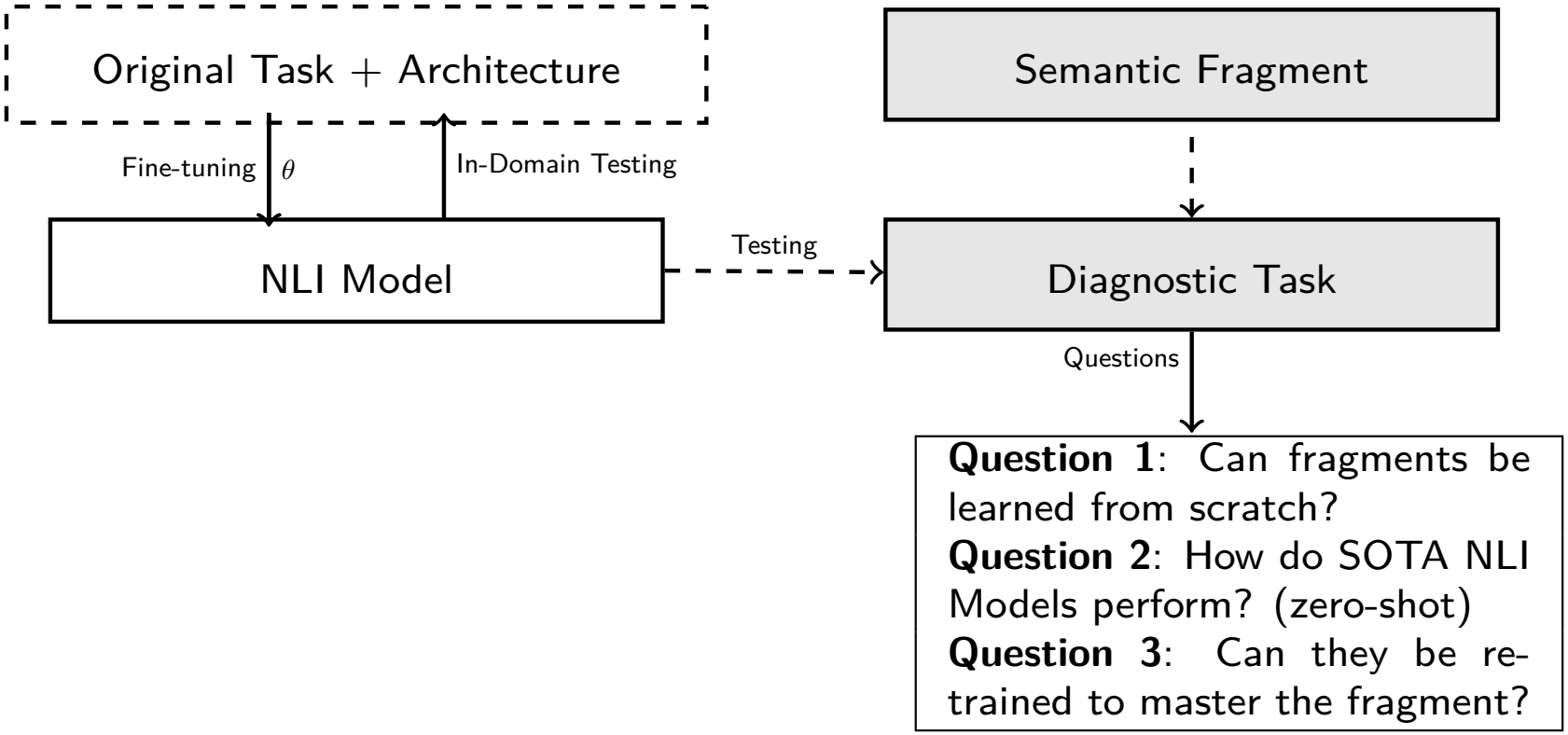
# 8 Example Fragments

- Straddle the boundaries of ordinary and pedantic English; can involve long and complicated examples (e.g., transitive reasoning):

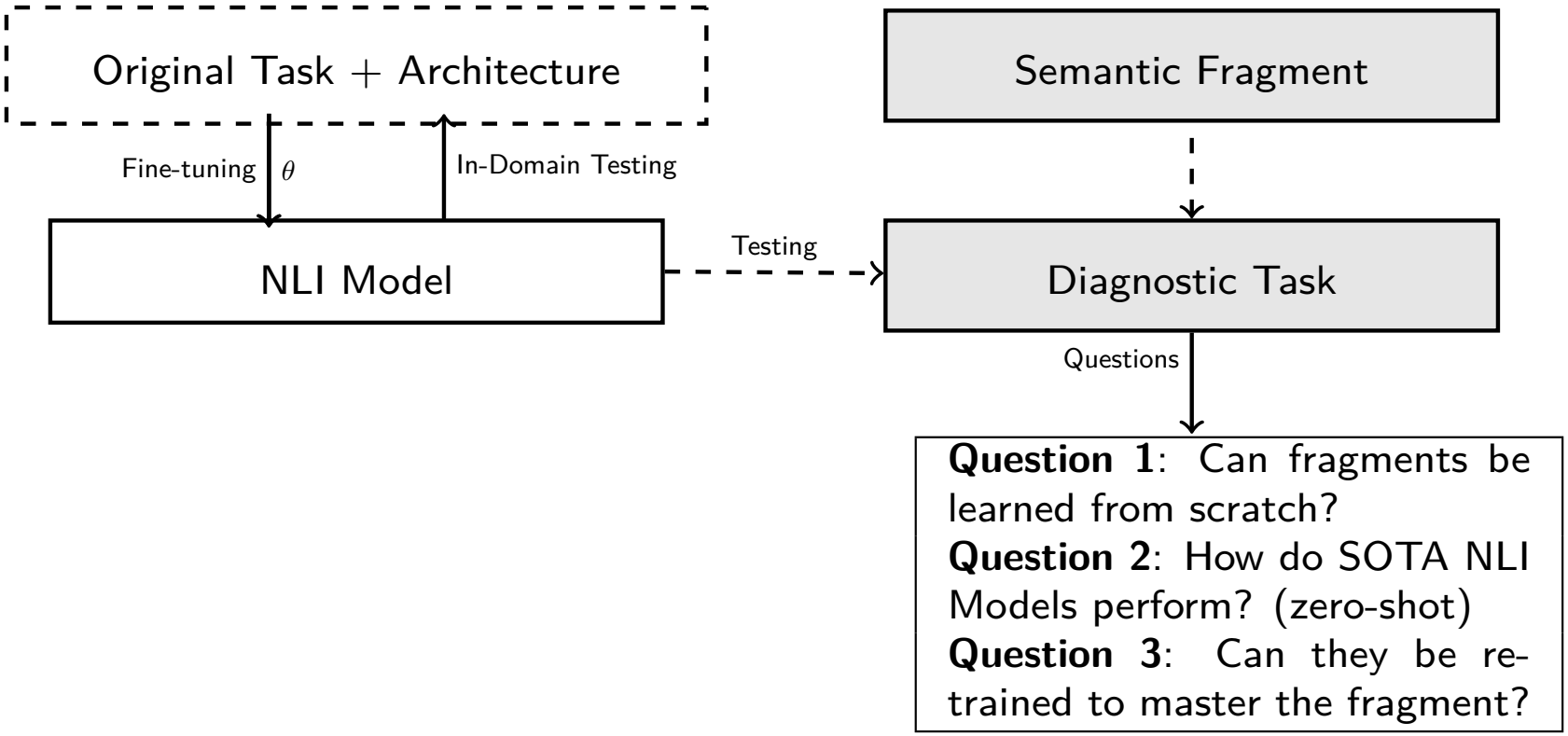
| Premise | Hypothesis | Label |
|---|---|---|
| Mitchell is as tall as Fred, Fred is as tall as Karl, Karl is as tall as Jon, Jon is as tall as Darryl, Darryl is as tall as Theodore, Theodore is as tall as Calvin, Calvin is as tall as Eddie , Eddie is as tall as Philip , Philip is taller than Travis | 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 |

  - **Out of Distribution Testing**: disjoint train and test vocabularies; investigate *lexical diversity* (Rozen et al. (2019)).

# Probing Methodology: Asking 3 Questions



# Probing Methodology: Asking 3 Questions




- Enforce certain **controls** (Hewitt and Liang (2019)); datasets should be demonstrably difficult, models should not forget.

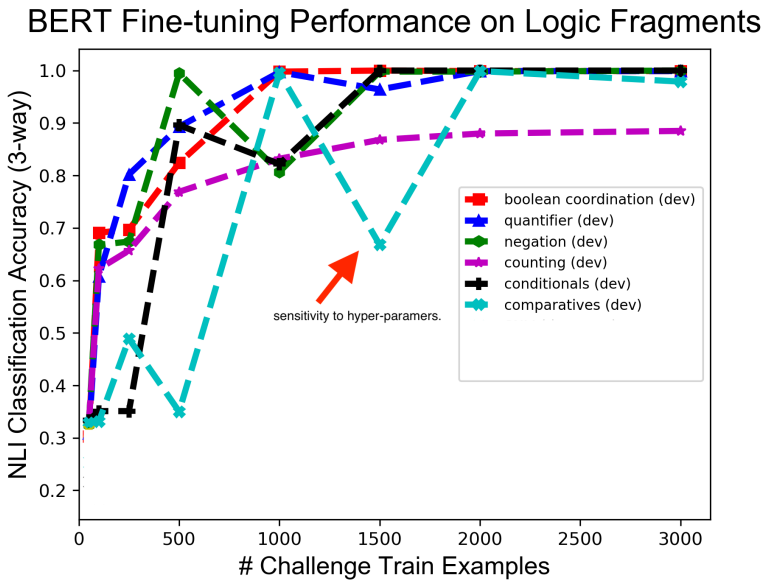
<Findings>

# Question 1: Can We Learn Fragments from Scratch?

- Training *task-specific* models without special NLI pre-training (i.e., the setup in Geiger et al. (2018)))

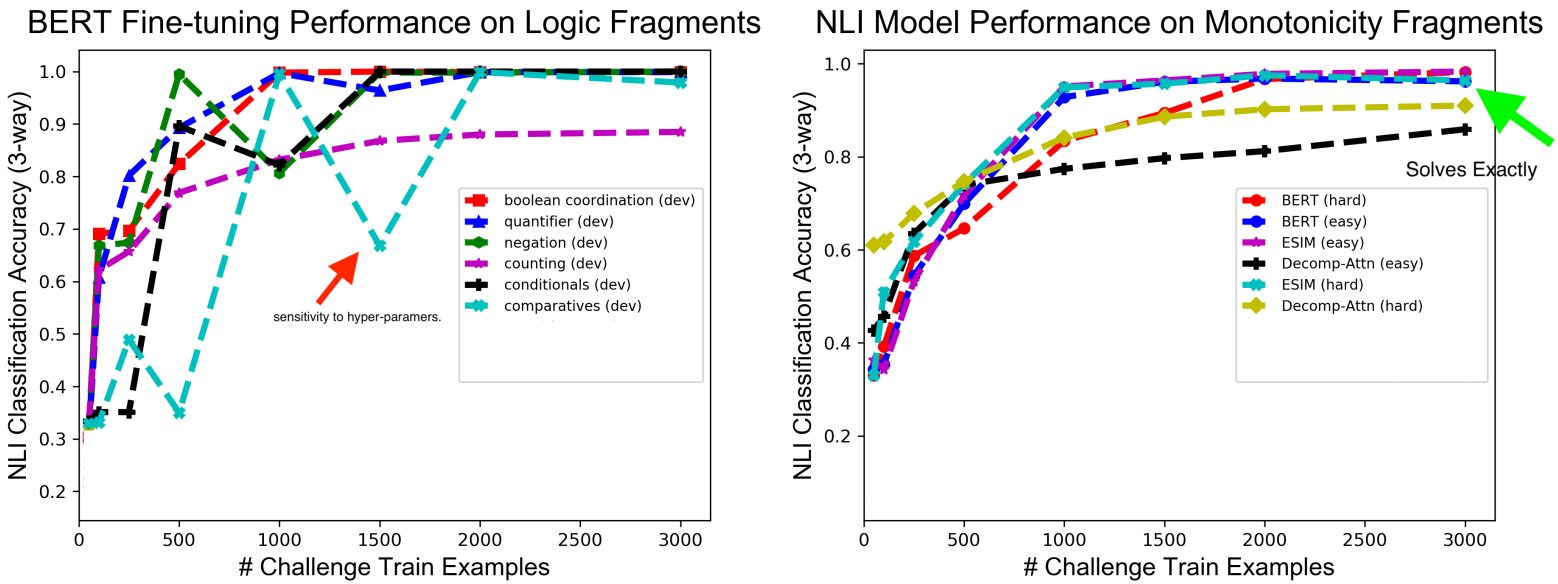
# Question 1: Can We Learn Fragments from Scratch?

- Training *task-specific* models without special NLI pre-training (i.e., the setup in Geiger et al. (2018)))



# Question 1: Can We Learn Fragments from Scratch?

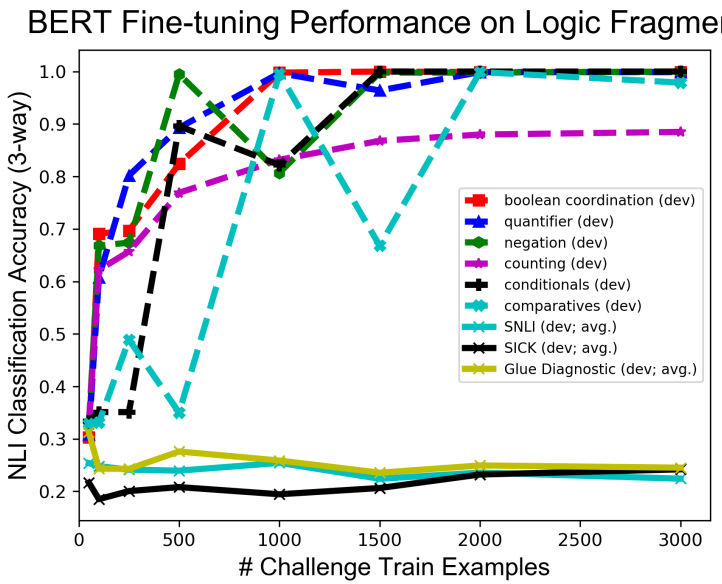
- Training *task-specific* models without special NLI pre-training (i.e., the setup in Geiger et al. (2018)))

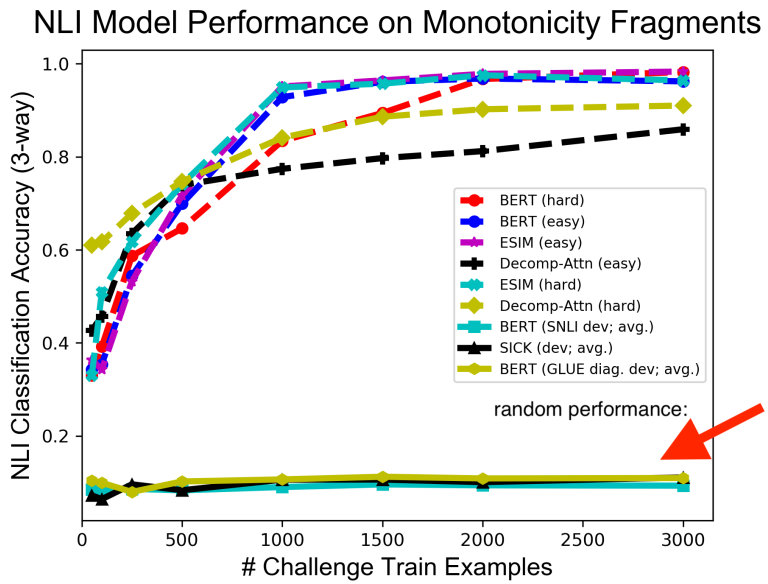



- BERT (+ ESIM, Decomposable-Attention) can easily learn most fragments. Difficult for other LSTM-based models/baselines.

# Question 1: Can We Learn Fragments from Scratch?

- Training *task-specific* models without special NLI pre-training (i.e., the setup in Geiger et al. (2018)))

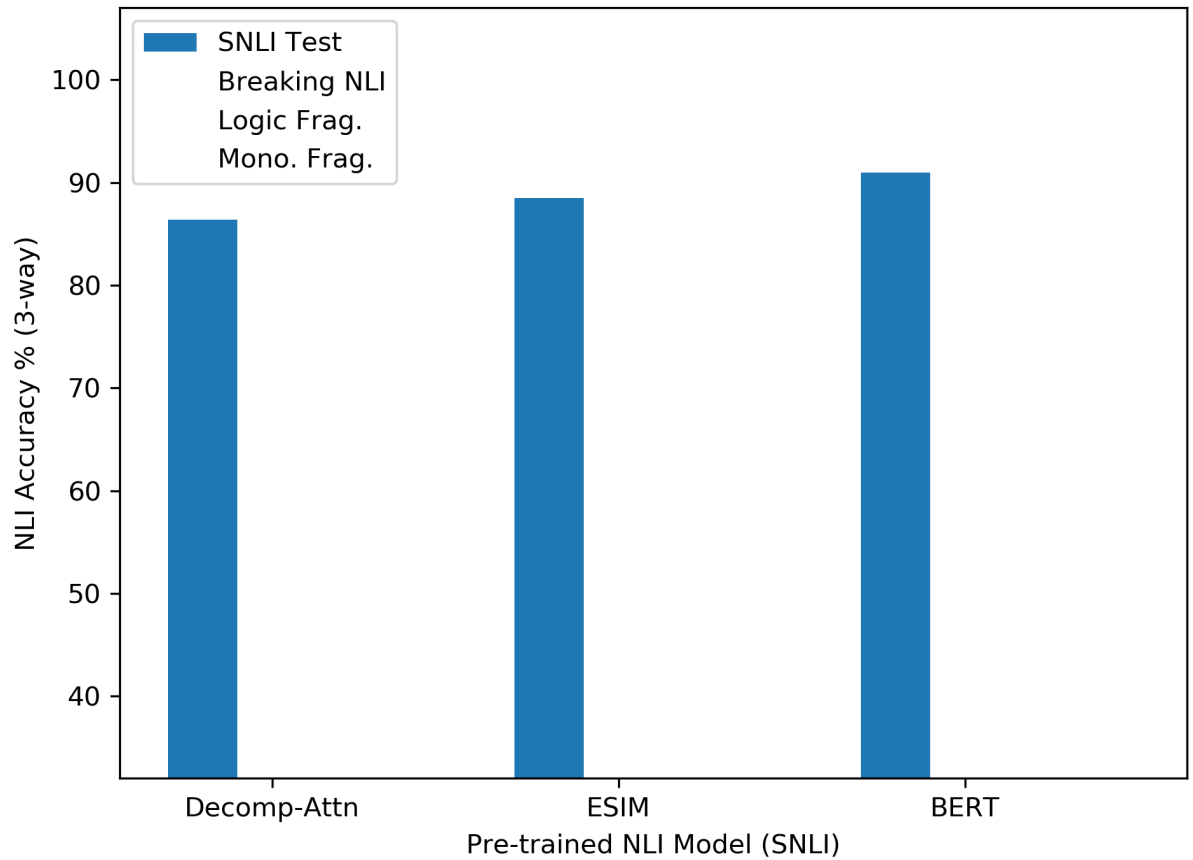



- **The Problem**: models are just **idiot savants**, cannot solve any other tasks (common probing strategy **but not always insightful**).

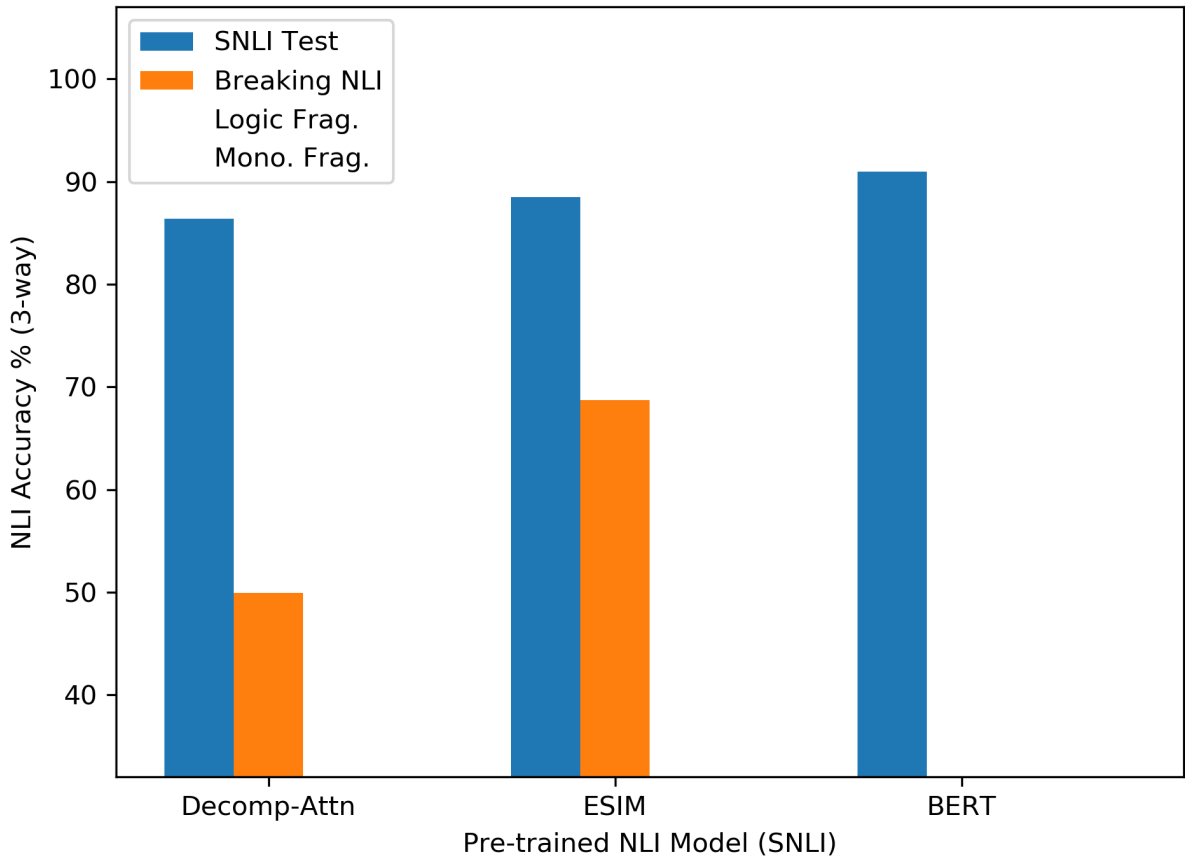
# Question 2: Zero-shot Evaluation

- How do models trained on NLI benchmarks perform? (i.e., the **BreakingNLI** (Glockner et al. (2018)) setup)



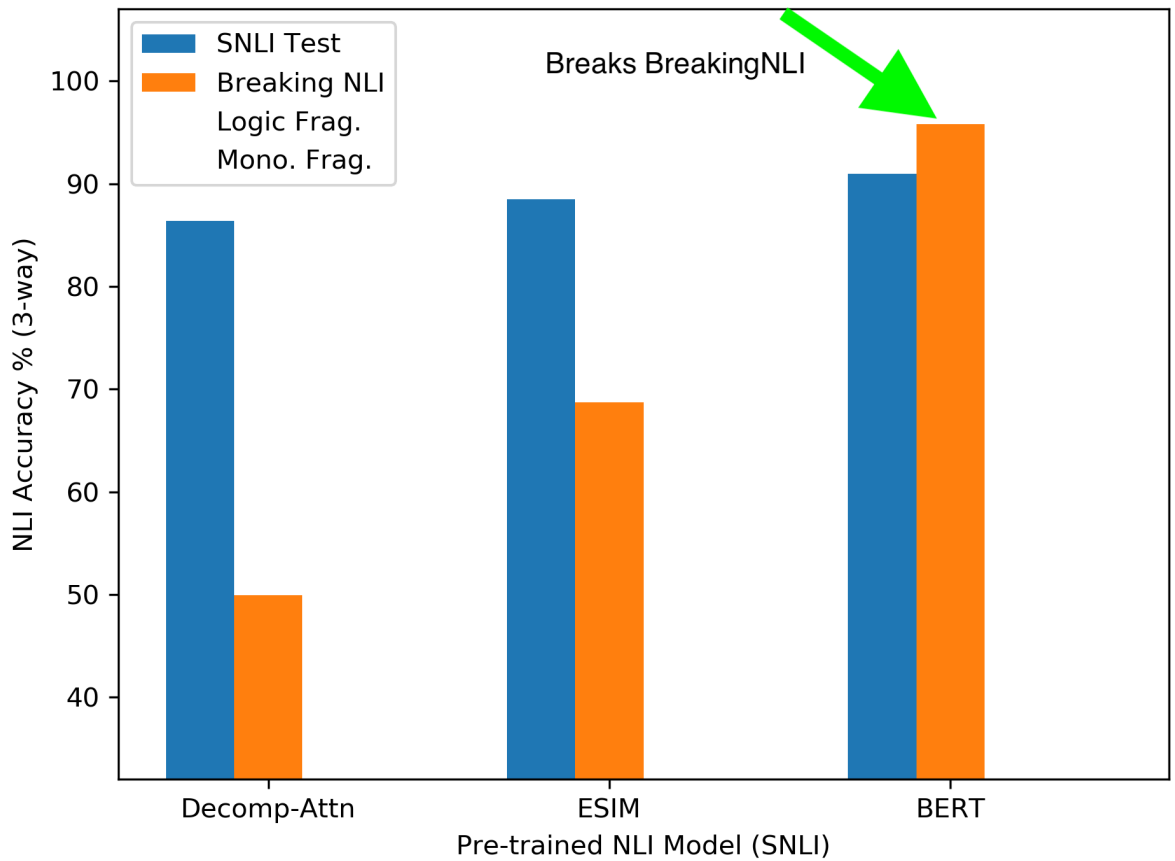
# Question 2: Zero-shot Evaluation

- How do models trained on NLI benchmarks perform? (i.e., the **BreakingNLI** (Glockner et al. (2018)) setup)



# Question 2: Zero-shot Evaluation

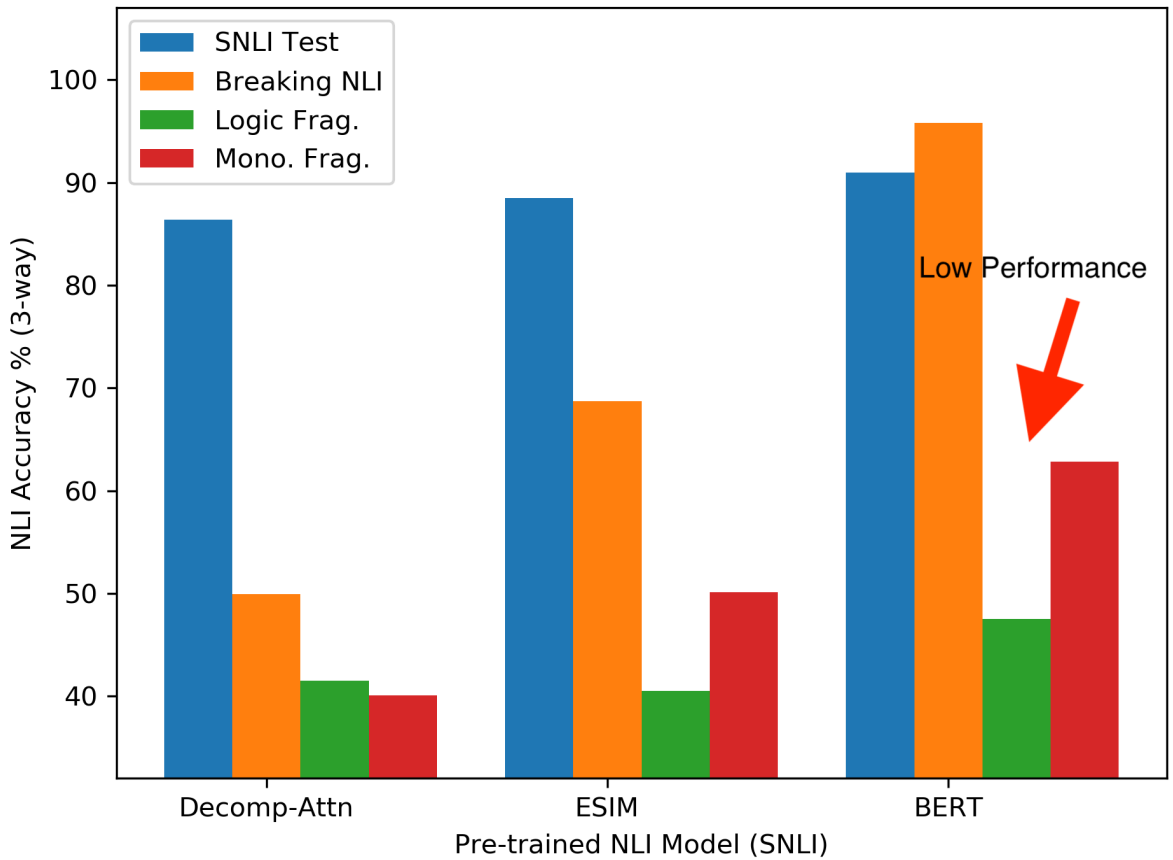
- How do models trained on NLI benchmarks perform? (i.e., the **BreakingNLI** (Glockner et al. (2018)) setup)




- Found BERT to solve **BreakingNLI**, showing need for more complex diagnostics.

# Question 2: Zero-shot Evaluation

- How do models trained on NLI benchmarks perform? (i.e., the **BreakingNLI** (Glockner et al. (2018)) setup)




- Pre-trained NLI **models perform poorly**, provides a new task that break models; but does this tell us much?

# The Biggest Challenge

Can we build models that are simultaneously good at our diagnostic tasks and their original benchmarks?

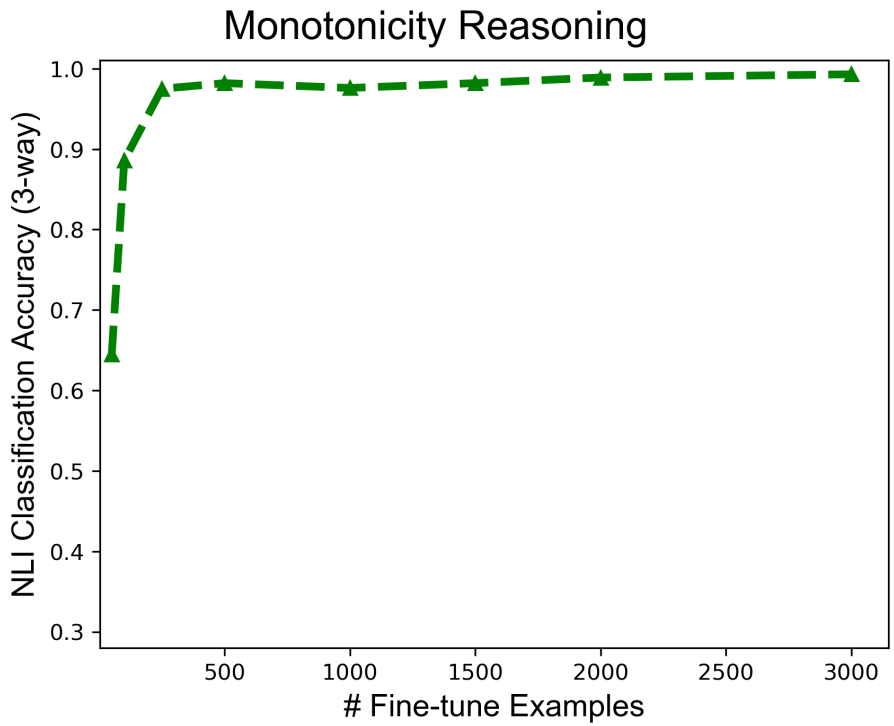
**Assumption**: A model's ability to quickly learn new tasks with limited *cost* (i.e., forgetting of original task) provides evidence of competence.

# Question 3: Can Models be Fixed? (Most interesting)

- **Model Inoculation** (Liu et al. (2019)): Continue training models on small amounts of diagnostic data; aim to (quickly/cheaply) fix model.

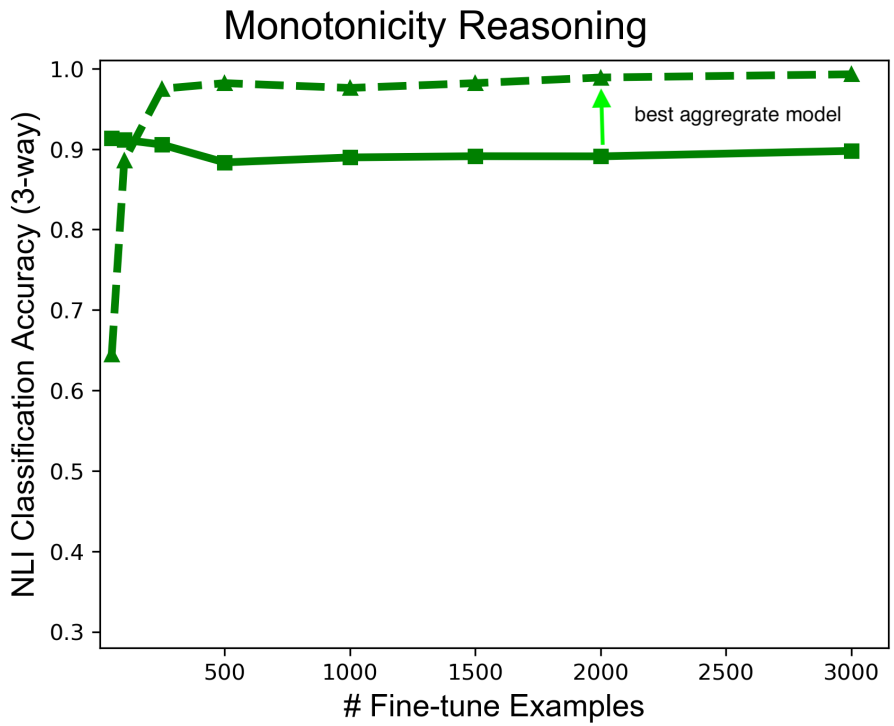
# Question 3: Can Models be Fixed? (Most interesting)

- **Model Inoculation** (Liu et al. (2019)): Continue training models on small amounts of diagnostic data; aim to (quickly/cheaply) fix model.



# Question 3: Can Models be Fixed? (Most interesting)

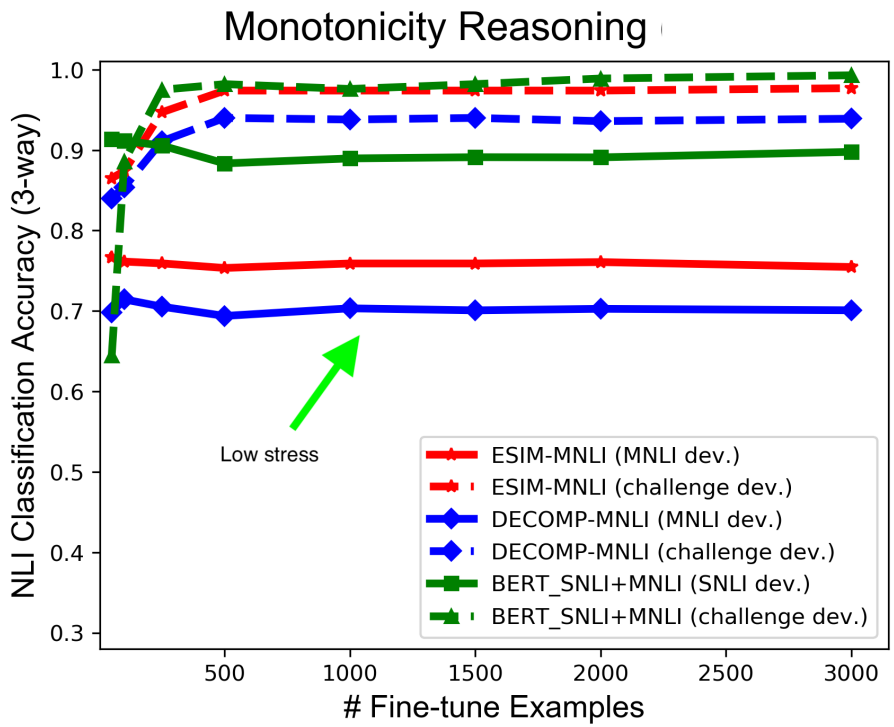
- **Model Inoculation** (Liu et al. (2019)): Continue training models on small amounts of diagnostic data; aim to (quickly/cheaply) fix model.




- **Loss-less Inoculation**: Models should be penalized for forgetting (a sign of stress), take best aggregate model.

# Question 3: Can Models be Fixed? (Most interesting)

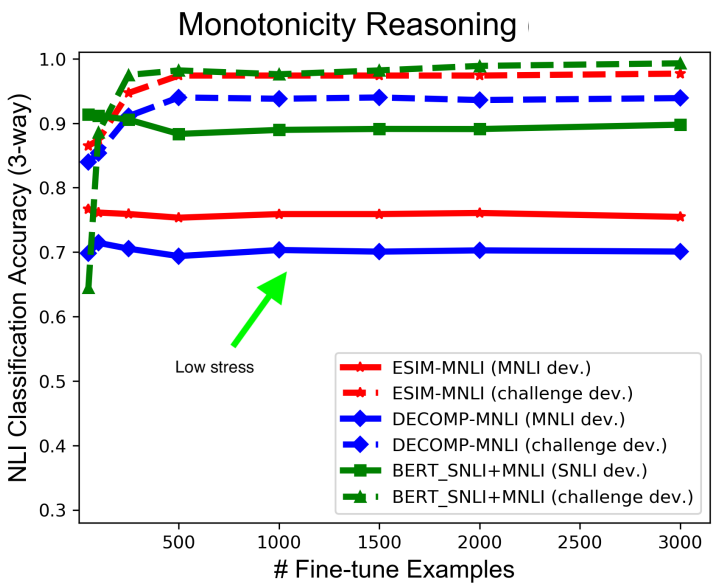
- **Model Inoculation** (Liu et al. (2019)): Continue training models on small amounts of diagnostic data; aim to (quickly/cheaply) fix model.

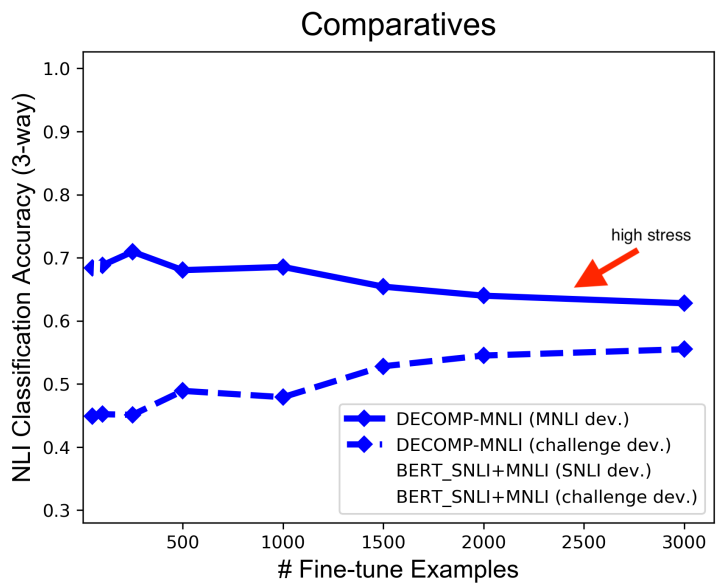



- Mastering diagnostic tasks with little loss gives evidence of competence and strong correspondence to training distribution.

# Question 3: Can Models be Fixed? (Most interesting)

- **Model Inoculation** (Liu et al. (2019)): Continue training models on small amounts of diagnostic data; aim to (quickly/cheaply) fix model.

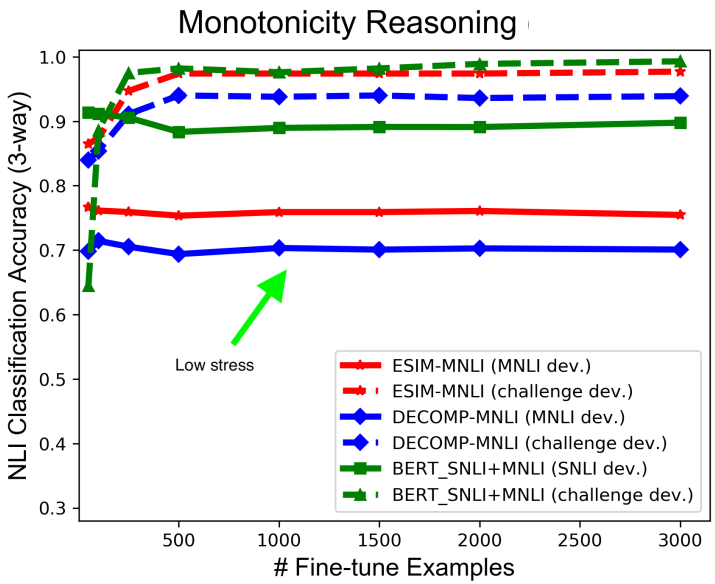


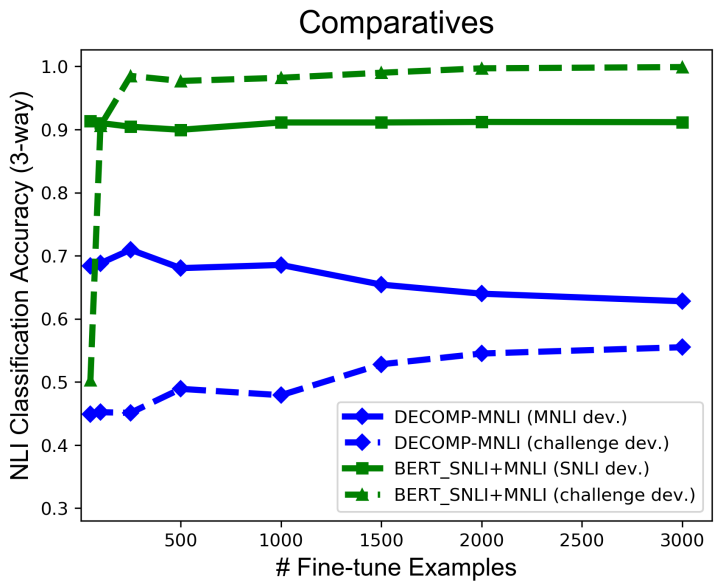




- **Not all fragments are the same**: some stress models (i.e., lead to forgetting) more than others; indicate lack of competence.

# Question 3: Can Models be Fixed? (Most interesting)

- **Model Inoculation** (Liu et al. (2019)): Continue training models on small amounts of diagnostic data; aim to (quickly/cheaply) fix model.

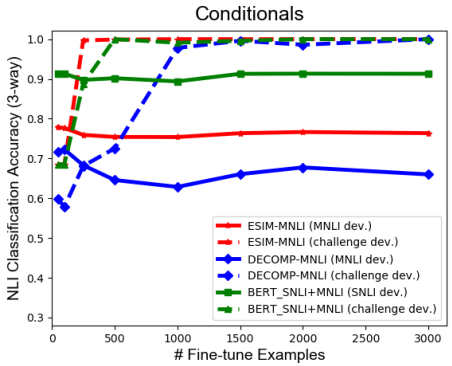


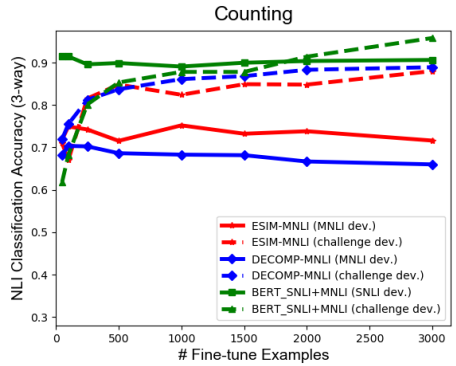

- **General finding**: more robust models (e.g., BERT) learn fast and with less forgetting; indication of higher competence.

# The Bigger Picture

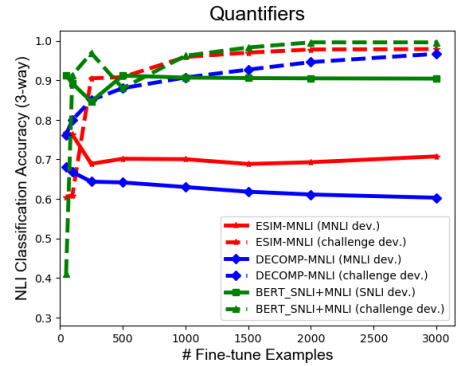
**BERT** ☺
**ESIM** ☺
**DecAttn** ☺



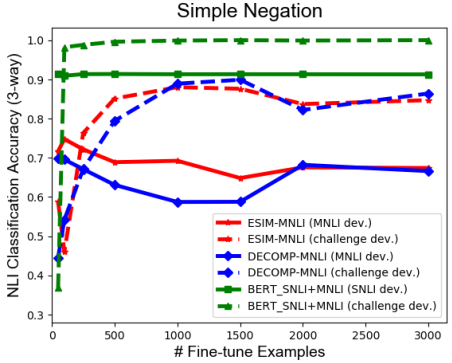

**BERT** ☺
**ESIM** ☺
**DecAttn** ☹



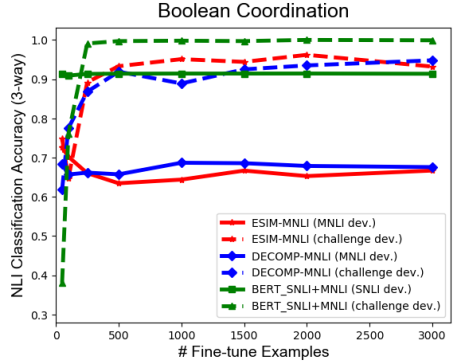

**BERT** ☺
**ESIM** ☹
**DecAttn** ☹



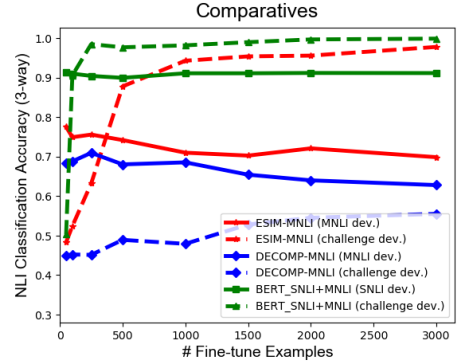

**BERT** ☺
**ESIM** ☹
**DecAttn** ☹




**BERT** ☺
**ESIM** ☹
**DecAttn** ☹




**BERT** ☺
**ESIM** ☹
**DecAttn** ☹




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

</Findings>

# Conclusions

- Proposed a linguistic approach to probing neural NLU models through **semantic fragments**.
  - Not a replacement for traditional dataset building, rather a supplement (**controlled experimentation**); a way to **engage linguists**.

# Conclusions

- Proposed a linguistic approach to probing neural NLU models through **semantic fragments**.
  - Not a replacement for traditional dataset building, rather a supplement (**controlled experimentation**); a way to **engage linguists**.
- Shows results on 8 diagnostic tasks, BERT models in particular show signs of high competence and high capacity for learning new phenomena.

# Conclusions

- Proposed a linguistic approach to probing neural NLU models through **semantic fragments**.
  - Not a replacement for traditional dataset building, rather a supplement (**controlled experimentation**); a way to **engage linguists**.
- Shows results on 8 diagnostic tasks, BERT models in particular show signs of high competence and high capacity for learning new phenomena.
- **Future**: Probing is difficult, need new semantic fragments.

Thanks

# References I


Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of EMNLP*.

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., Cases, I., Karttunen, L., and Potts, C. (2018). Stress-Testing Neural Models of Natural Language Inference with Multiply-Quantified Sentences. *arXiv preprint arXiv:1810.13033*.

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*.

Hewitt, J. and Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. *arXiv preprint arXiv:1909.03368*.

Hu, H. and Moss, L. S. (2018). Polarity computations in flexible categorial grammar. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 124–129.

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*.

# References II


McCoy, R. T., Pavlick, E., and Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *arXiv preprint arXiv:1902.01007*.

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

Poliak, A., Haldar, A., Rudinger, R., Hu, J. E., Pavlick, E., White, A. S., and Van Durme, B. (2018). Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81.

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

Richardson, K. and Sabharwal, A. (2019). What Does My QA Model Know? Devising Controlled Probes using Expert Knowledge. *arXiv preprint arXiv:1912.13337*.

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*.

Salvatore, F., Finger, M., and Hirata Jr, R. (2019). Using Syntactical and Logical Forms to Evaluate Textual Inference Competence. *arXiv preprint arXiv:1905.05704*.

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

# References III


Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A Multi-task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv preprint arXiv:1804.07461*.

Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., Alsop, A., Bordia, S., Liu, H., Parrish, A., et al. (2019). Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. *arXiv preprint arXiv:1909.02597*.

White, A. S., Rastogi, P., Duh, K., and Van Durme, B. (2017). Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005.