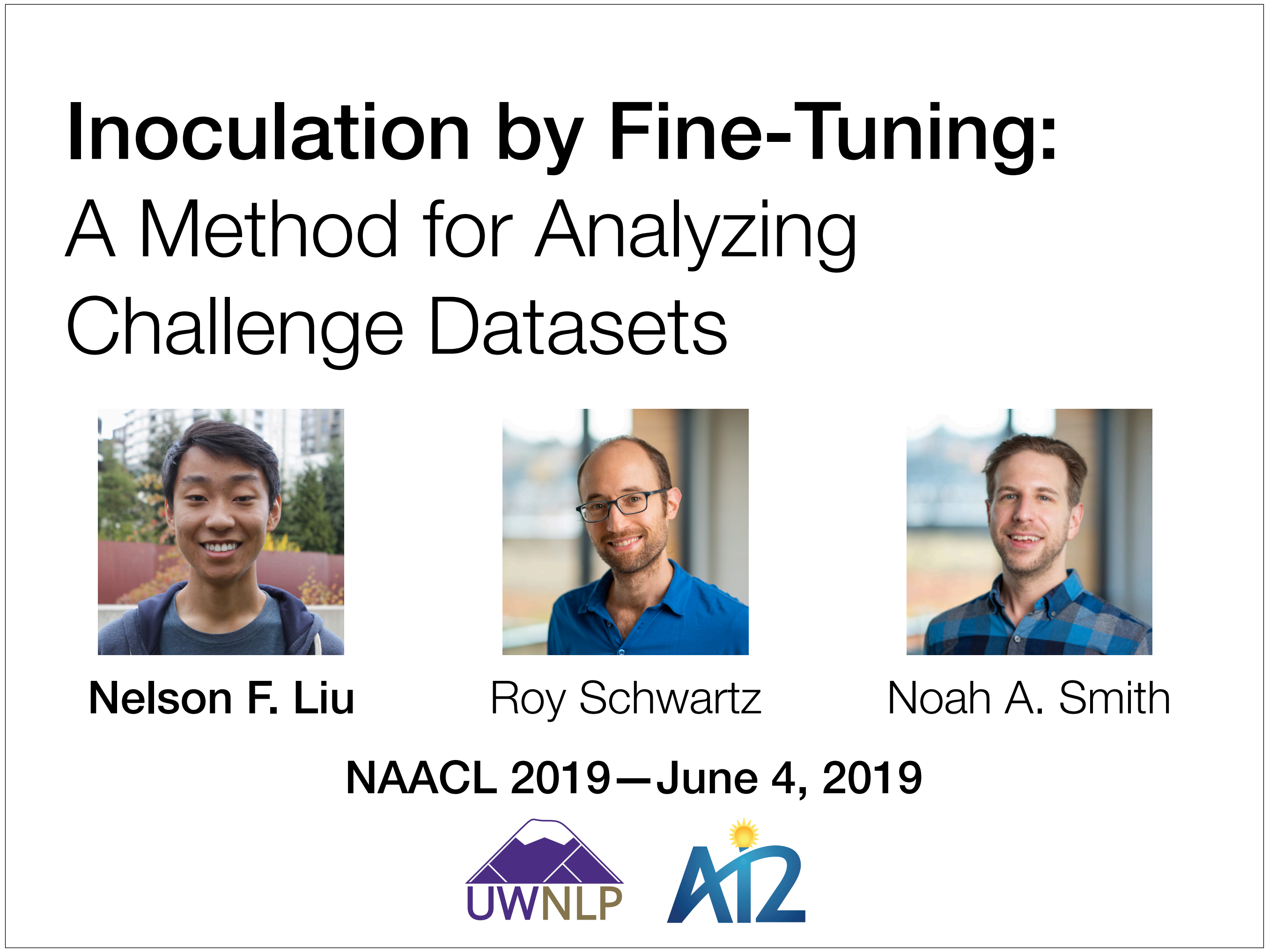


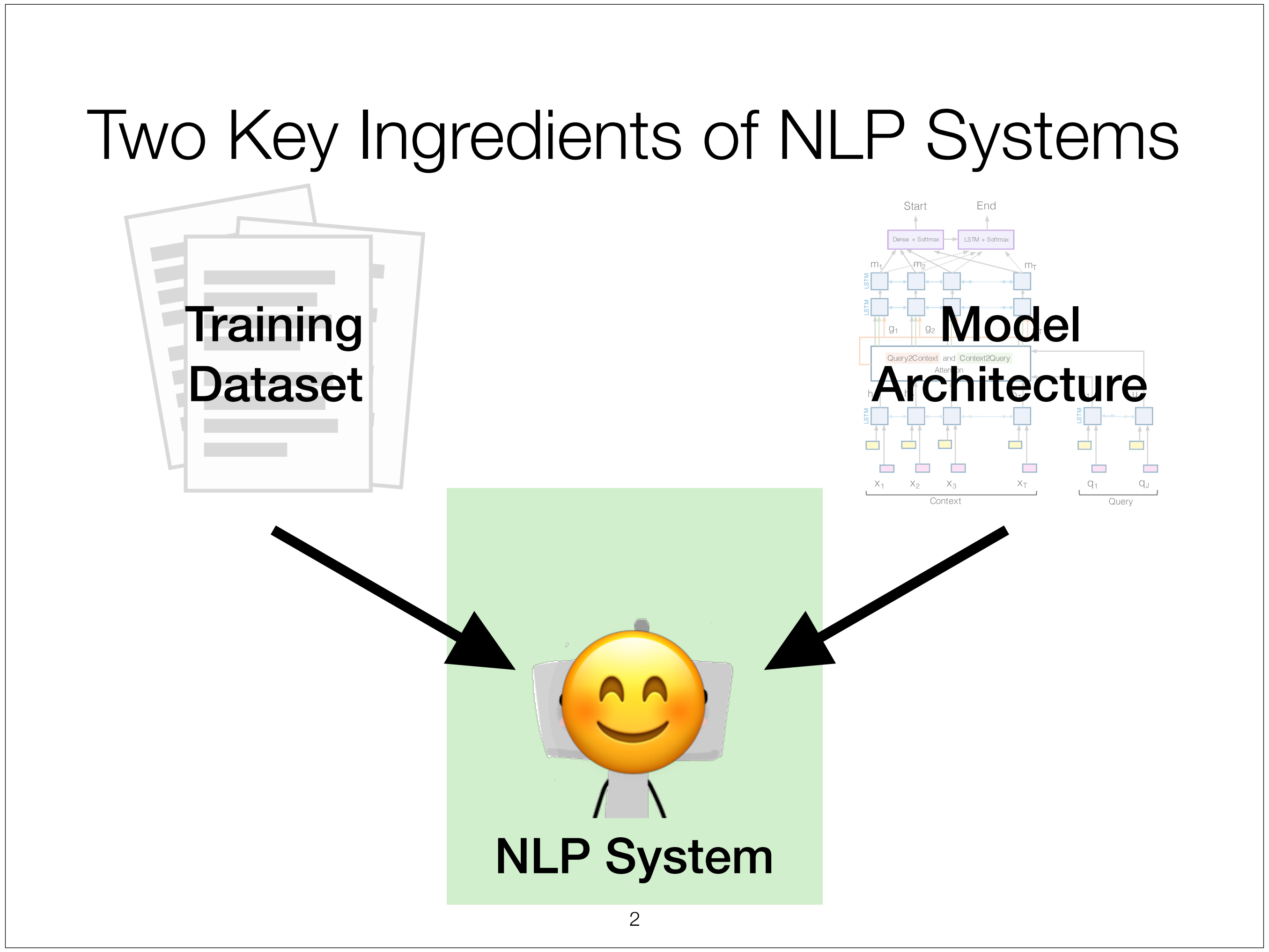
Inoculation by Fine-Tuning:
A Method for Analyzing Challenge Datasets

Nelson F. Liu Roy Schwartz Noah A. Smith

NAACL 2019—June 4, 2019

Thanks. Today, I'll be presenting "Inoculation by Fine-Tuning", our method for characterizing the lack of robustness in neural NLP models.

This is joint work with Roy Schwartz and Noah Smith.



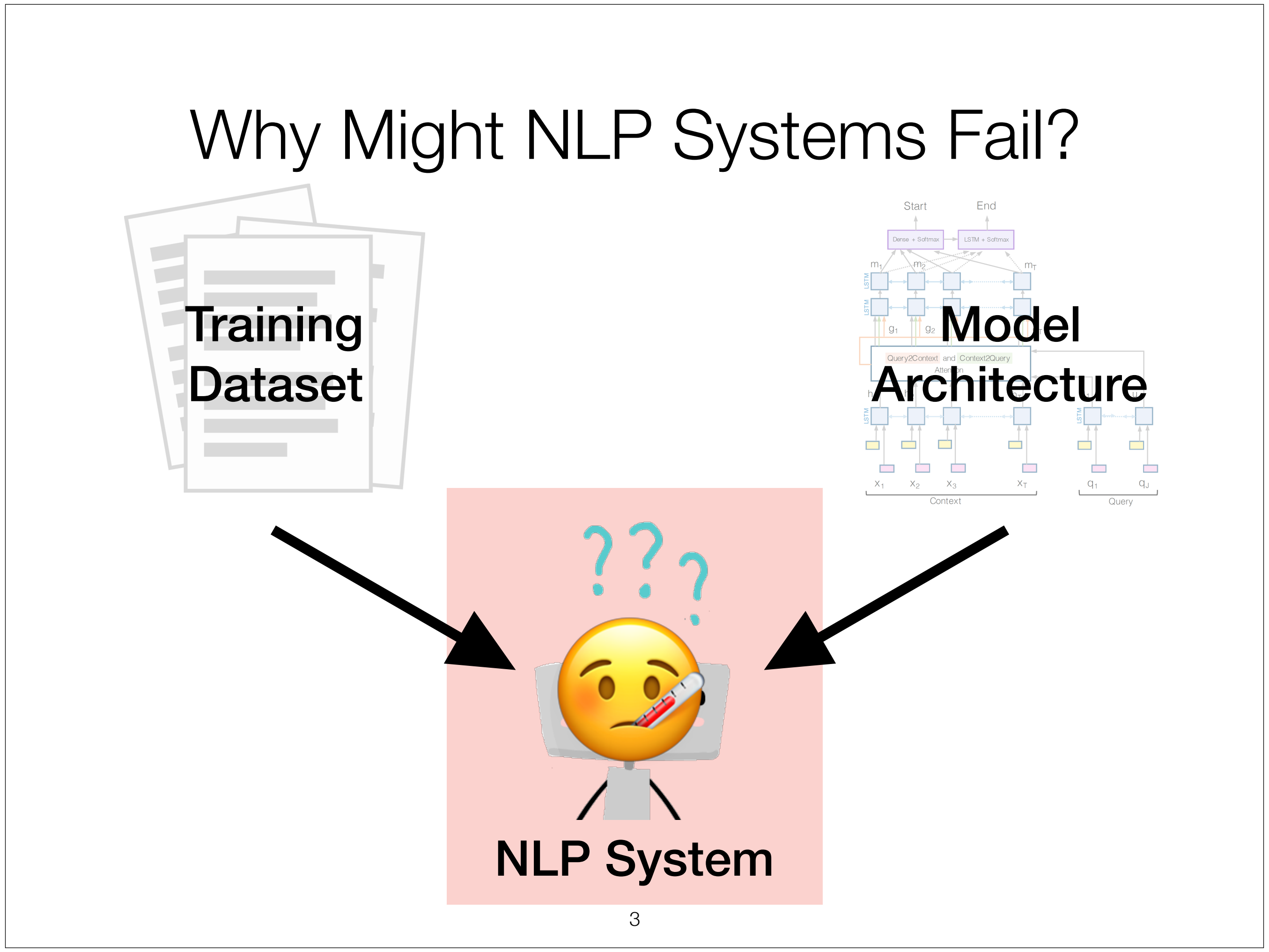
Two Key Ingredients of NLP Systems

Training Dataset

Model Architecture

NLP System

When we build NLP systems, we often require two key components---a training dataset, and a model architecture.



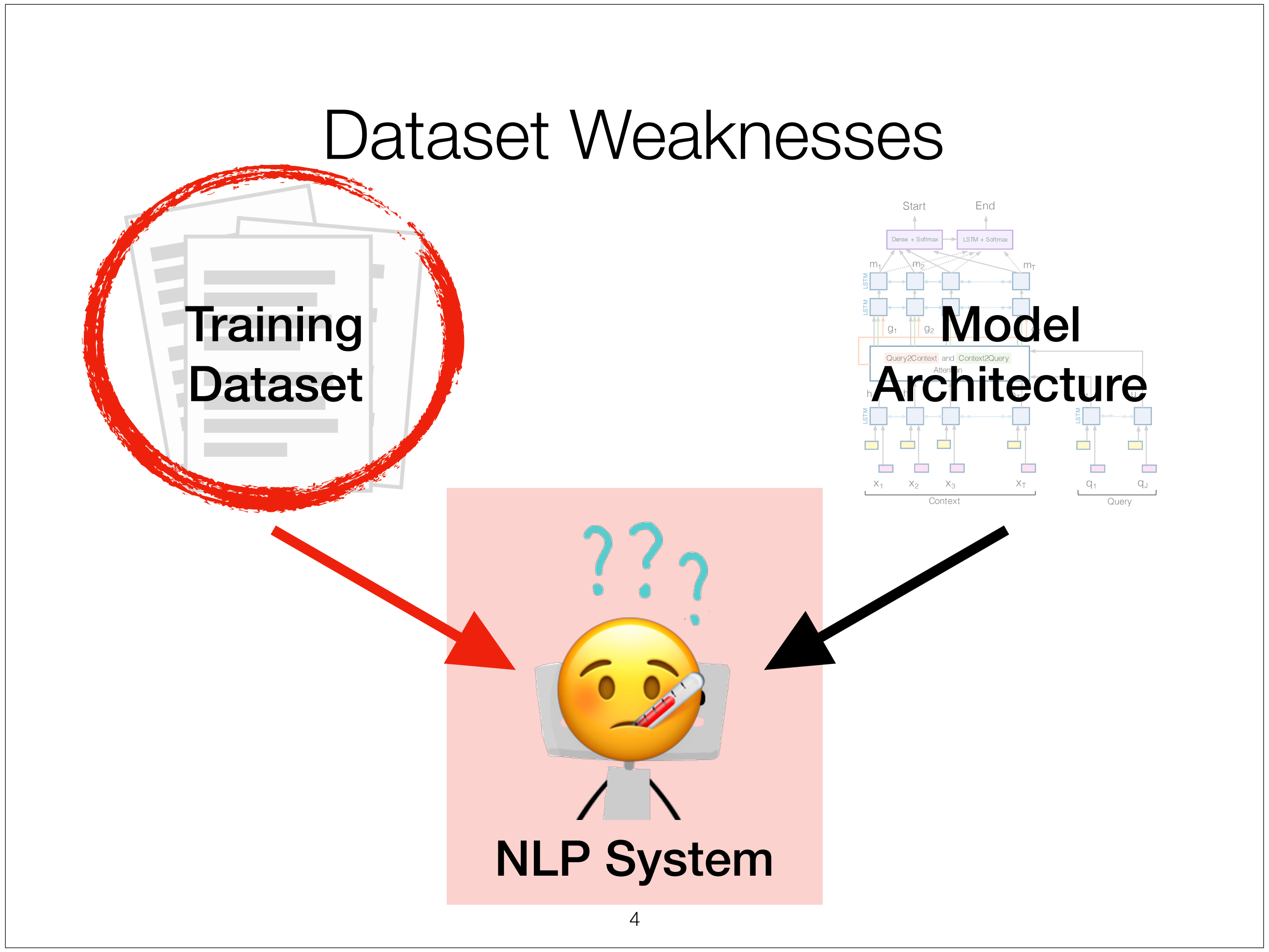
Why Might NLP Systems Fail?

Training Dataset

Model Architecture

NLP System

Given these two components, when we use or evaluate our NLP systems, they might fail for a variety of reasons.



Dataset Weaknesses

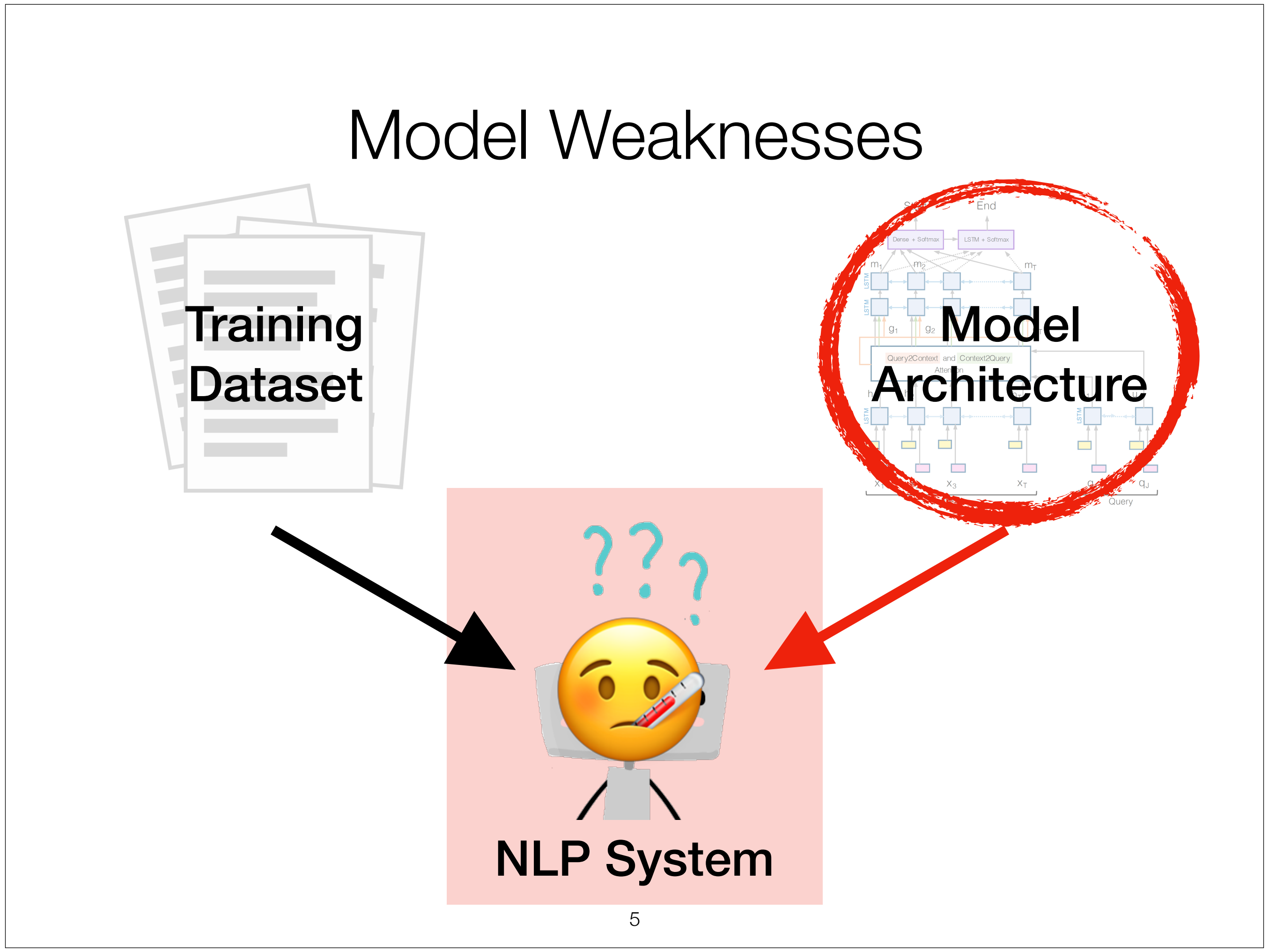
Training Dataset

Model Architecture

NLP System

For example, the test data might be so far removed from the training data that the model has no idea what to do.

Broadly, the test data might exploit blind spots in the training dataset, a deficiency that we call a "dataset weakness".



Model Weaknesses

Training Dataset

Model Architecture

NLP System

Alternatively, the test dataset might expose an inherent inability of a particular model family to handle certain natural language phenomena.

For instance, a bag-of-words model throws away the word order, and this can lead to incorrect predictions.

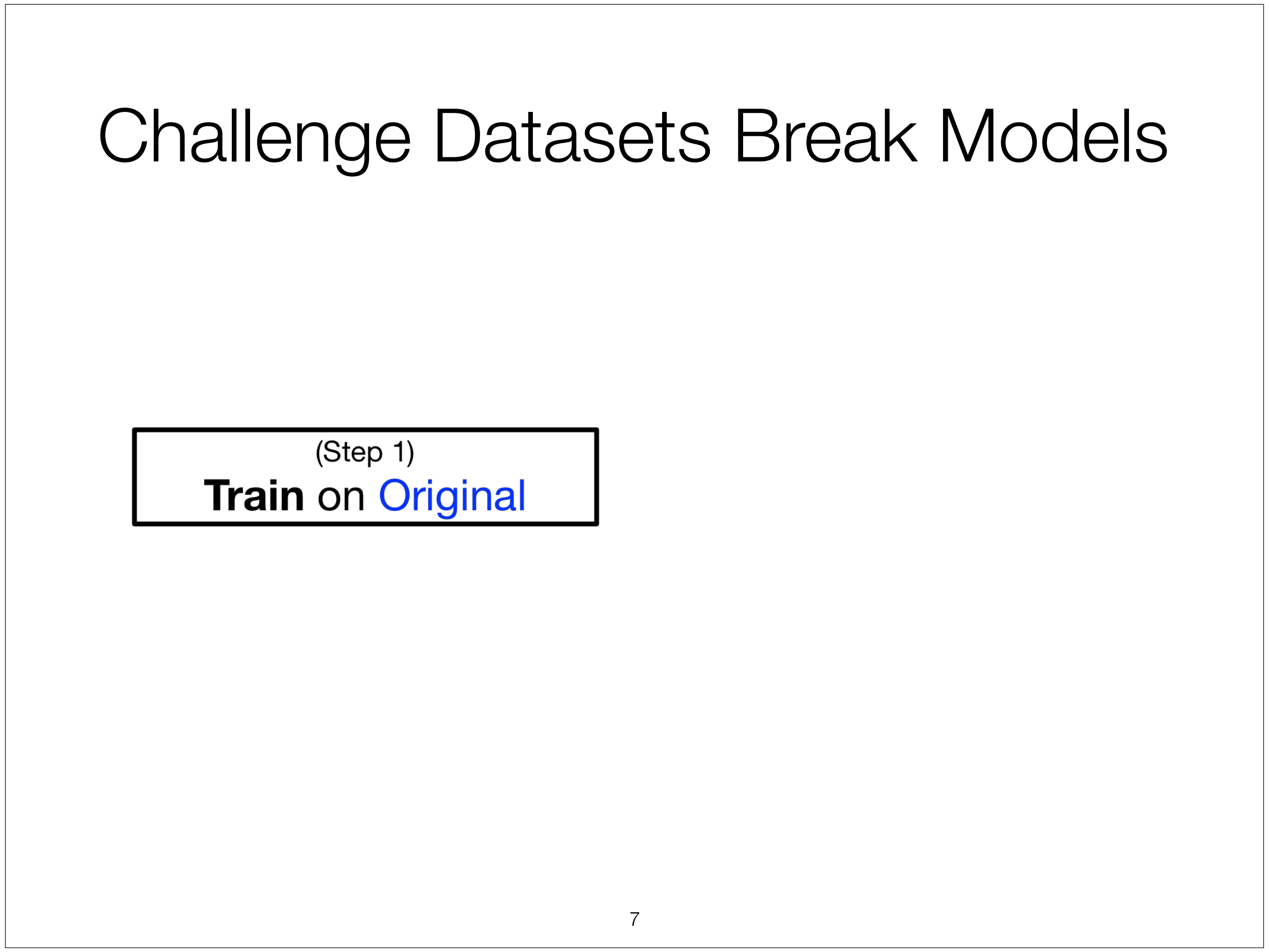
We call failures arising from these limitations, "model weaknesses".

This is certainly not a comprehensive list, and these cases aren't mutually exclusive either.
I just wanted to start by making the point that we should be cognizant that models fail for a variety of reasons.

Challenge Datasets Break Models

Indeed, we've gotten very good at making our models fail, and there's been a whole line of work on challenge datasets---simple perturbations to input data that break our models.

The typical challenge evaluation procedure looks something like this:

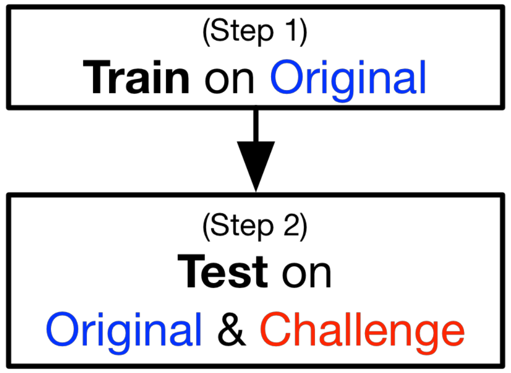


Challenge Datasets Break Models

(Step 1)
Train on Original

First, a model is trained on some original dataset.

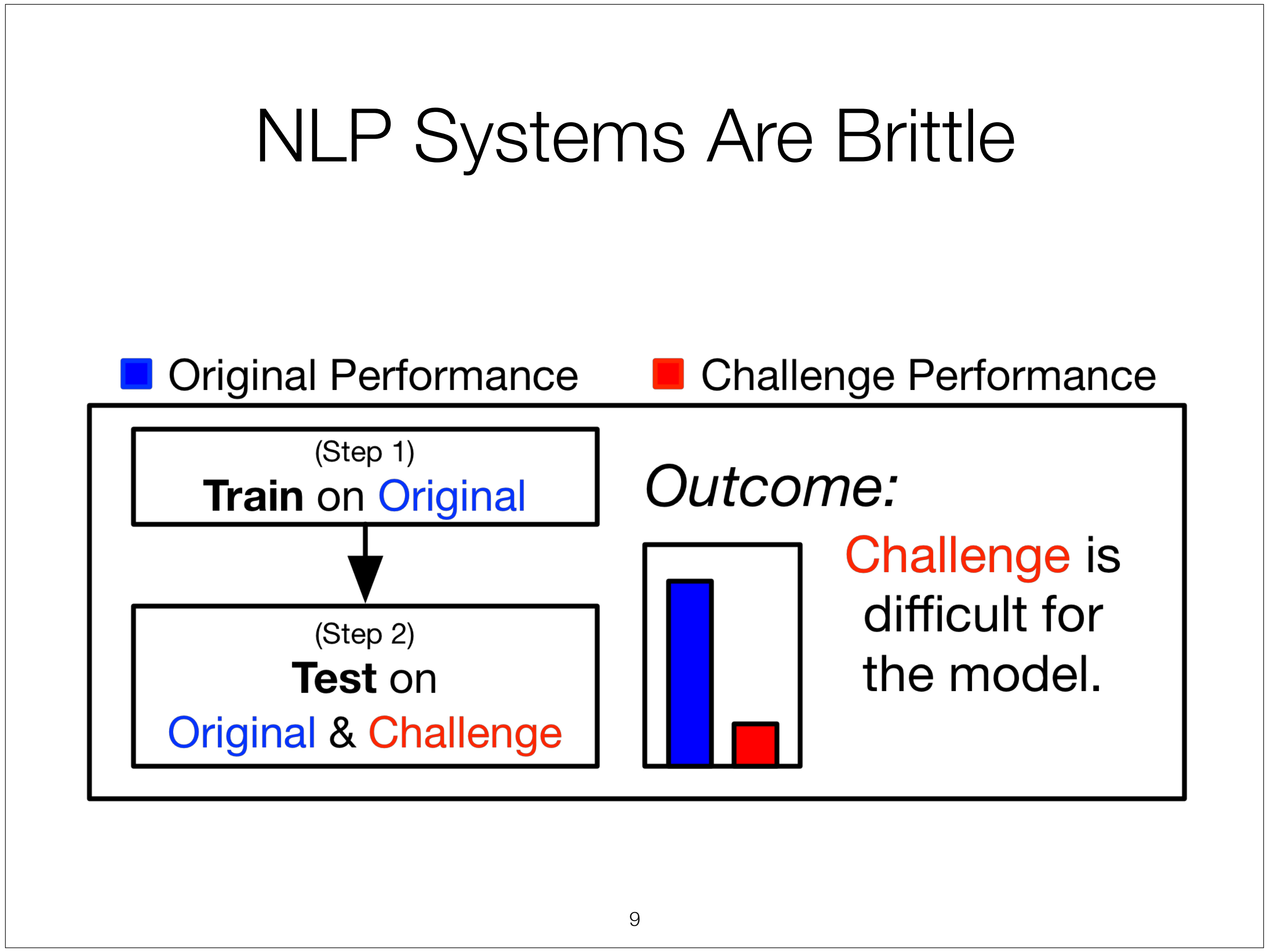
Challenge Datasets Break Models



(Step 1)
Train on Original

(Step 2)
Test on
Original & Challenge

Then, we test the model on both the original and the challenge dataset.



NLP Systems Are Brittle

🟦 Original Performance 🟥 Challenge Performance

(Step 1)
Train on Original

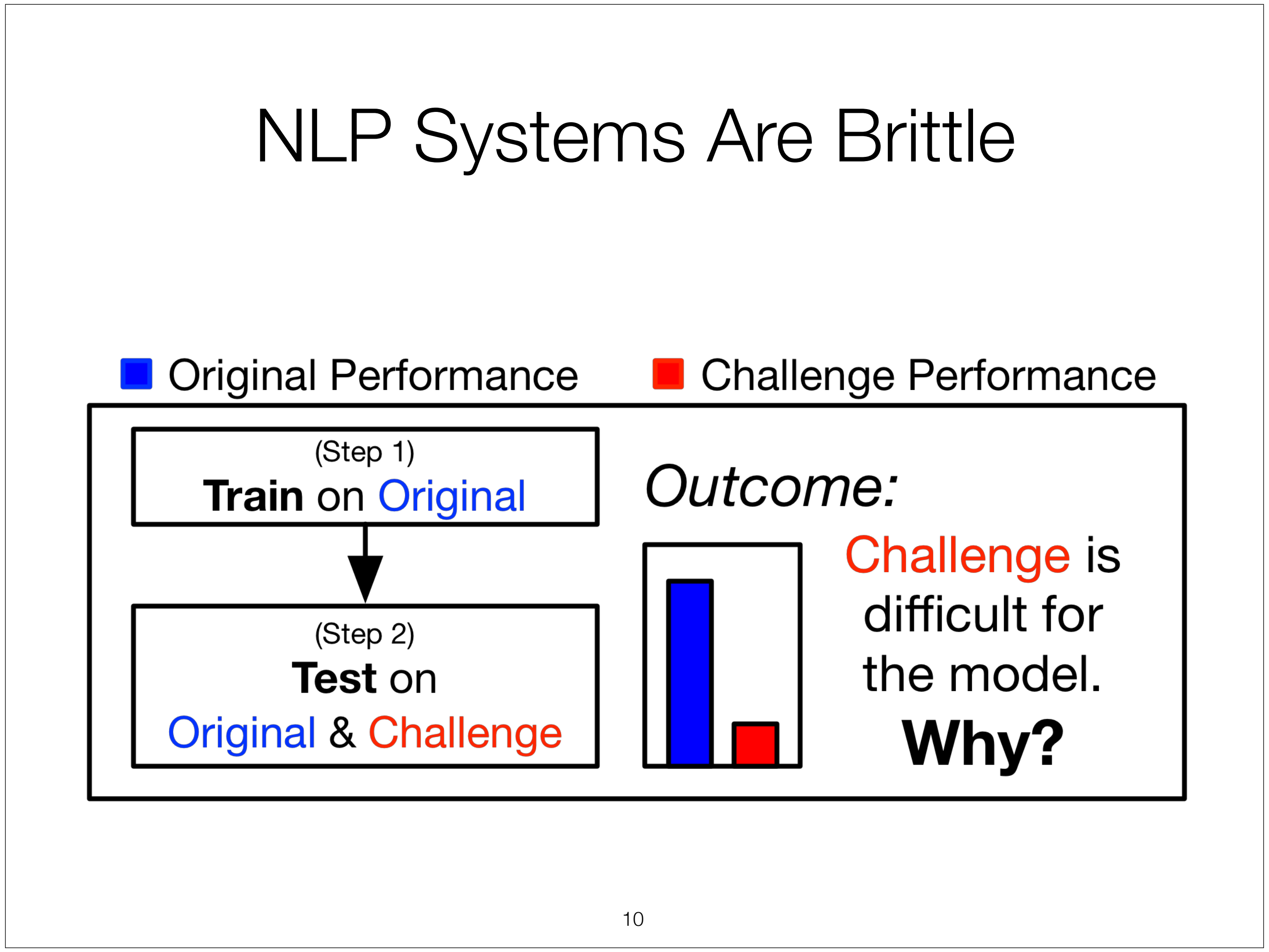
(Step 2)
Test on
Original & Challenge

Outcome:

Challenge is difficult for the model.

The result is that the performance on the original dataset, in blue, is far higher than performance on the challenge dataset, in red.

These challenges have been used as evidence that current systems are brittle, since systems that achieve state-of-the-art performance on standard benchmarks fail to generalize to even simple perturbations to their input.



NLP Systems Are Brittle

🟦 Original Performance 🟥 Challenge Performance

(Step 1)
Train on Original

(Step 2)
Test on
Original & Challenge

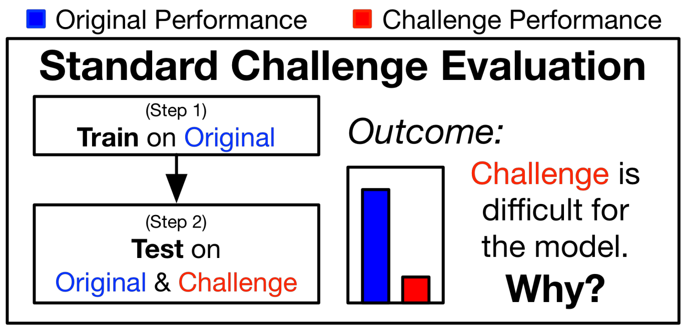
Outcome:

Challenge is difficult for the model.

Why?

Challenge datasets break our models, but it's difficult to glean insights beyond this---we don't know what particular weaknesses the challenge dataset reveals.

Inoculation by Fine-Tuning



🟦 Original Performance 🟥 Challenge Performance

Standard Challenge Evaluation

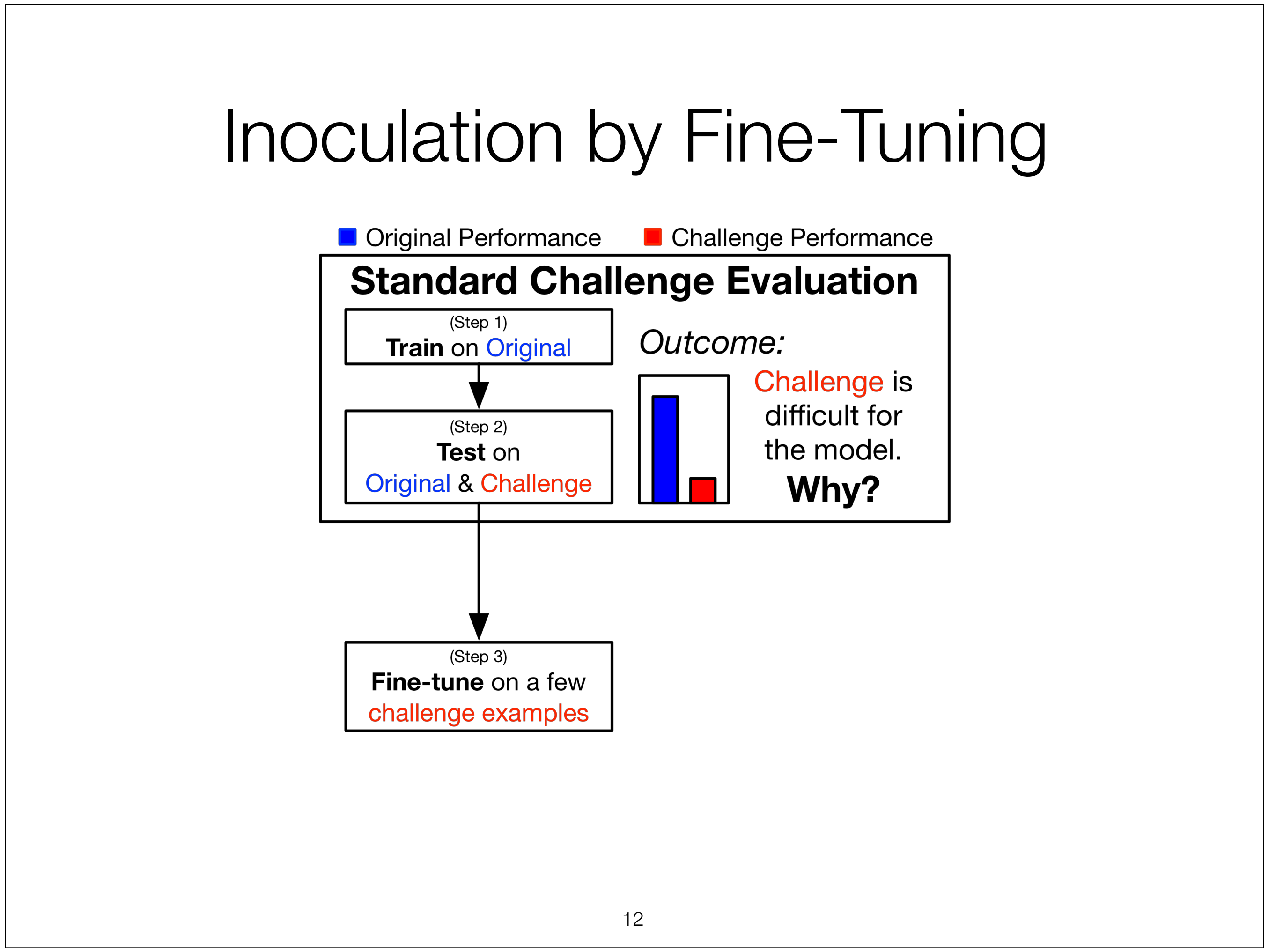
(Step 1)
Train on Original

(Step 2)
Test on
Original & Challenge

Outcome:

Challenge is difficult for the model. **Why?**

We introduce "Inoculation by Fine-tuning", a method that seeks to better understand *why* challenge datasets are difficult for particular models.



Inoculation by Fine-Tuning

🟦 Original Performance 🟥 Challenge Performance

Standard Challenge Evaluation

(Step 1)
Train on Original

(Step 2)
Test on
Original & Challenge

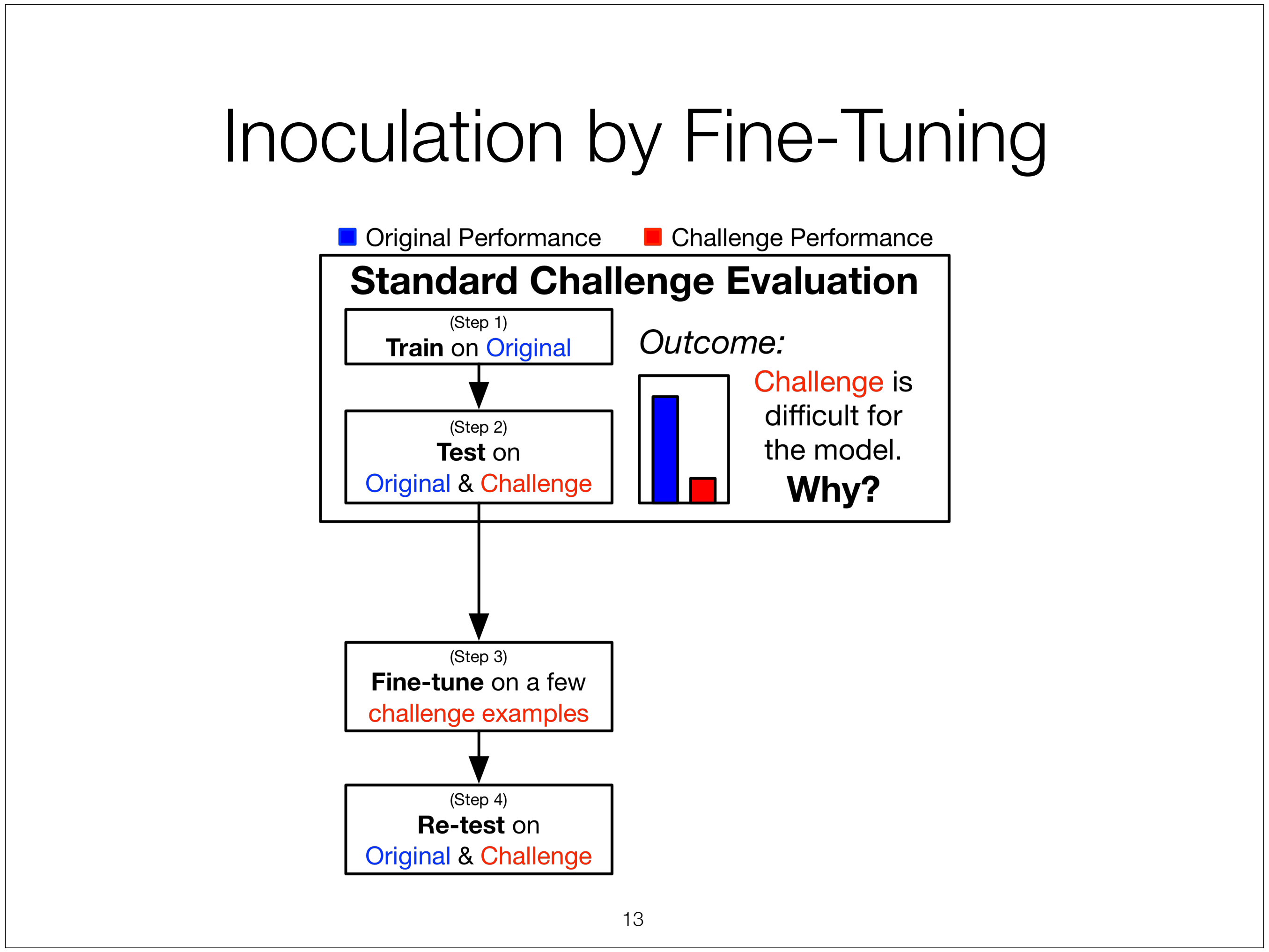
Outcome:

Challenge is difficult for the model.
Why?

(Step 3)
Fine-tune on a few
challenge examples

At a high level, the method exposes models trained on some original dataset (our metaphorical patient) to a small amount of data from the challenge dataset (our metaphorical pathogen), allowing learning to continue.

I should note that, while "fine-tuning" has been a popular way to *improve* models, in this case, we use it to better *understand* **models**, their **training** datasets, and **challenge** datasets.



Inoculation by Fine-Tuning

🟦 Original Performance 🟥 Challenge Performance

Standard Challenge Evaluation

(Step 1)
Train on Original

(Step 2)
Test on
Original & Challenge

Outcome:

Challenge is difficult for the model.
Why?

(Step 3)
Fine-tune on a few
challenge examples

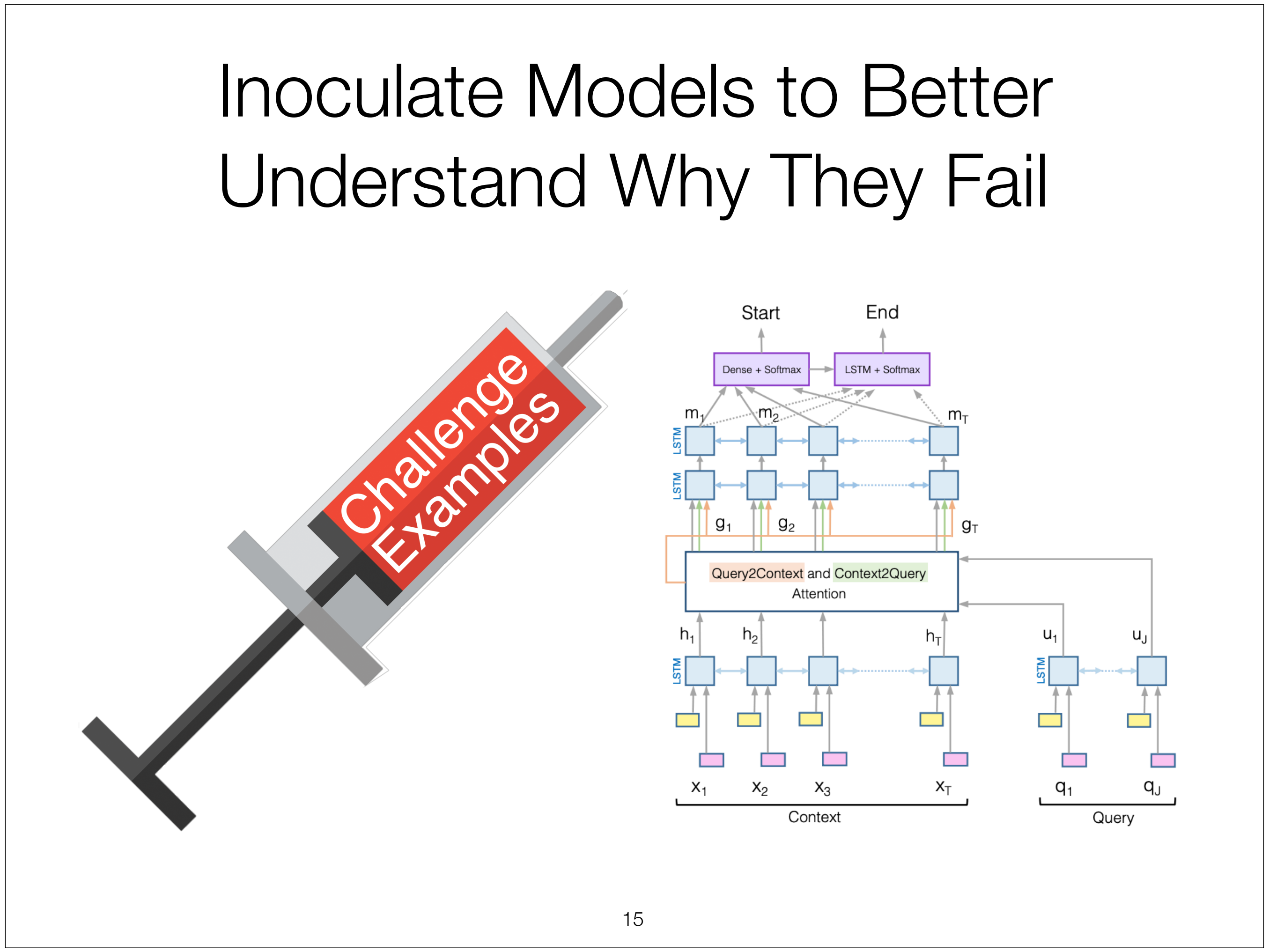
(Step 4)
Re-test on
Original & Challenge

By seeing how performance changes on the original and challenge datasets, we can get a better sense of how they stress models.

Inoculation

So, while I'm happy to argue for children being inoculated...

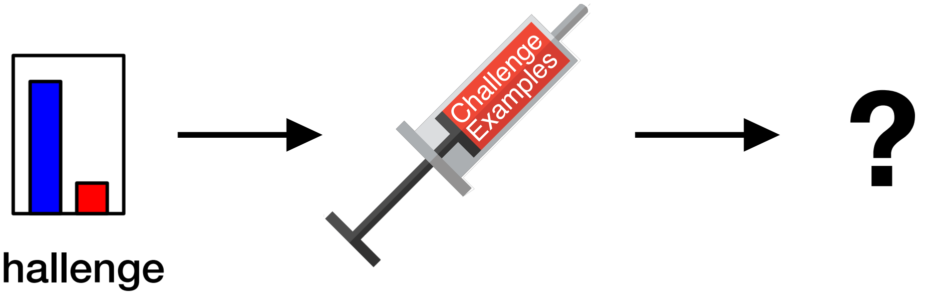
Inoculate Models to Better Understand Why They Fail



...this talk is about inoculating *models* to better understand why they fail. The model is our patient, and the challenge dataset examples are the pathogens.

Three Clear Outcomes of Interest

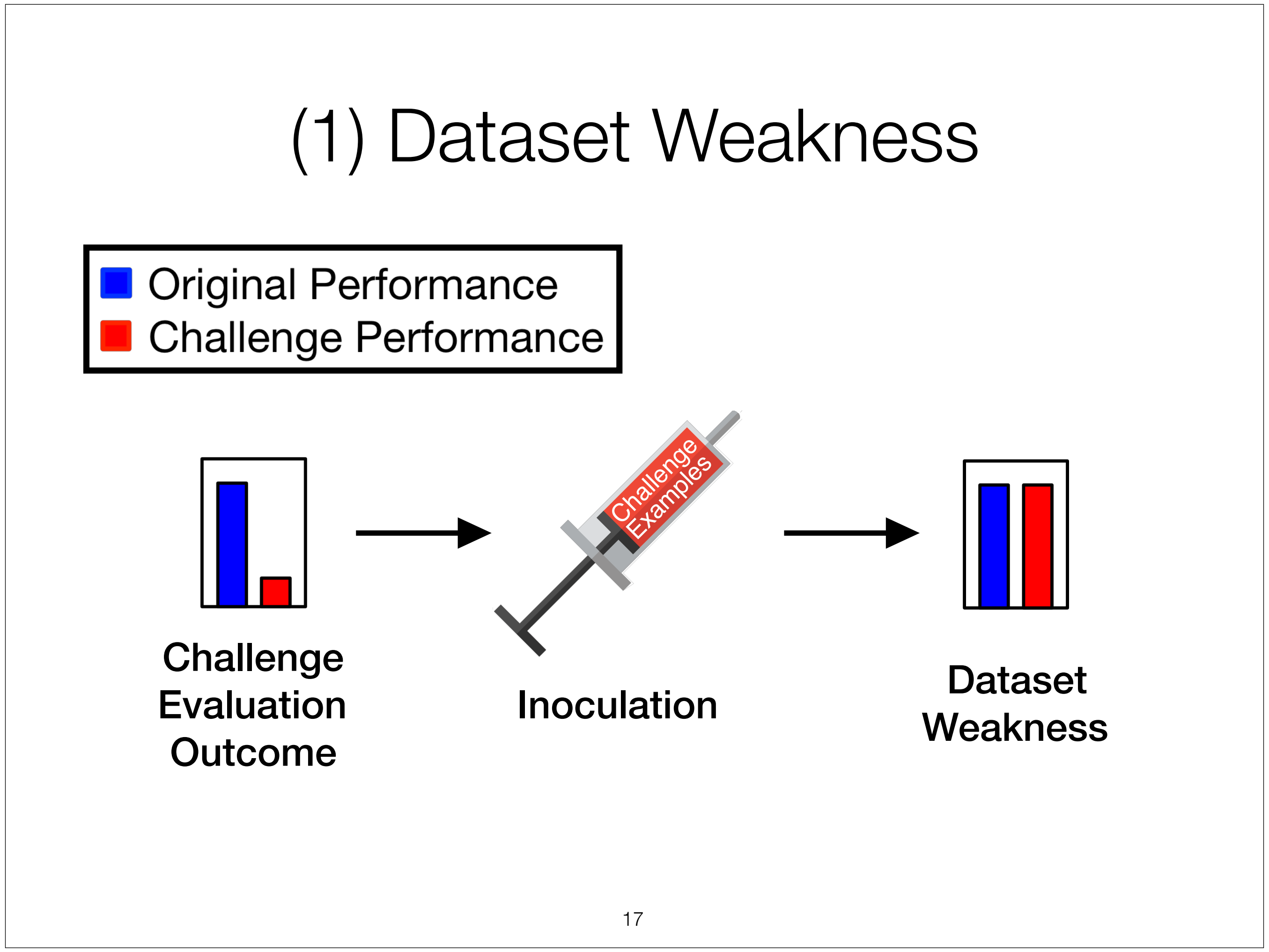
🟦 Original Performance
🟥 Challenge Performance



Challenge Evaluation Outcome

Inoculation

From inoculation, we studied three clear outcomes of interest.



(1) Dataset Weakness

🟦 Original Performance
🟥 Challenge Performance

Challenge Evaluation Outcome

Inoculation

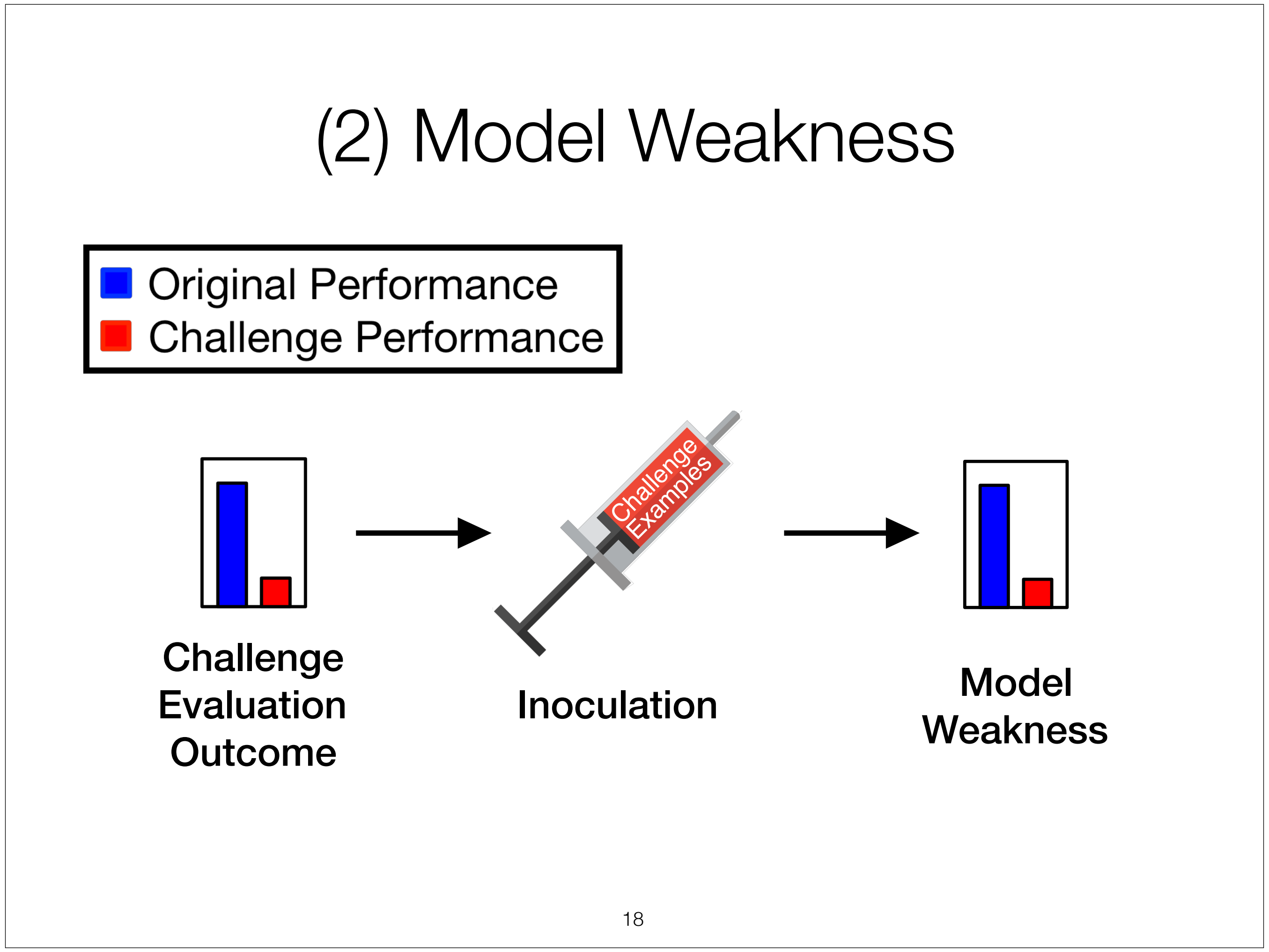
Dataset Weakness

In the first outcome, fine-tuning on a few challenge examples closes the gap, and the inoculated system performs well on both the original and challenge datasets.

This case suggests that the challenge dataset did not reveal a weakness in the model family, since we were able to overcome the challenge with just a bit of fine-tuning.

Instead, the challenge has likely revealed a lack of diversity in the original dataset.

When we see this outcome, we characterize it as a dataset weakness.



(2) Model Weakness

🟦 Original Performance
🟥 Challenge Performance

Challenge Evaluation Outcome

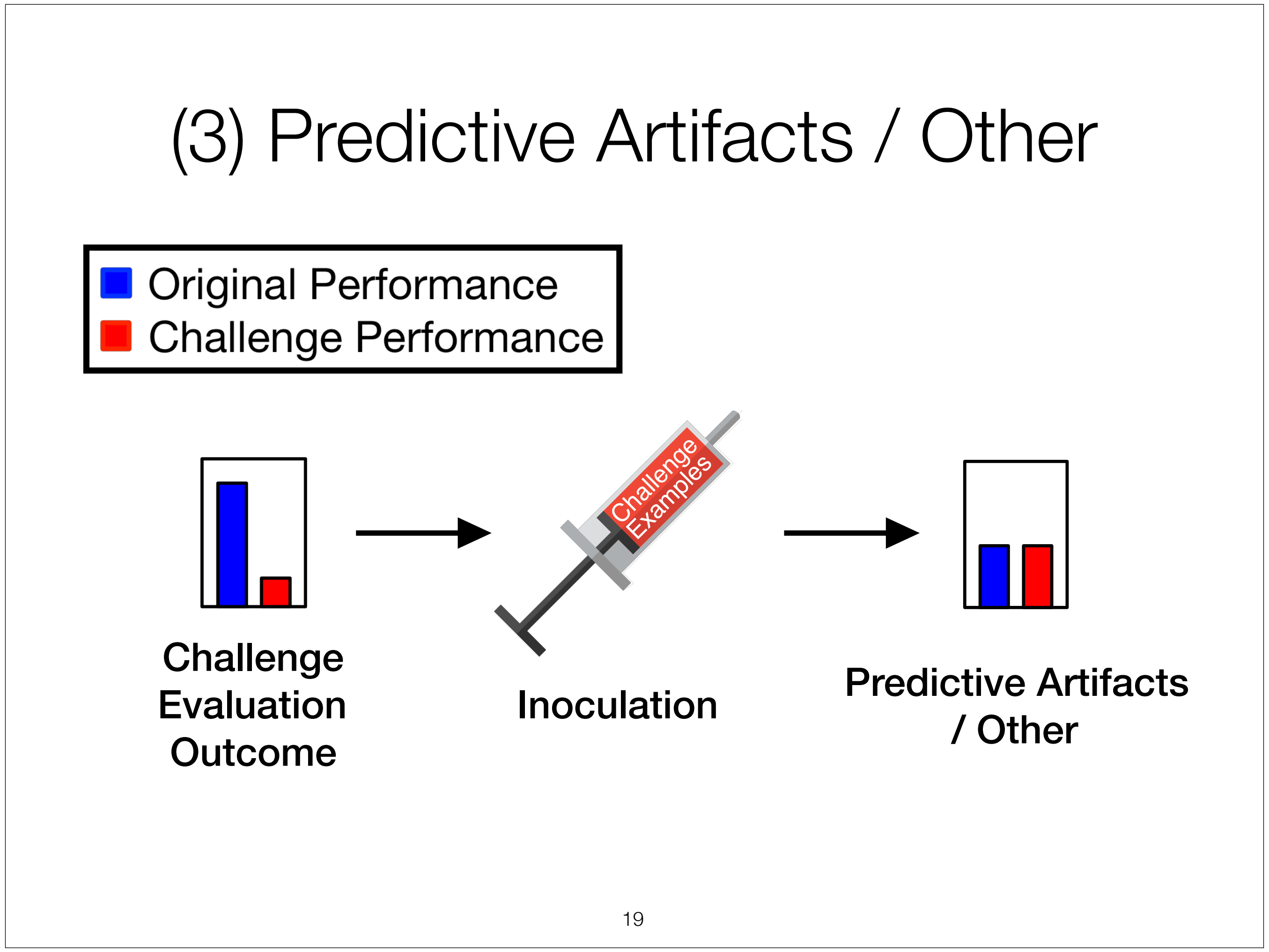
Inoculation

Model Weakness

In the second outcome, fine-tuning on a few challenge examples does not affect performance on either test set.

This indicates that the challenge dataset has revealed a fundamental weakness of the model; it is unable to adapt to the challenge data distribution, even with some exposure.

When we see this result, we classify it as a "model weakness".



(3) Predictive Artifacts / Other

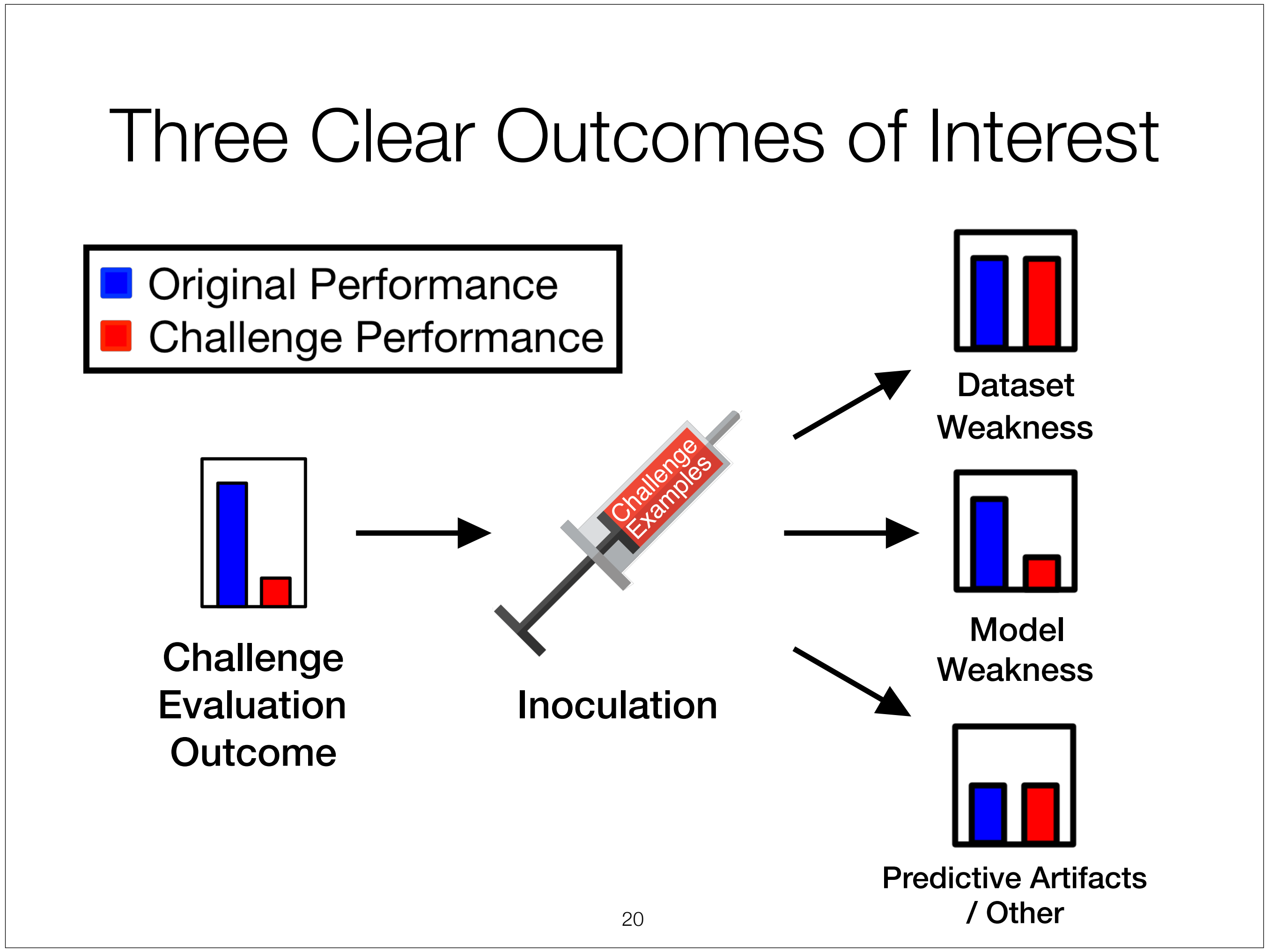
🟦 Original Performance
🟥 Challenge Performance

Challenge Evaluation Outcome → Inoculation → Predictive Artifacts / Other

In the third outcome, we see that inoculation damages performance on the original test set.

The main difference between this outcome and the previous two outcomes is that here, by fine-tuning, the model is shifting towards a challenge distribution that somehow contradicts the original distribution.

For instance, this could result from predictive features that exist in one dataset but not in the other.



Three Clear Outcomes of Interest

🟦 Original Performance
🟥 Challenge Performance

Challenge Evaluation Outcome

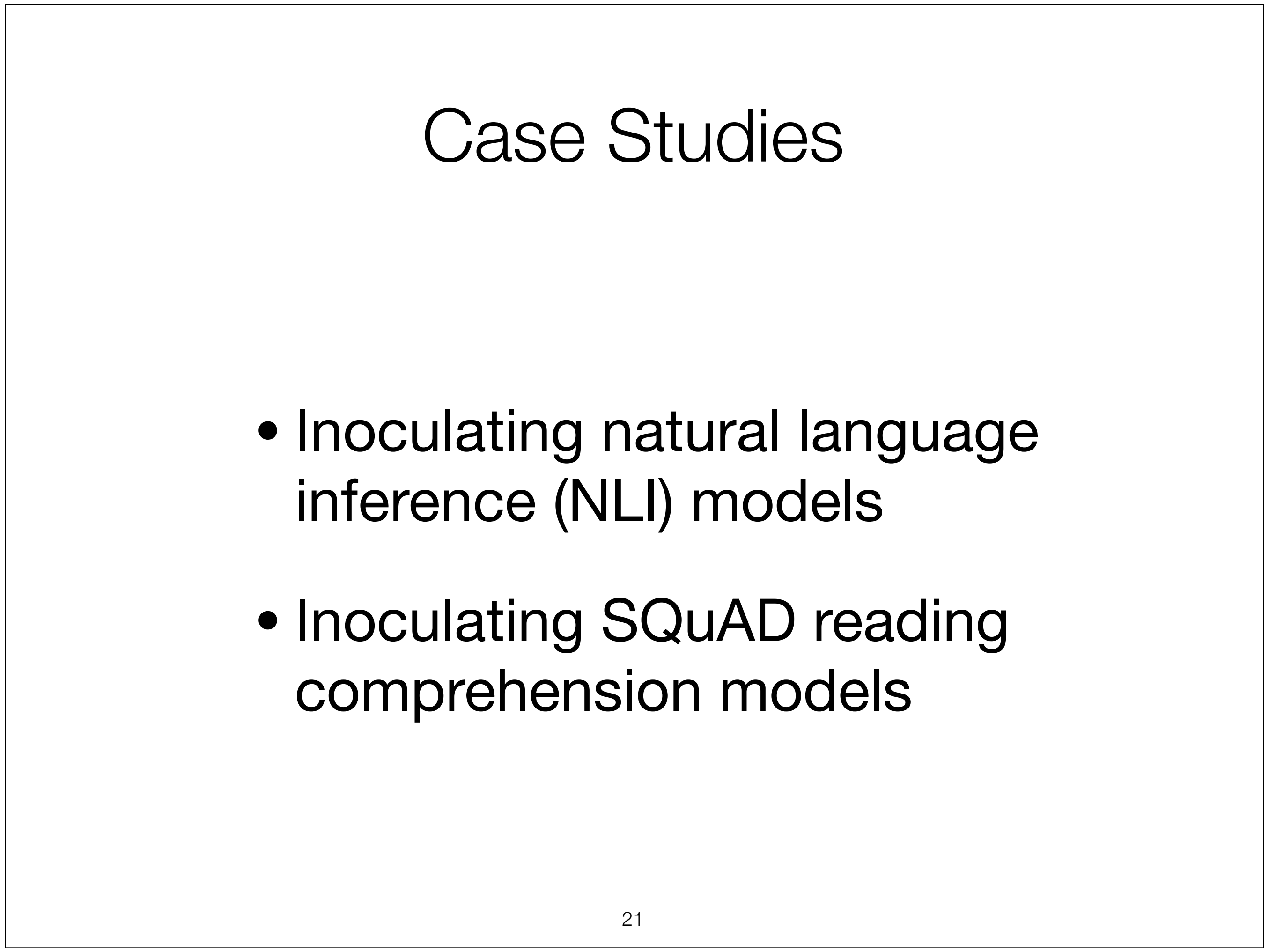
Inoculation

Dataset Weakness

Model Weakness

Predictive Artifacts / Other

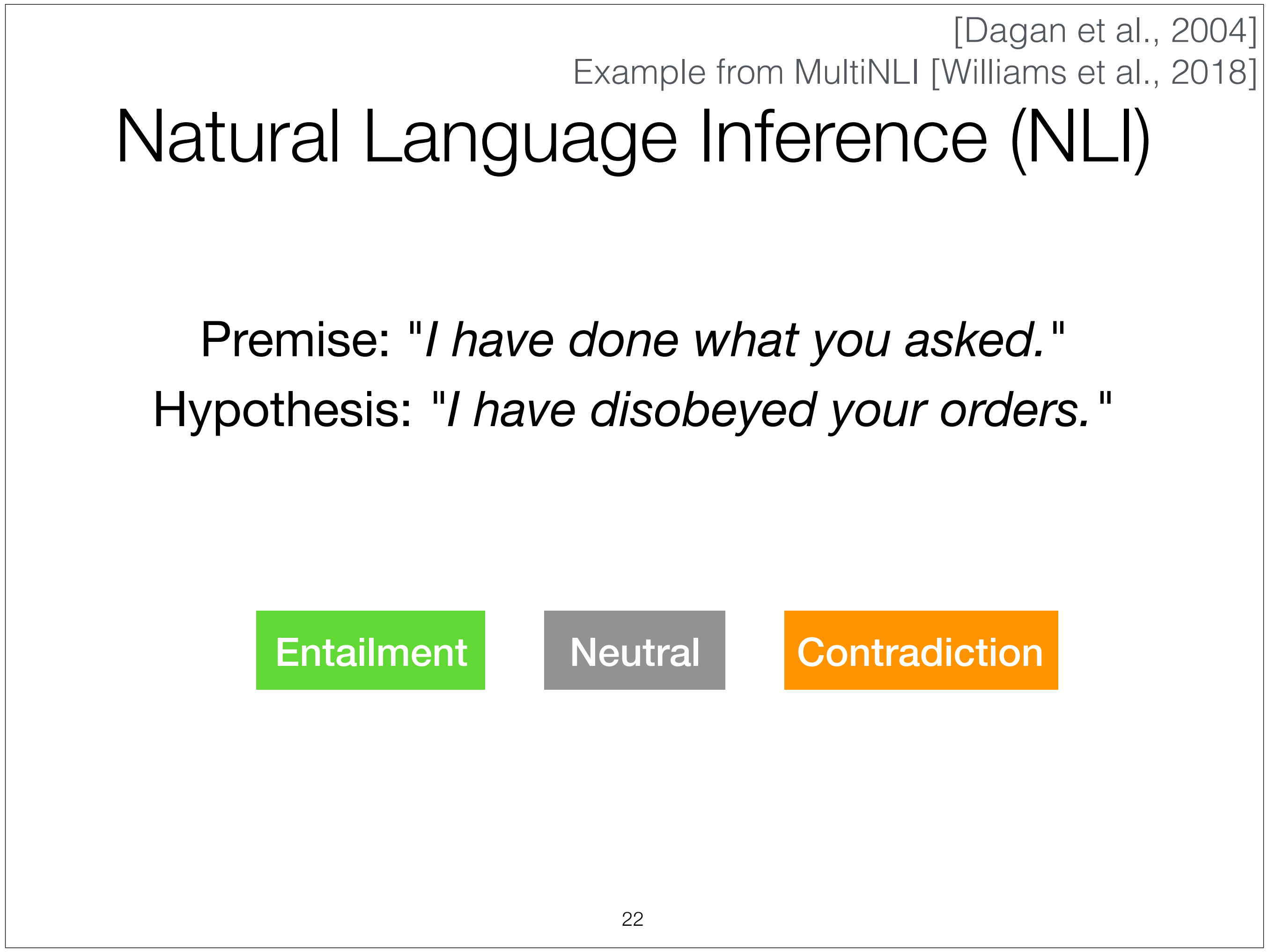
To summarize, inoculation by fine-tuning exposes models trained on some original dataset to additional examples from a challenge dataset of interest. Here are the three clear outcomes of interest that we looked at in this work. Of course, the outcome may also lie between these extremes.
(take a drink of water, pause)



Case Studies

- Inoculating natural language inference (NLI) models
- Inoculating SQuAD reading comprehension models

Now that I've described Inoculation by Fine-Tuning, I'm going to show how we used it to characterize the lack of robustness in NLI and SQuAD reading comprehension models.



[Dagan et al., 2004]
Example from MultiNLI [Williams et al., 2018]

Natural Language Inference (NLI)

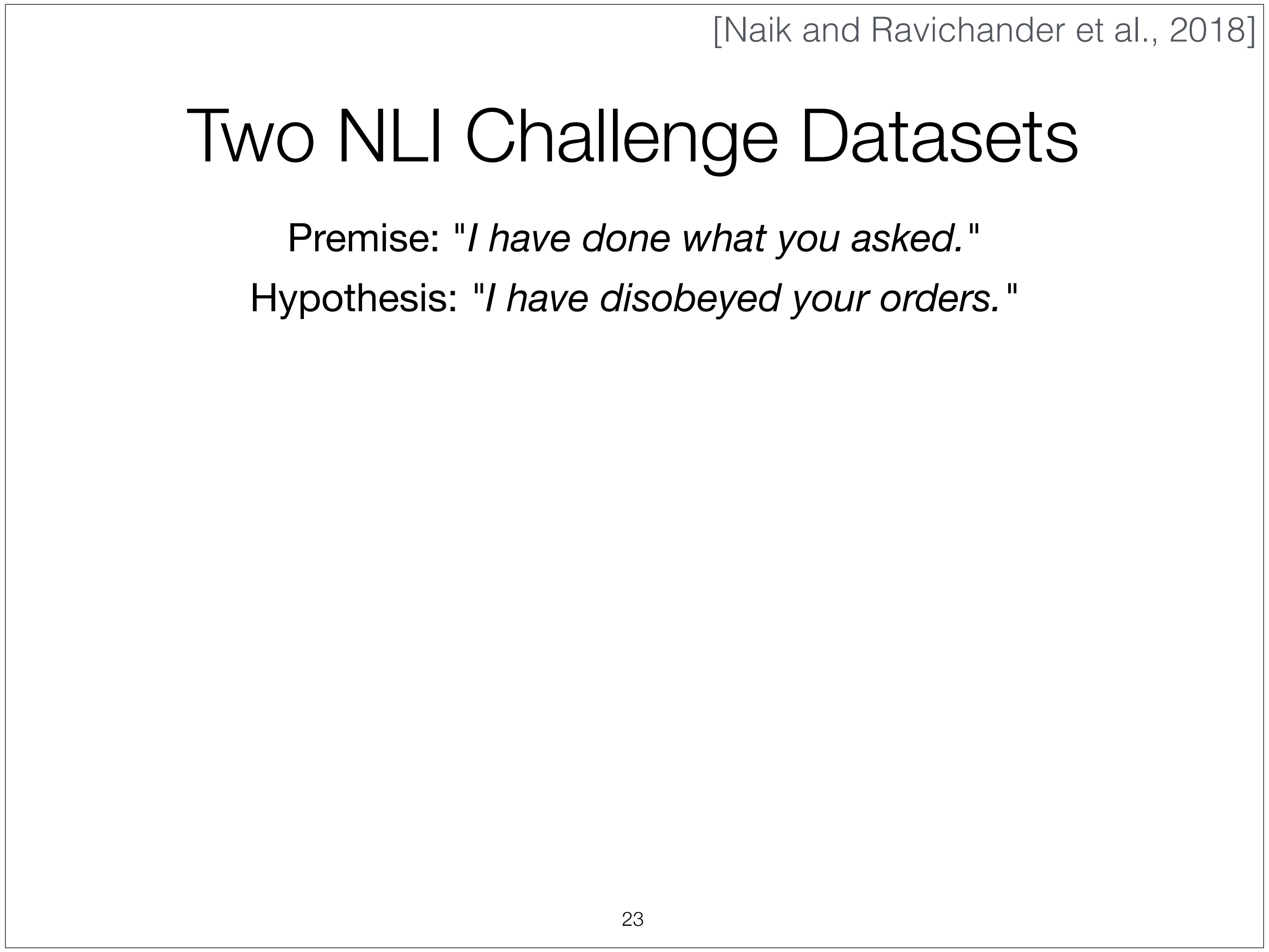
Premise: "*I have done what you asked.*"
Hypothesis: "*I have disobeyed your orders.*"

Entailment

Neutral

Contradiction

As a quick reminder, in the natural language inference, or NLI, task, the model is given a pair of sentences and asked to judge their relationship as entailment, neutral, or contradiction.



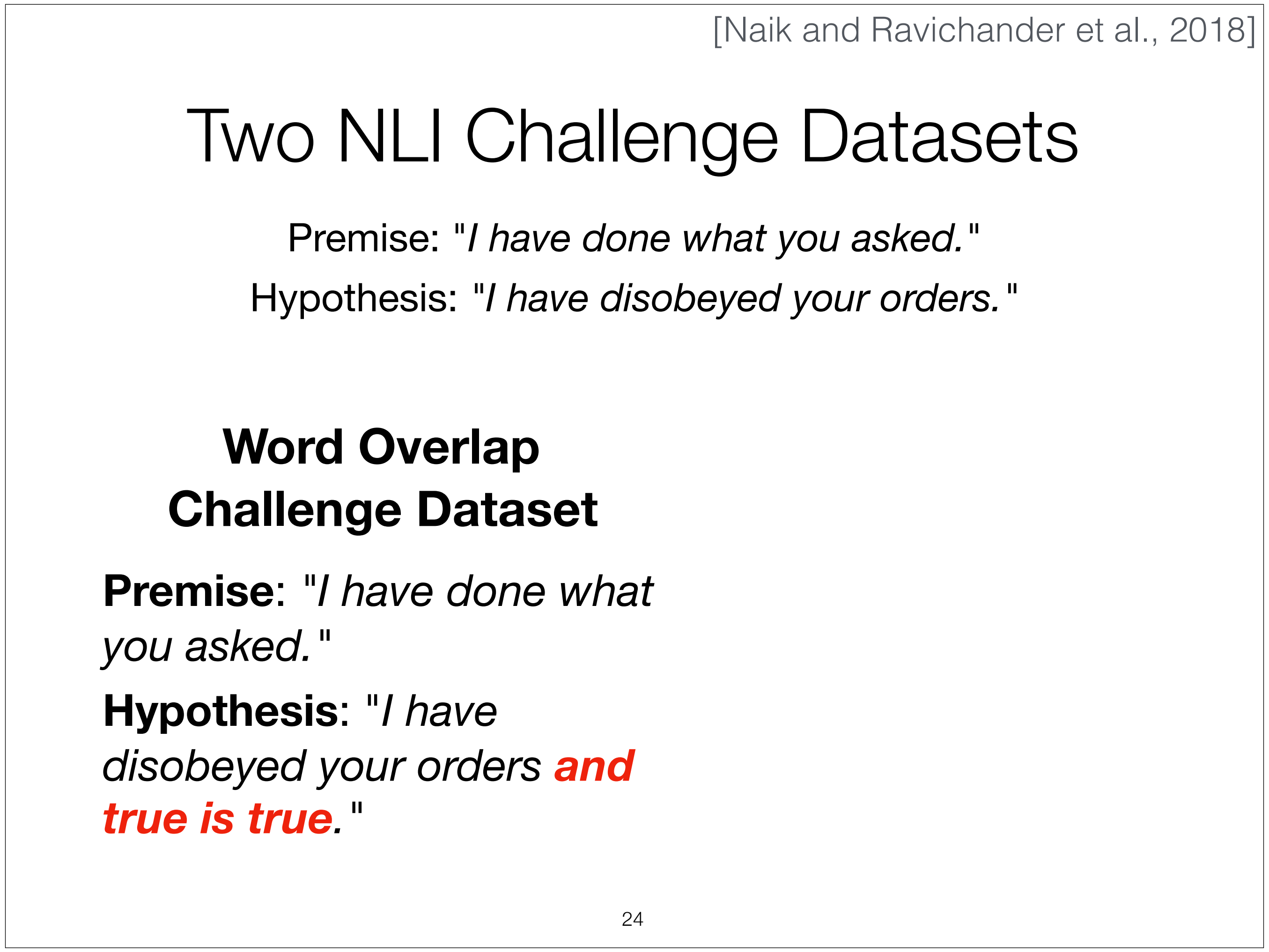
[Naik and Ravichander et al., 2018]

Two NLI Challenge Datasets

Premise: "*I have done what you asked.*"

Hypothesis: "*I have disobeyed your orders.*"

Last year, Naik and Ravichander et al proposed several challenge datasets for NLI. I'll present two here and show how inoculation by fine-tuning helps us draw conclusions about why they're difficult for models.



[Naik and Ravichander et al., 2018]

Two NLI Challenge Datasets

Premise: "*I have done what you asked.*"

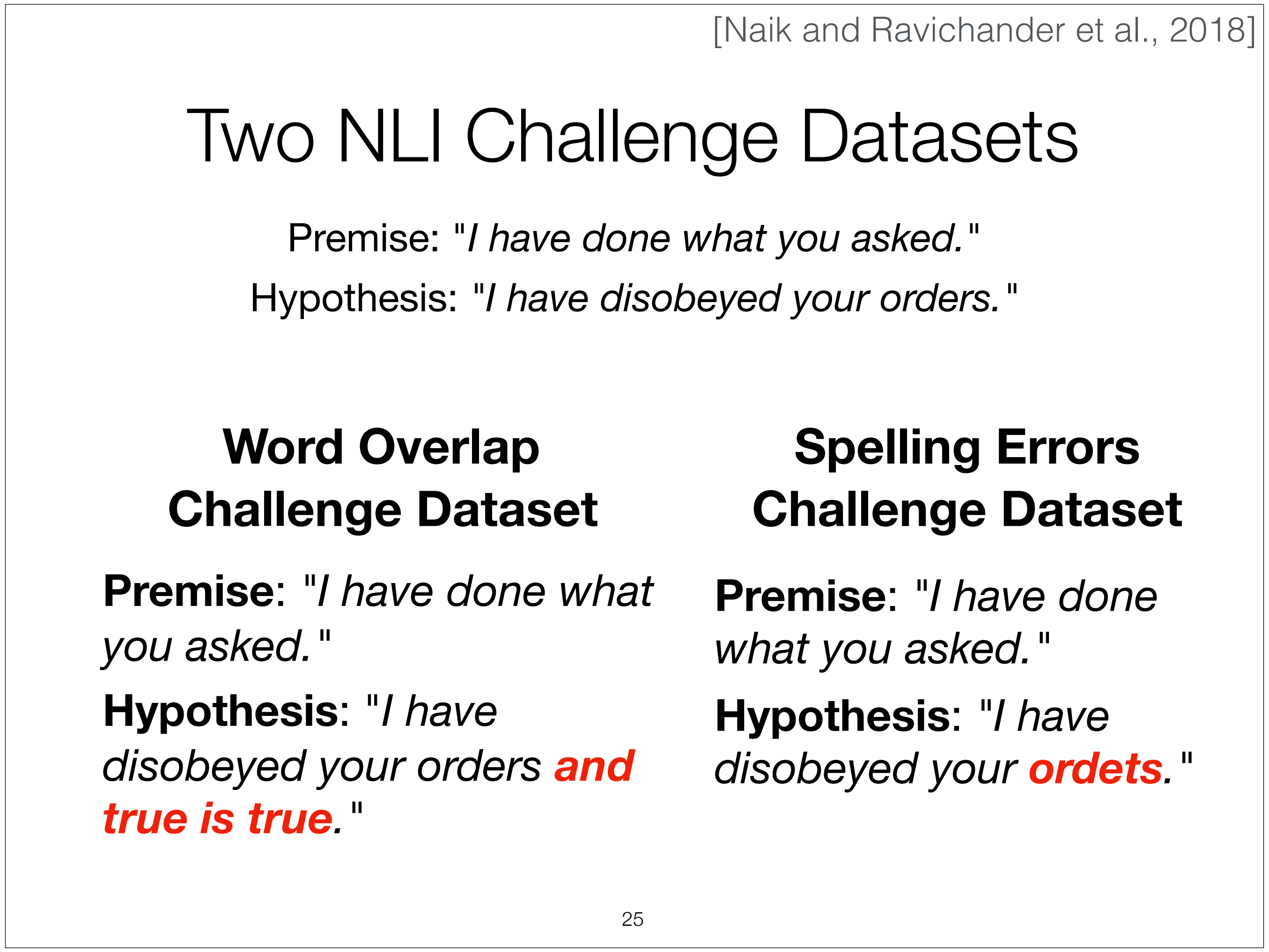
Hypothesis: "*I have disobeyed your orders.*"

Word Overlap Challenge Dataset

Premise: "*I have done what you asked.*"

Hypothesis: "*I have disobeyed your orders* ***and true is true.***"

In the first challenge dataset, called the word overlap challenge, the phrase "and true is true" is appended to the end of every hypothesis.



[Naik and Ravichander et al., 2018]

Two NLI Challenge Datasets

Premise: "*I have done what you asked.*"

Hypothesis: "*I have disobeyed your orders.*"

Word Overlap Challenge Dataset

Premise: "*I have done what you asked.*"

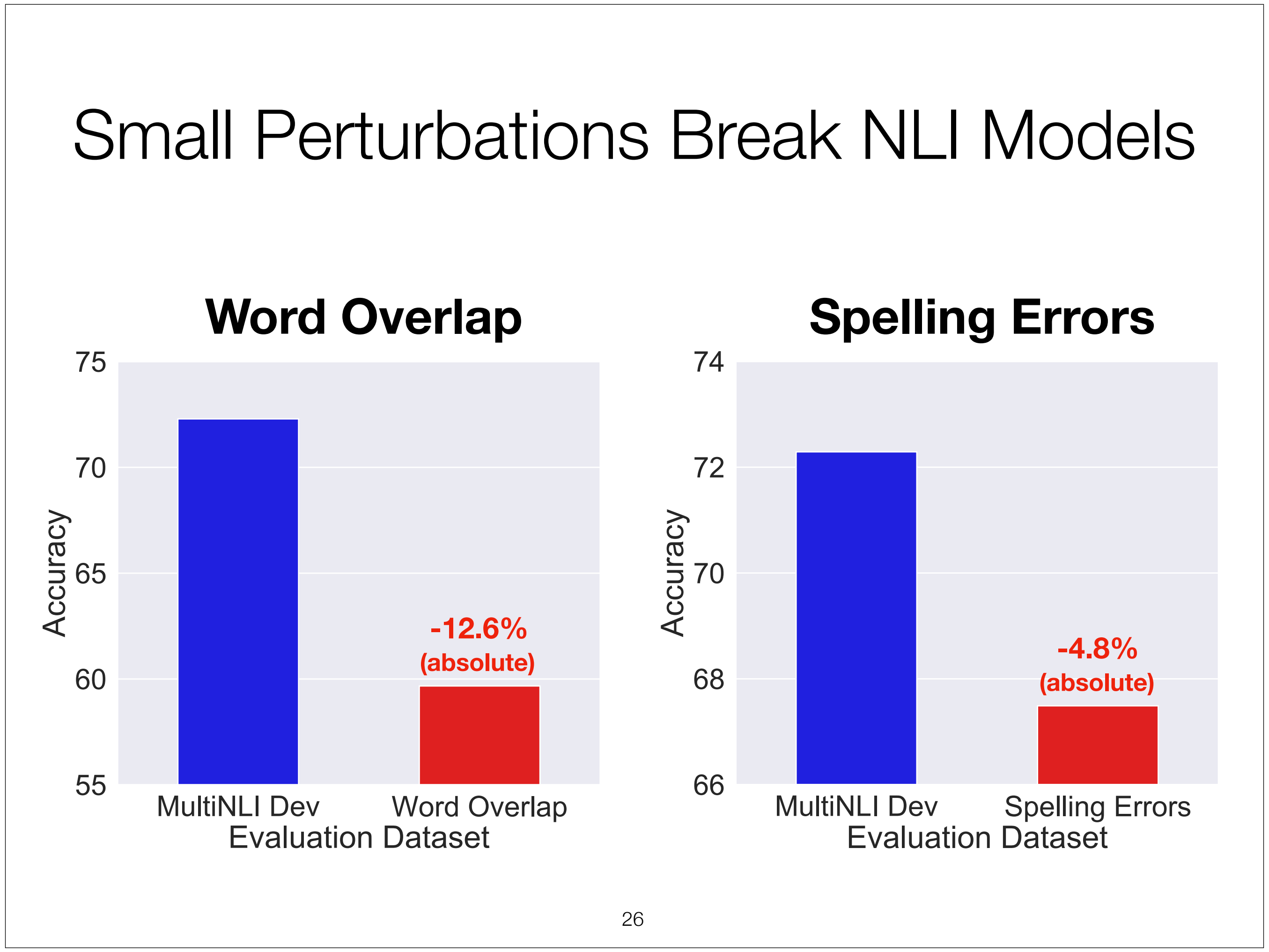
Hypothesis: "*I have disobeyed your orders **and true is true.***"

Spelling Errors Challenge Dataset

Premise: "*I have done what you asked.*"

Hypothesis: "*I have disobeyed your **ordets**.*"

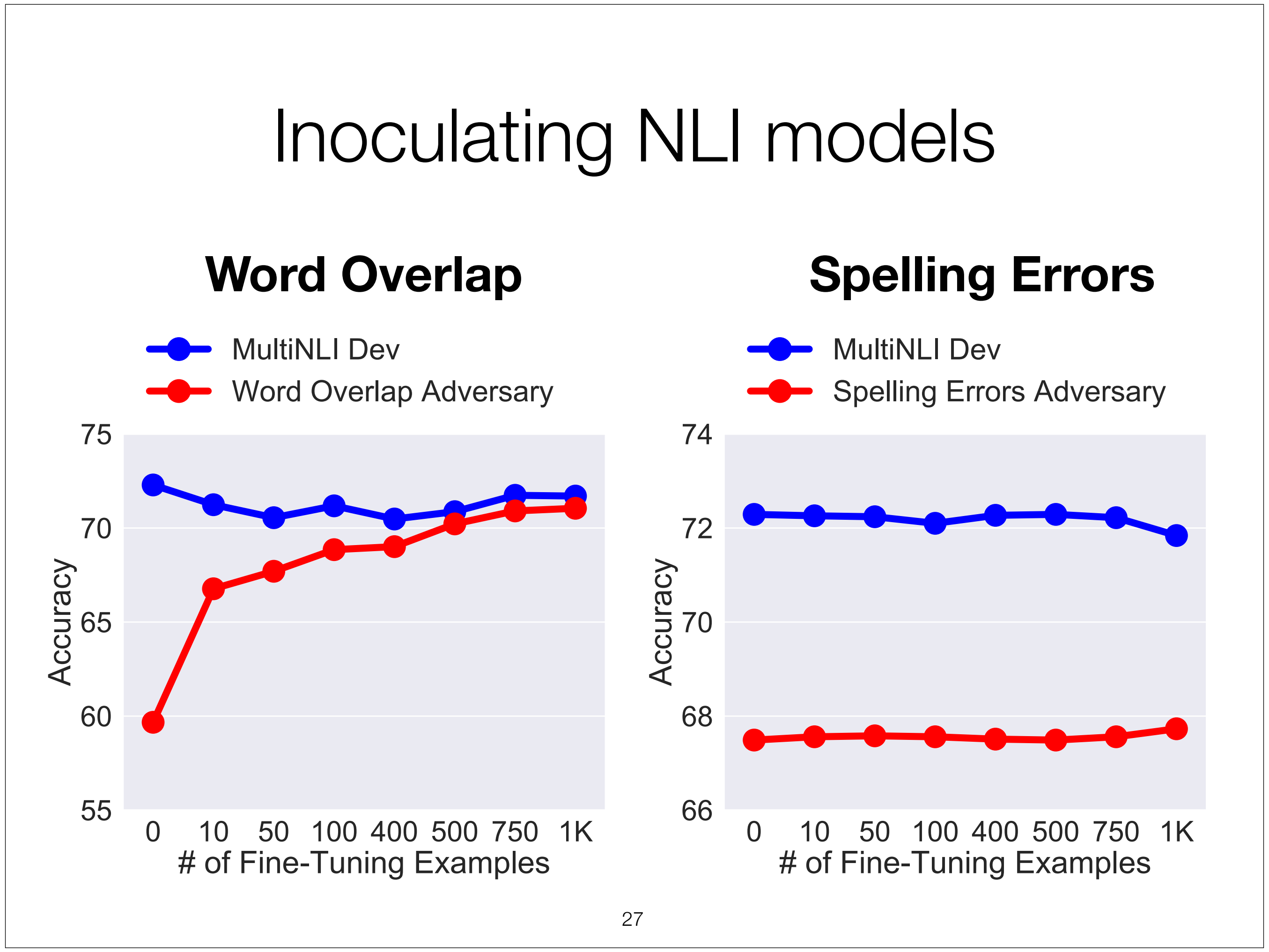
In the second challenge dataset, the spelling errors challenge, a random letter in a random word is swapped with one that is close by on the keyboard---this is meant to simulate natural human typos.



Small Perturbations Break NLI Models

These perturbations break NLI models.

Note that the y axes on these two plots are different. The accuracy of the decomposable attention model suffers an absolute drop of 12.6 points when the word overlap challenge is applied to the MultiNLI dev set, and simply swapping one character in one word results in a loss of nearly 5 accuracy points as well.



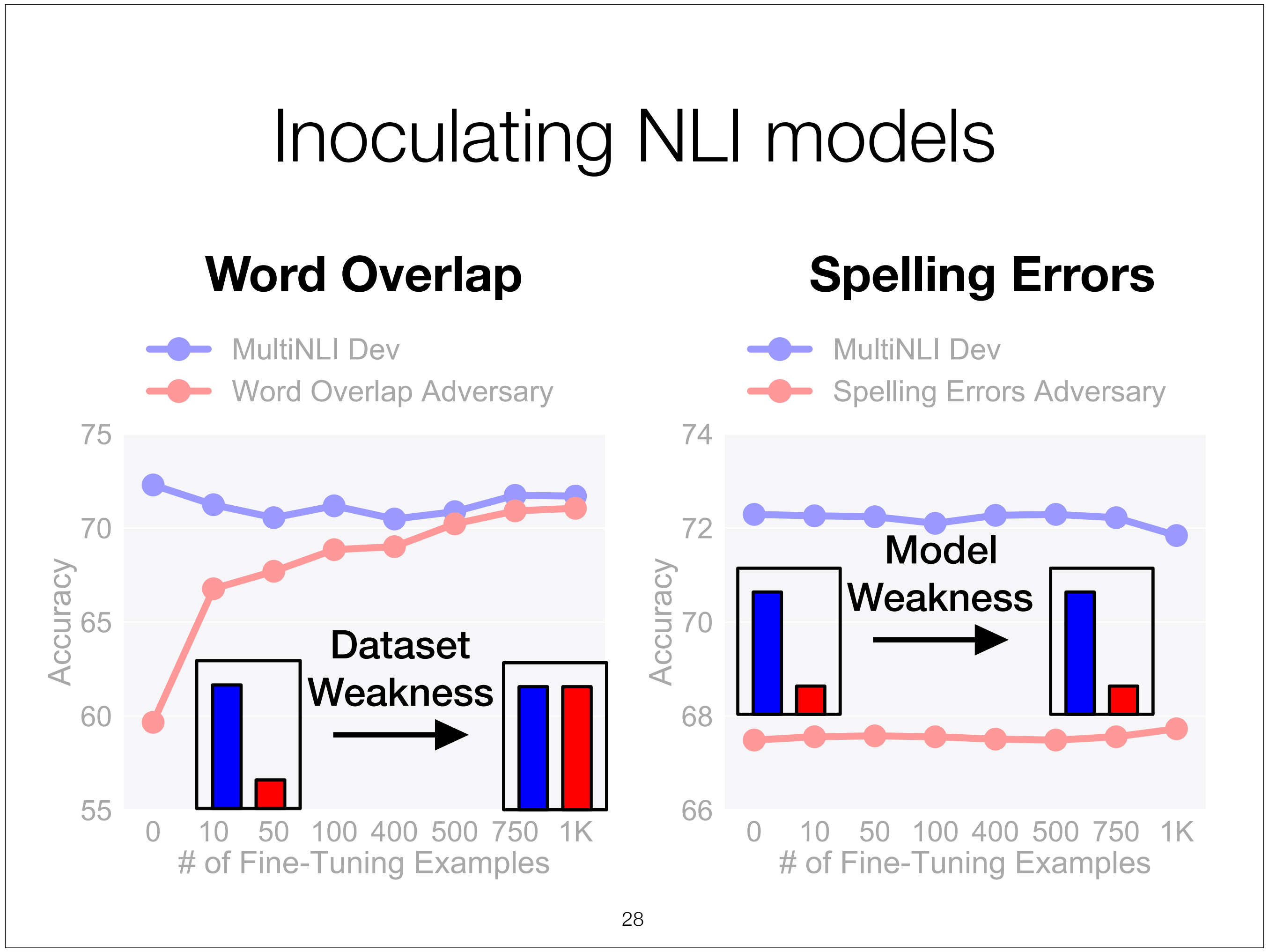
Inoculating NLI models

This slide shows the results of applying inoculation against these two challenge datasets. On the left, we've got the word overlap results and on the right, we've got the spelling error results. The x axis shows the number of challenge dataset examples we fine-tune on, and the y axis shows the performance. Note that the y axes on these two plots have different ranges.

So, looking first at the word overlap results, we can see that the model is able to quickly close the performance gap when it's fine-tuned on a small number of challenge examples.

On the other hand, fine-tuning on spelling errors doesn't make too much of a difference---the gap remains more or less constant.

So, despite the fact that these two challenge datasets both break our models, the behavior after we fine-tune on some challenge examples is clearly very different.

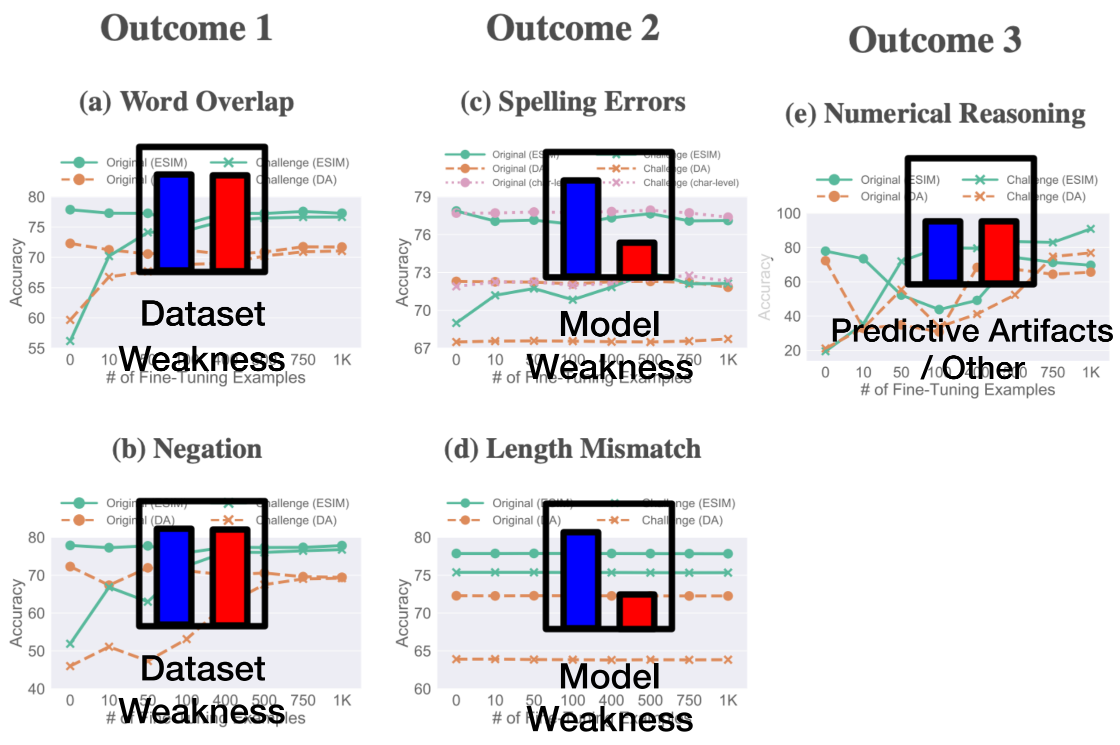


Inoculating NLI models

Putting these results in the context of the three inoculation outcomes that I described earlier, we can categorize the failures on the word overlap challenge as a dataset weakness. The challenge dataset couldn't have been stressing an inherent limitation of the model, because we closed the gap with just a bit of fine-tuning.

On the other hand, the failures on spelling errors are more of a model weakness---even after fine-tuning, we're unable to improve on the challenge dataset, although performance on the original dataset does not degrade either.

More Examples in the Paper!



Outcome 1

(a) Word Overlap

Dataset Weakness

(b) Negation

Dataset Weakness

Outcome 2

(c) Spelling Errors

Model Weakness

(d) Length Mismatch

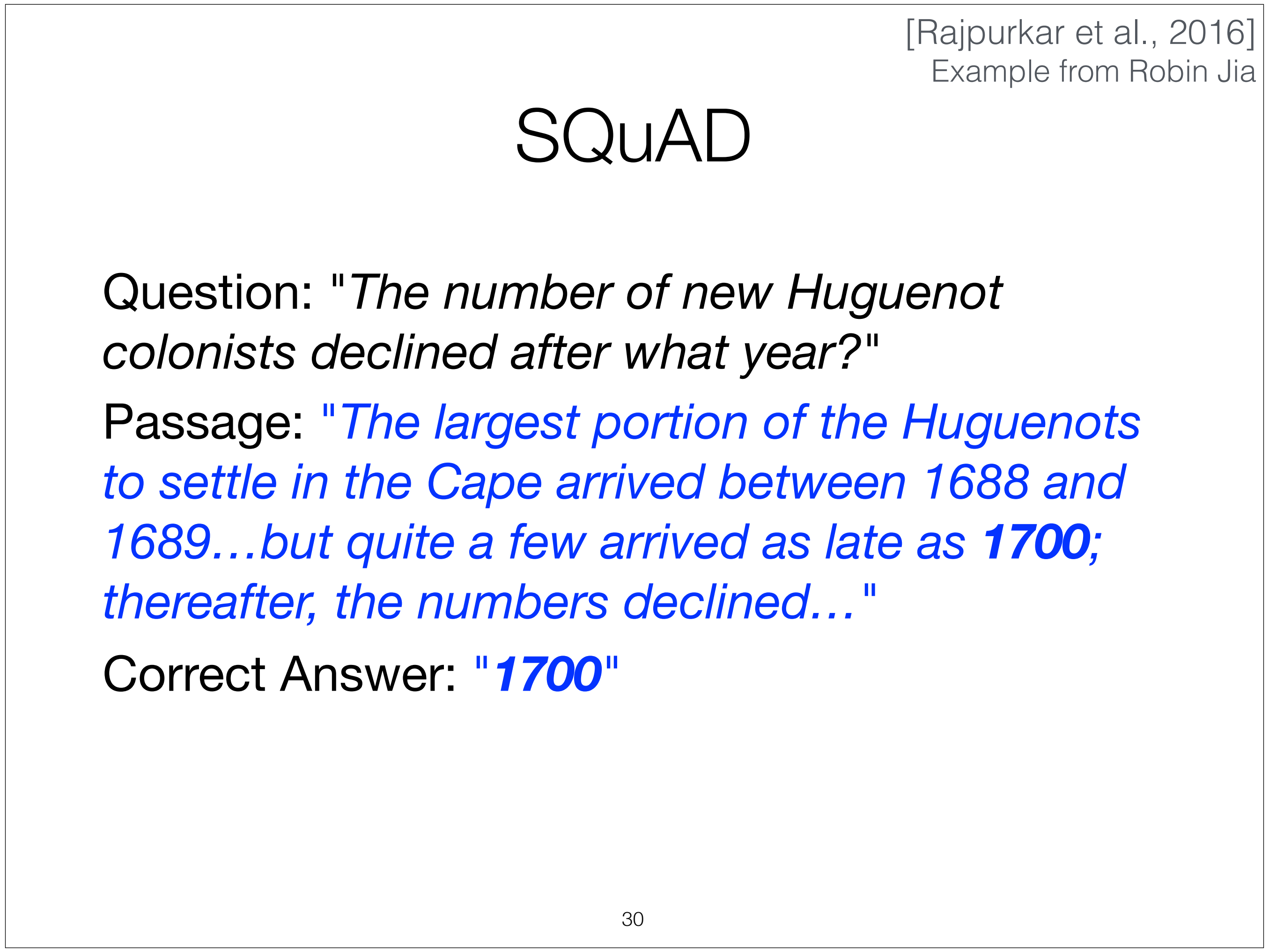
Model Weakness

Outcome 3

(e) Numerical Reasoning

Predictive Artifacts / Other

We saw similar trends when inoculating the ESIM model, and the paper also has experiments with more NLI challenge datasets.



[Rajpurkar et al., 2016]
Example from Robin Jia

SQuAD

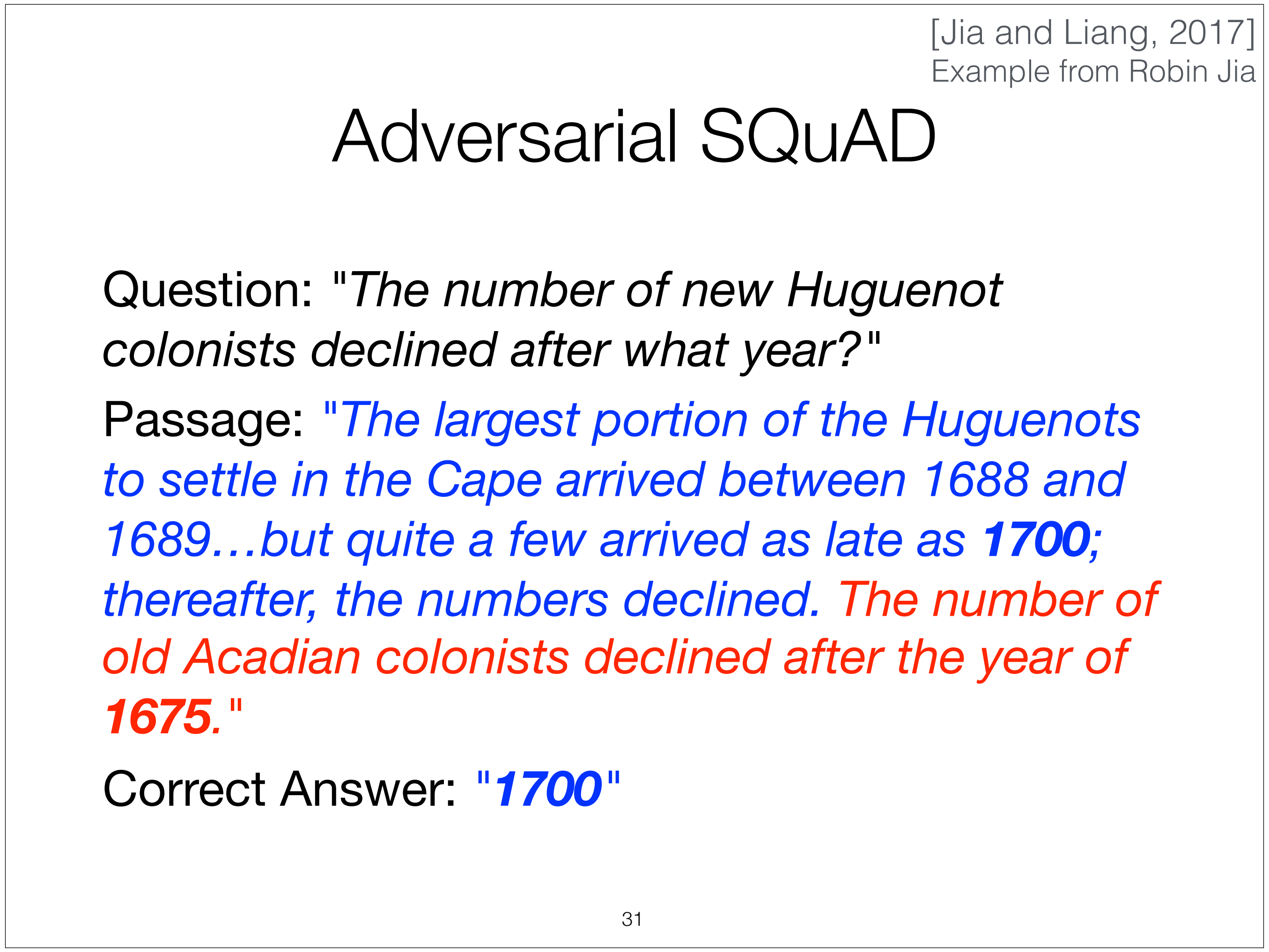
Question: "*The number of new Huguenot colonists declined after what year?*"

Passage: "*The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as **1700**; thereafter, the numbers declined…*"

Correct Answer: "***1700***"

We also looked at inoculating SQuAD reading comprehension models.

In the SQuAD dataset, a model is given a question and a passage, and is trained to answer the question by selecting a span from within the passage.



[Jia and Liang, 2017]
Example from Robin Jia

Adversarial SQuAD

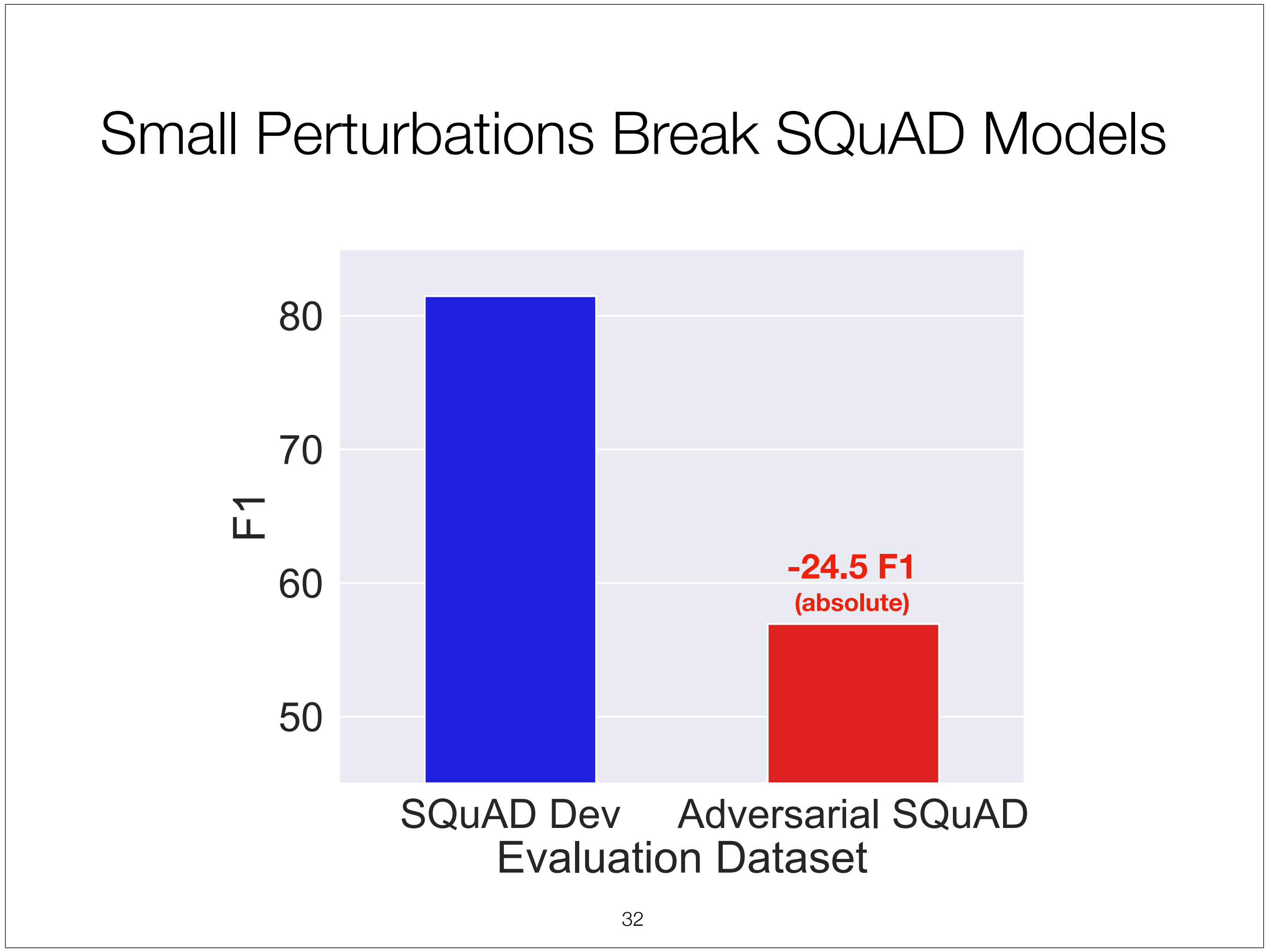
Question: "*The number of new Huguenot colonists declined after what year?*"

Passage: "*The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as **1700**; thereafter, the numbers declined. The number of old Acadian colonists declined after the year of **1675**.*"

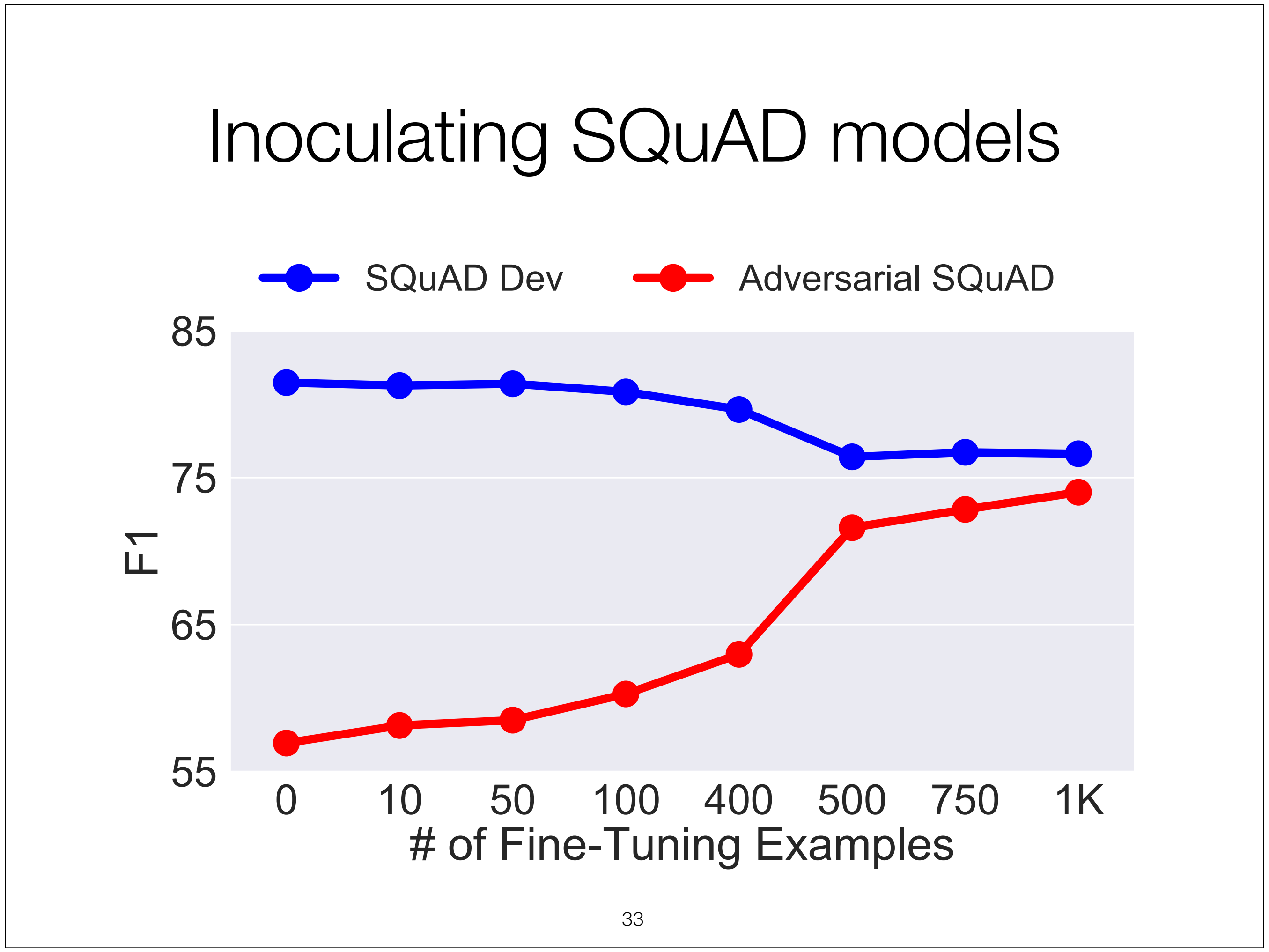
Correct Answer: "***1700***"

Jia and Liang proposed an adversarial SQuAD dataset that appends a distracting sentence, in red, to the end of every passage. This distracting sentence generally has high lexical overlap with the question, and models are often fooled into predicting an incorrect answer.

Small Perturbations Break SQuAD Models

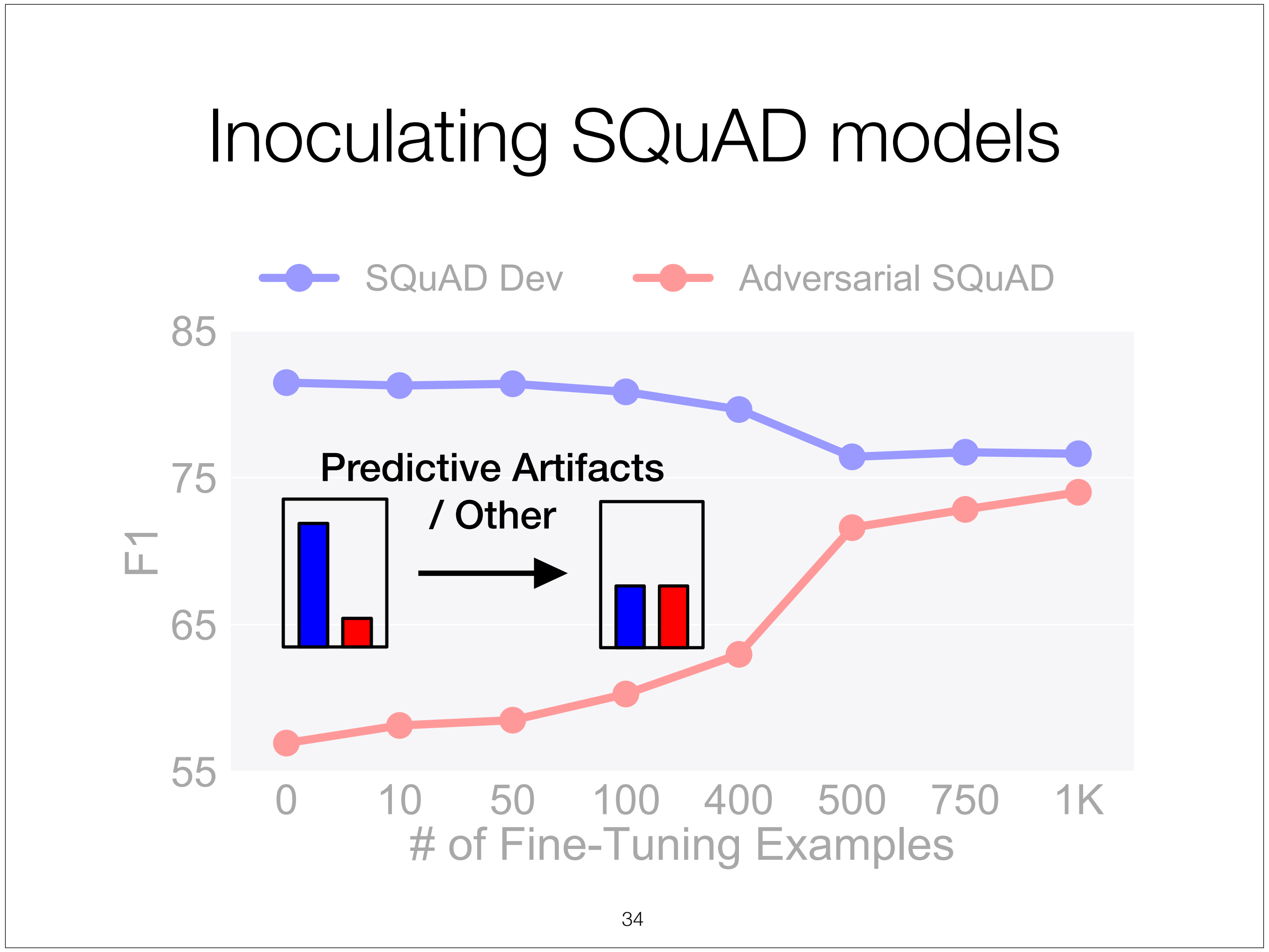


In the model that we looked at, QANet, appending these distractor sentences degraded performance on the SQuAD development set by around 24.5 F1 points.



Inoculating SQuAD models

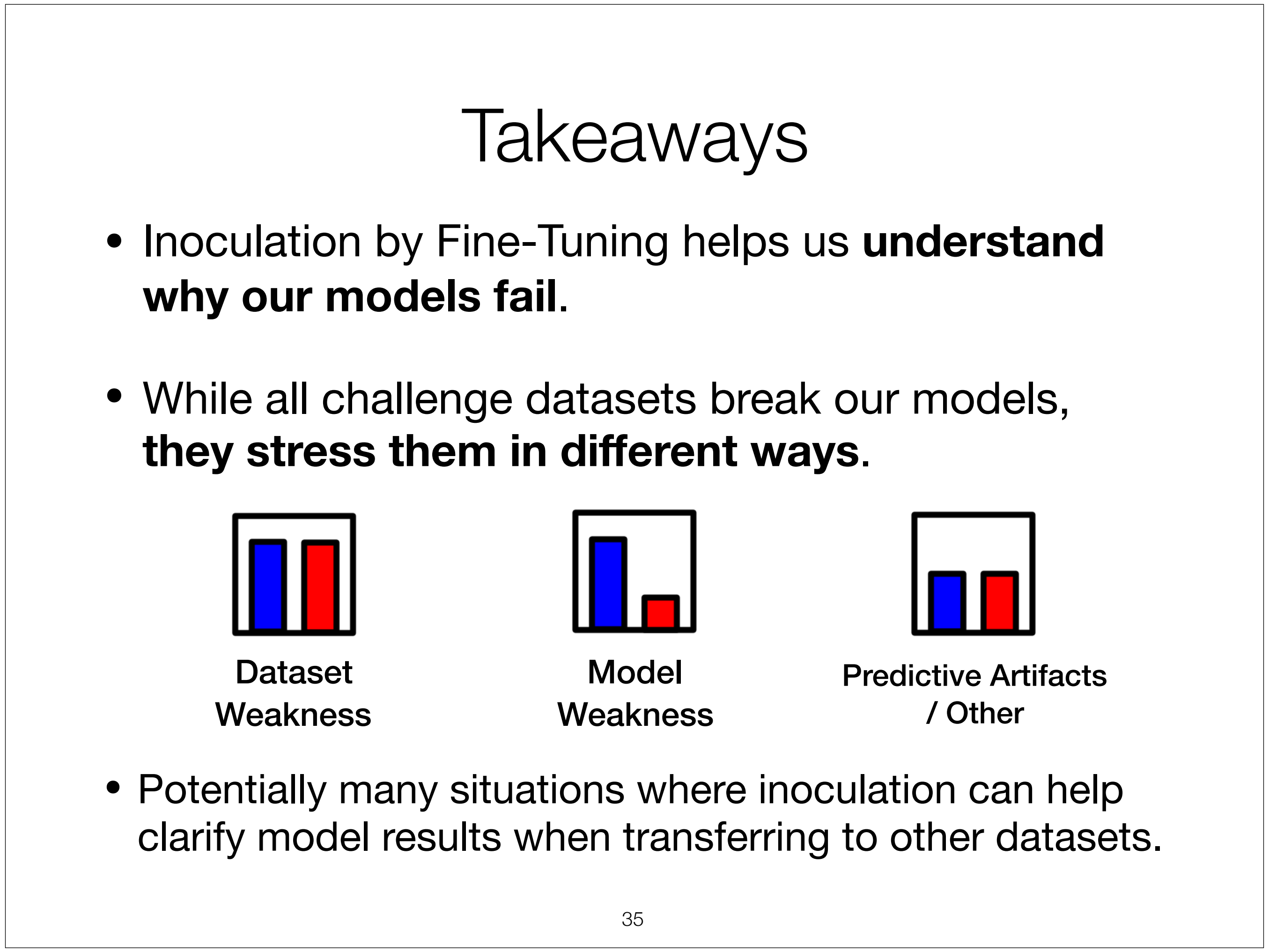
When inoculating our model with examples from the adversarial SQuAD dataset, we see that fine-tuning on more challenge examples degrades our performance on the original dataset while improving performance on the challenge dataset.



Inoculating SQuAD models

This corresponds to the third outcome that I presented. In this case, the model has learned to exploit predictive features in the adversarial SQuAD dataset that don't generalize to the original SQuAD dataset---namely, it learns to ignore the last sentence in the passage. We saw similar results with the BiDAF model.

This outcome yields insights into the challenge datasets themselves. Specifically, a challenge dataset that a model *can* recover from when fine-tuning, at the expense of performance on the original dataset, probably isn't testing the full breadth of a linguistic phenomenon, but rather a *particular exploitable* aspect of it.



Takeaways

- Inoculation by Fine-Tuning helps us **understand why our models fail**.
- While all challenge datasets break our models, **they stress them in different ways**.

Dataset Weakness

Model Weakness

Predictive Artifacts / Other

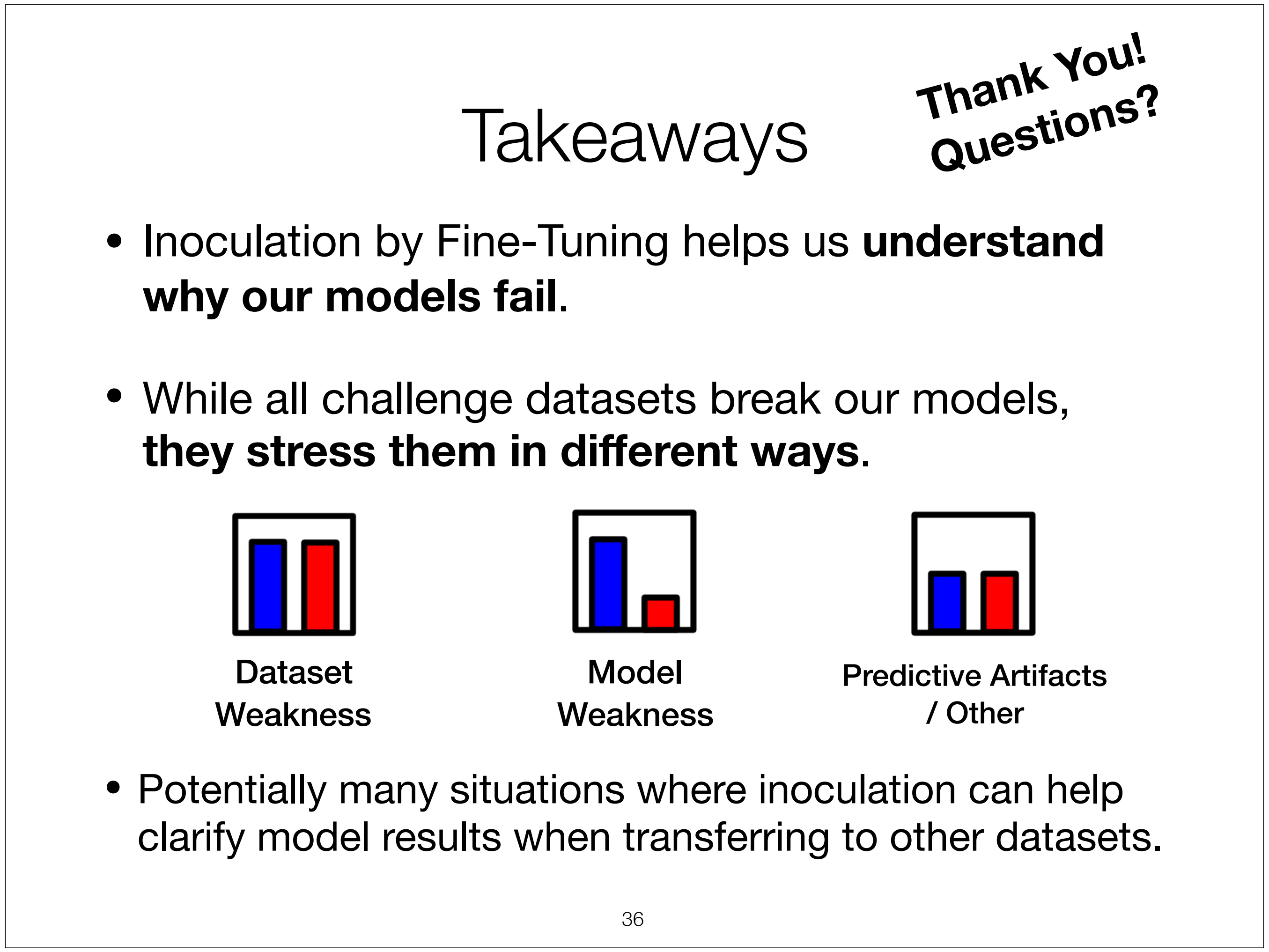
- Potentially many situations where inoculation can help clarify model results when transferring to other datasets.

In terms of takeaways:

Inoculation by Fine-Tuning is a method for better understanding why our models fail, giving us information we need to make them better.

While all challenge datasets break our models, they are certainly not all the same. Challenge datasets are not always difficult for the reasons we think they are, and failures on challenge datasets may lead to very different conclusions about models, training datasets, and the challenge datasets themselves. We've shown that different challenge datasets stress our models in different, and often unintuitive, ways, so they're all unique tools for helping us build better models.

Finally, while we specifically focused on better understanding why *challenge datasets* are difficult in this work, inoculation by fine-tuning is far more general than that --- it can be used in any case with a train-test distribution mismatch.



Takeaways

Thank You!
Questions?

- Inoculation by Fine-Tuning helps us **understand why our models fail**.
- While all challenge datasets break our models, **they stress them in different ways**.

Dataset Weakness

Model Weakness

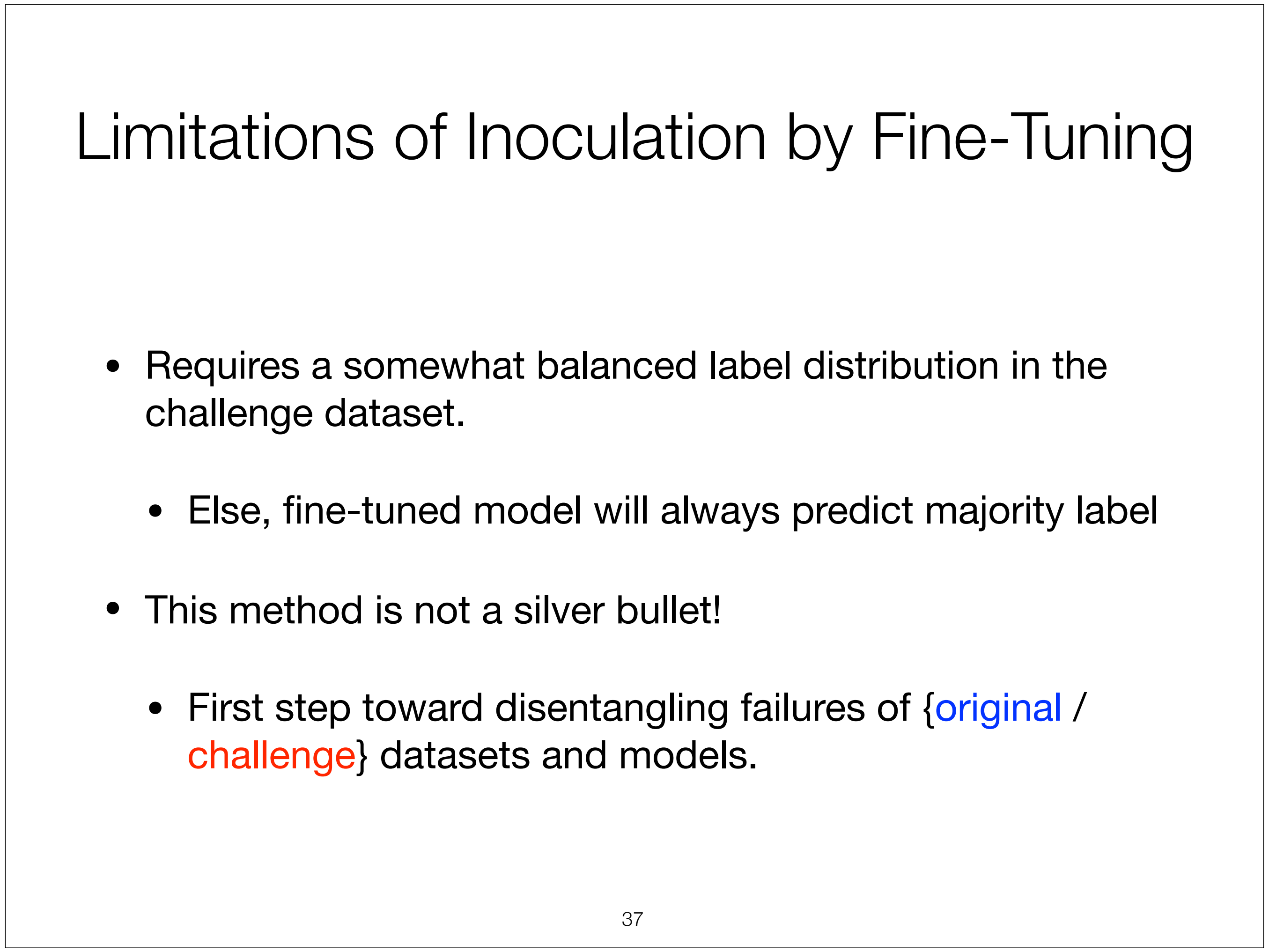
Predictive Artifacts / Other

- Potentially many situations where inoculation can help clarify model results when transferring to other datasets.

And that concludes my talk. Thanks for listening, and I'm happy to take questions now.

REPEAT THE QUESTION!!

3 examples of train-test mismatch: domain, language change over time.

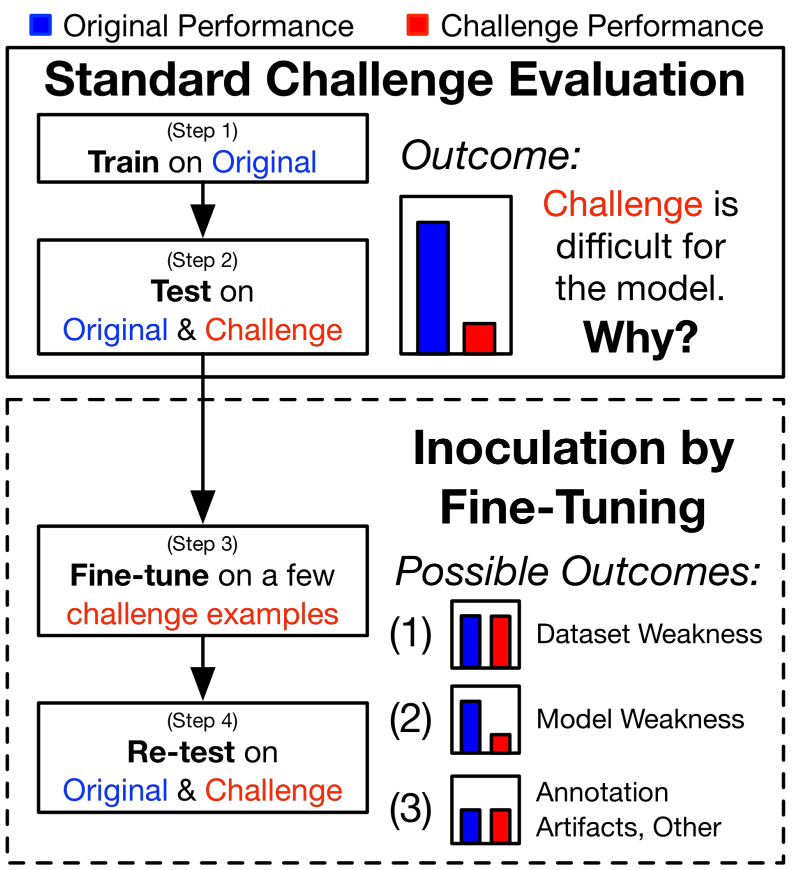


Limitations of Inoculation by Fine-Tuning

- Requires a somewhat balanced label distribution in the challenge dataset.
  - Else, fine-tuned model will always predict majority label
- This method is not a silver bullet!
  - First step toward disentangling failures of {original / challenge} datasets and models.

Now that I've told you a bit about inoculation by fine-tuning, I want to discuss some of the shortcomings.
For one, the method assumes a somewhat balanced label distribution in the challenge dataset. If a challenge dataset is highly skewed toward a certain label, fine-tuning might result in the model always predicting that label. In this case, you can't really draw any conclusions, because the model has completely deviated from the task.

This method isn't a silver bullet, but we think that it's a first step toward disentangling failures of datasets and models.



🟦 Original Performance 🟥 Challenge Performance

Standard Challenge Evaluation

(Step 1)
Train on Original

(Step 2)
Test on
Original & Challenge

Outcome:

Challenge is difficult for the model.
Why?

Inoculation by Fine-Tuning

(Step 3)
Fine-tune on a few challenge examples

(Step 4)
Re-test on
Original & Challenge

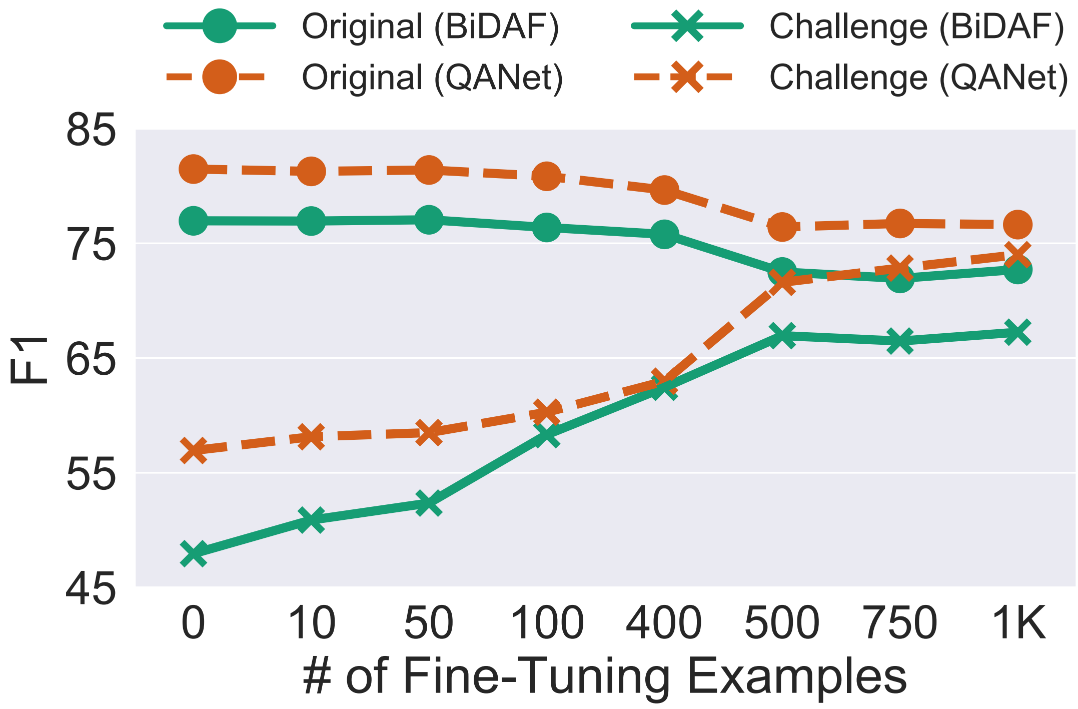
Possible Outcomes:

(1) Dataset Weakness

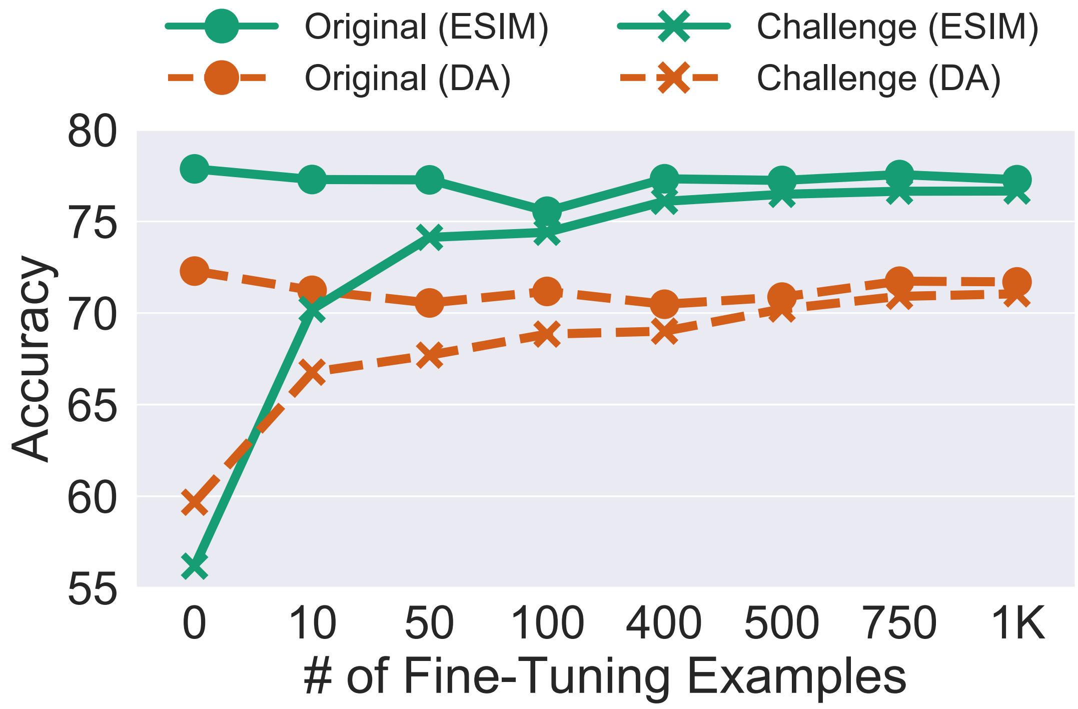
(2) Model Weakness

(3) Annotation Artifacts, Other

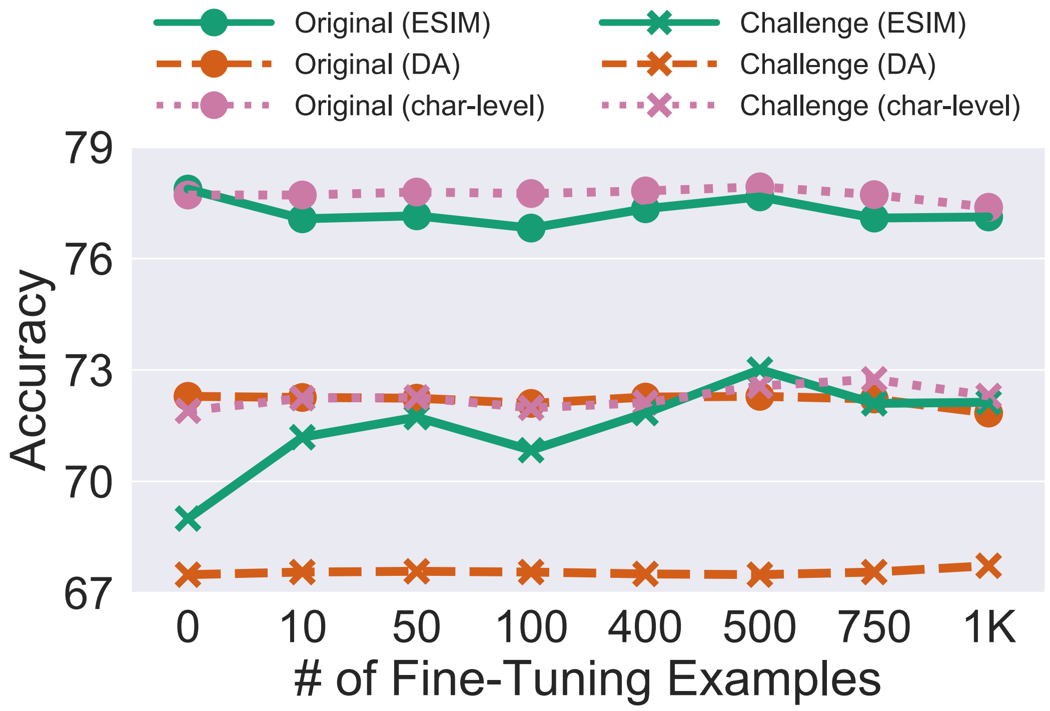
Inoculating Multiple SQuAD Reading Comprehension Models



Inoculating Multiple NLI Models Against Word Overlap Adversary



Inoculating Multiple NLI Models Against Spelling Errors


Original (ESIM)
Original (DA)
Original (char-level)
Challenge (ESIM)
Challenge (DA)
Challenge (char-level)
Accuracy
1K
# of Fine-Tuning Examples