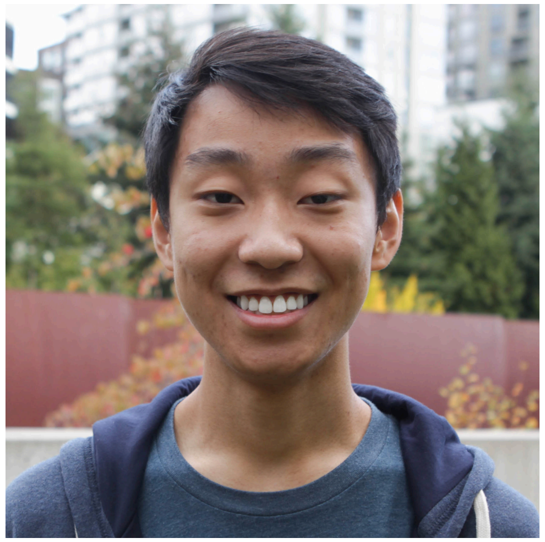
Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets



Nelson F. Liu

Roy Schwartz

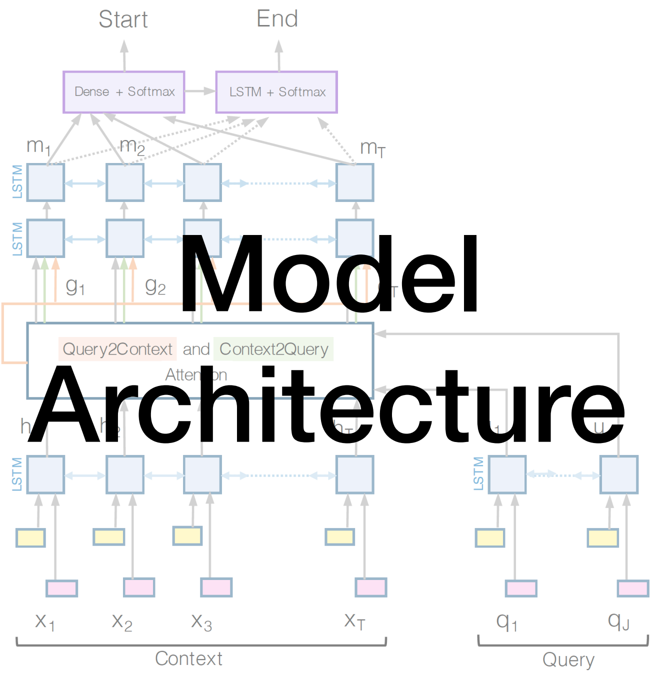
Noah A. Smith

NAACL 2019—June 4, 2019

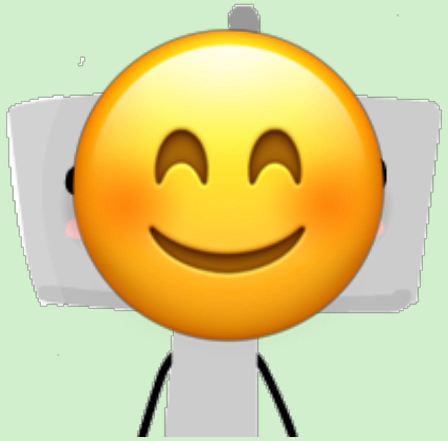
UWNLP AI2

Two Key Ingredients of NLP Systems

Training Dataset

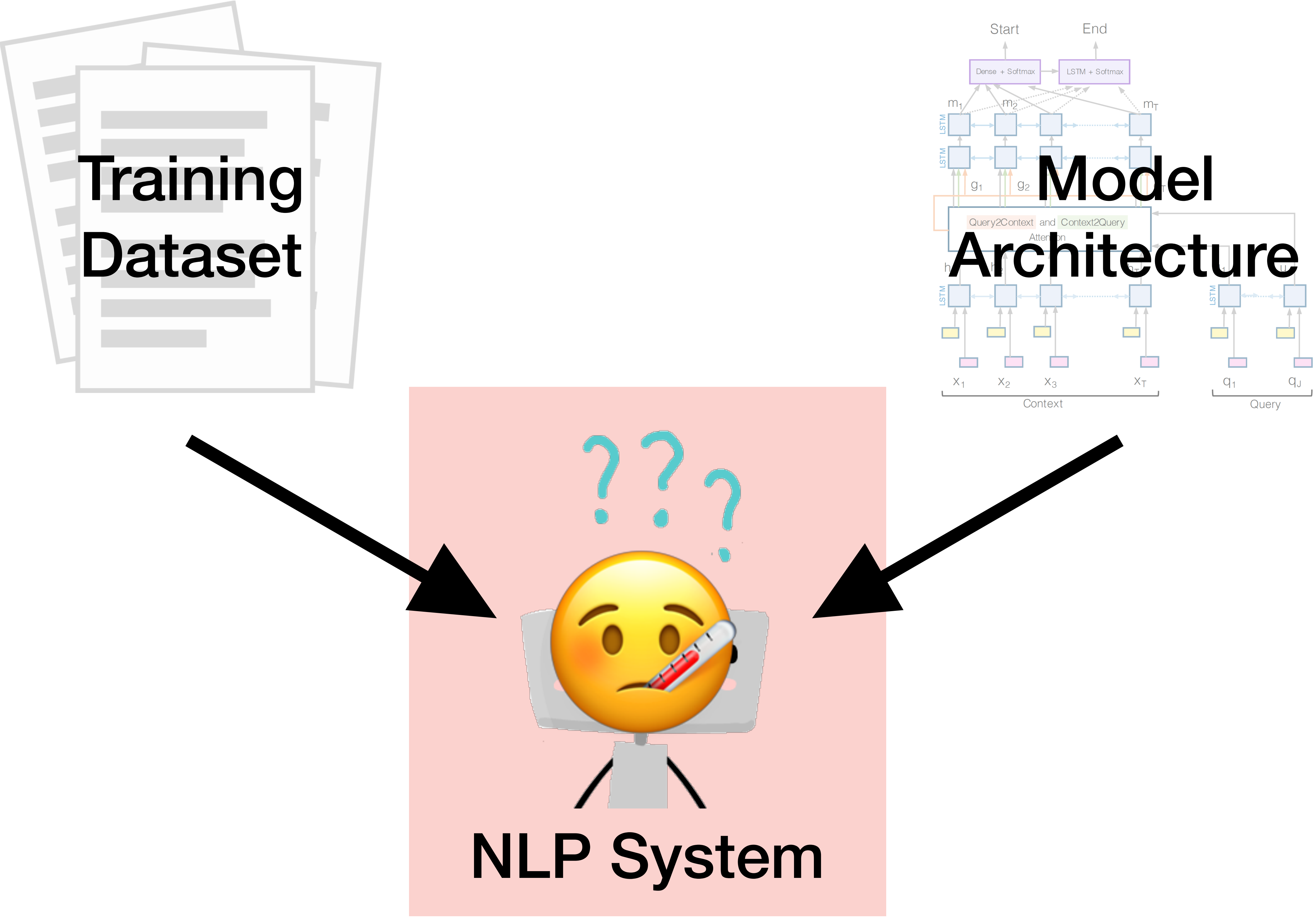


Model Architecture



NLP System

Why Might NLP Systems Fail?

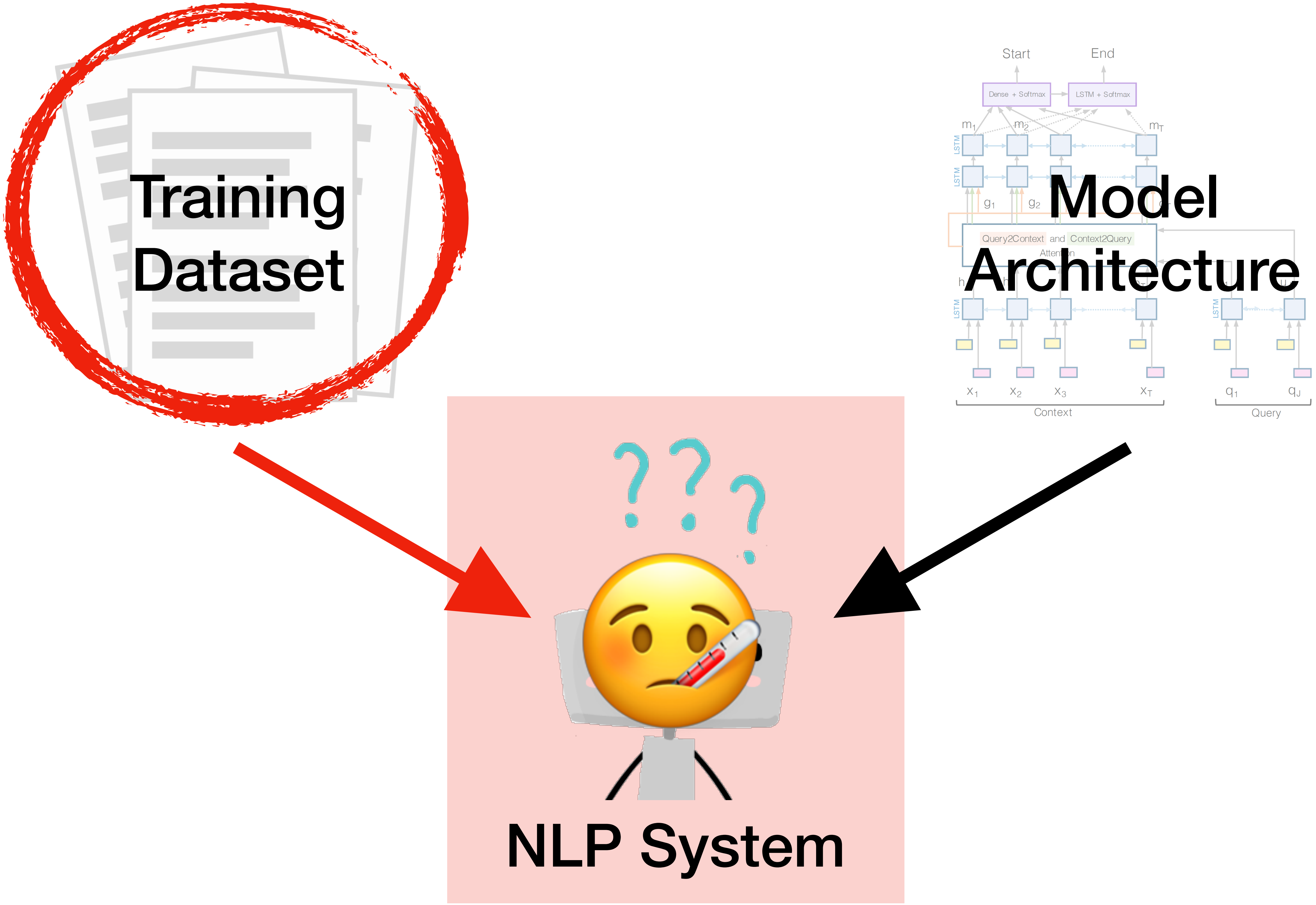


Training Dataset

Model Architecture

NLP System

Dataset Weaknesses

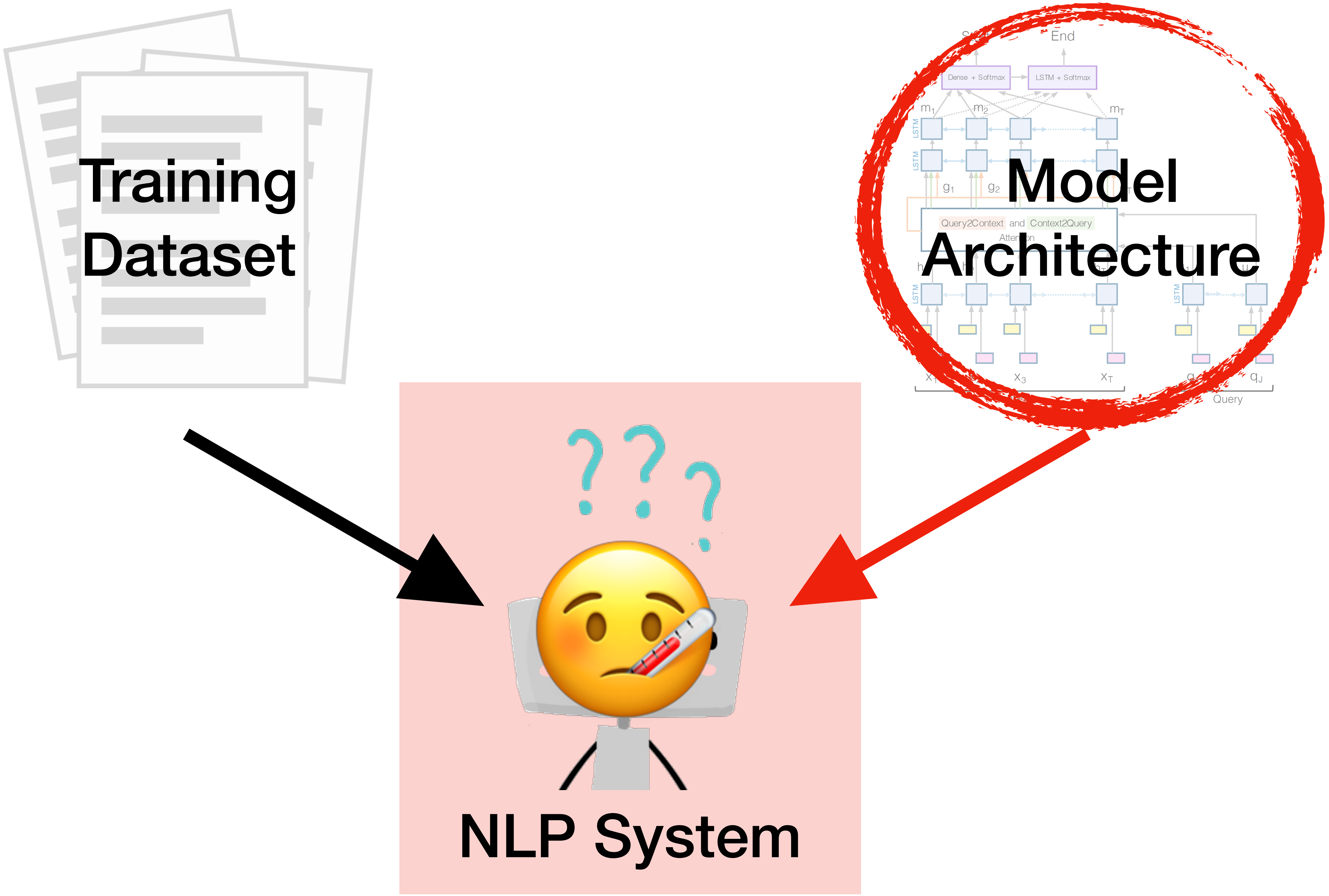


Training Dataset

Model Architecture

NLP System

Model Weaknesses



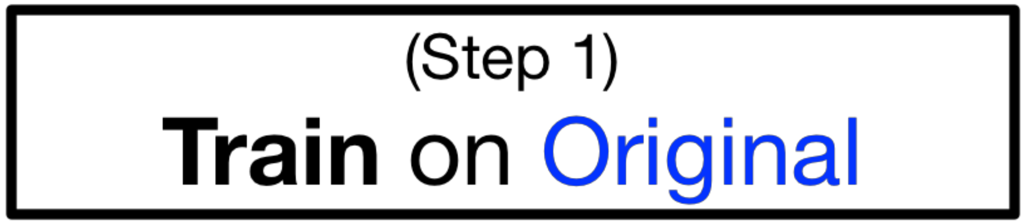
Training Dataset

Model Architecture

NLP System

Challenge Datasets Break Models

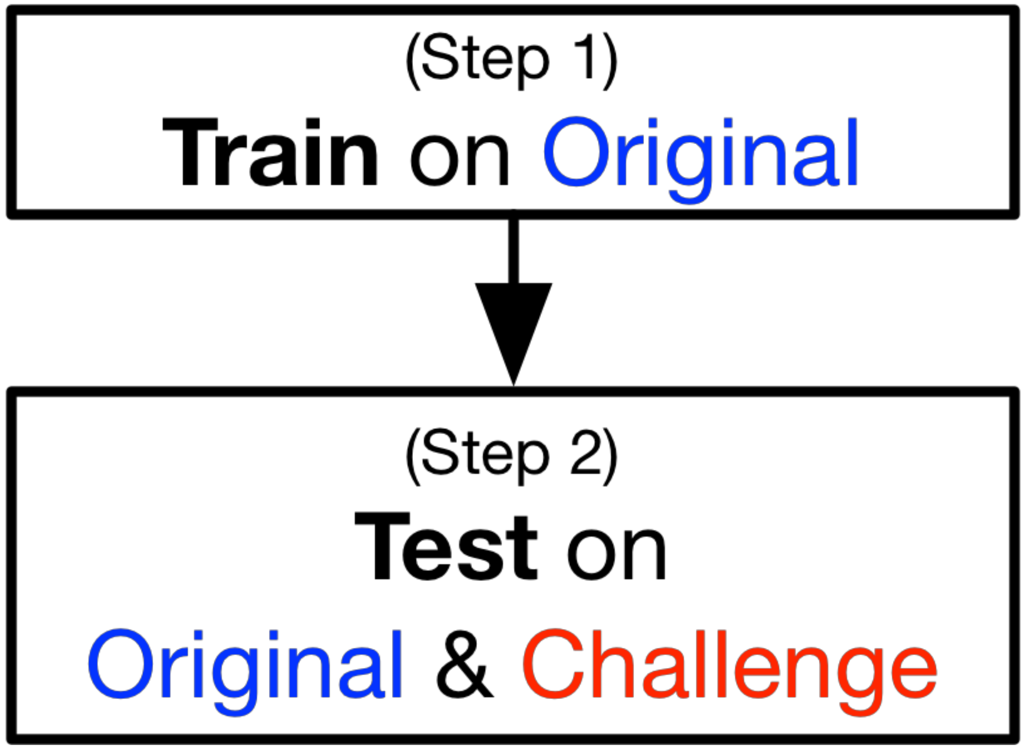
Challenge Datasets Break Models



(Step 1)

Train on Original

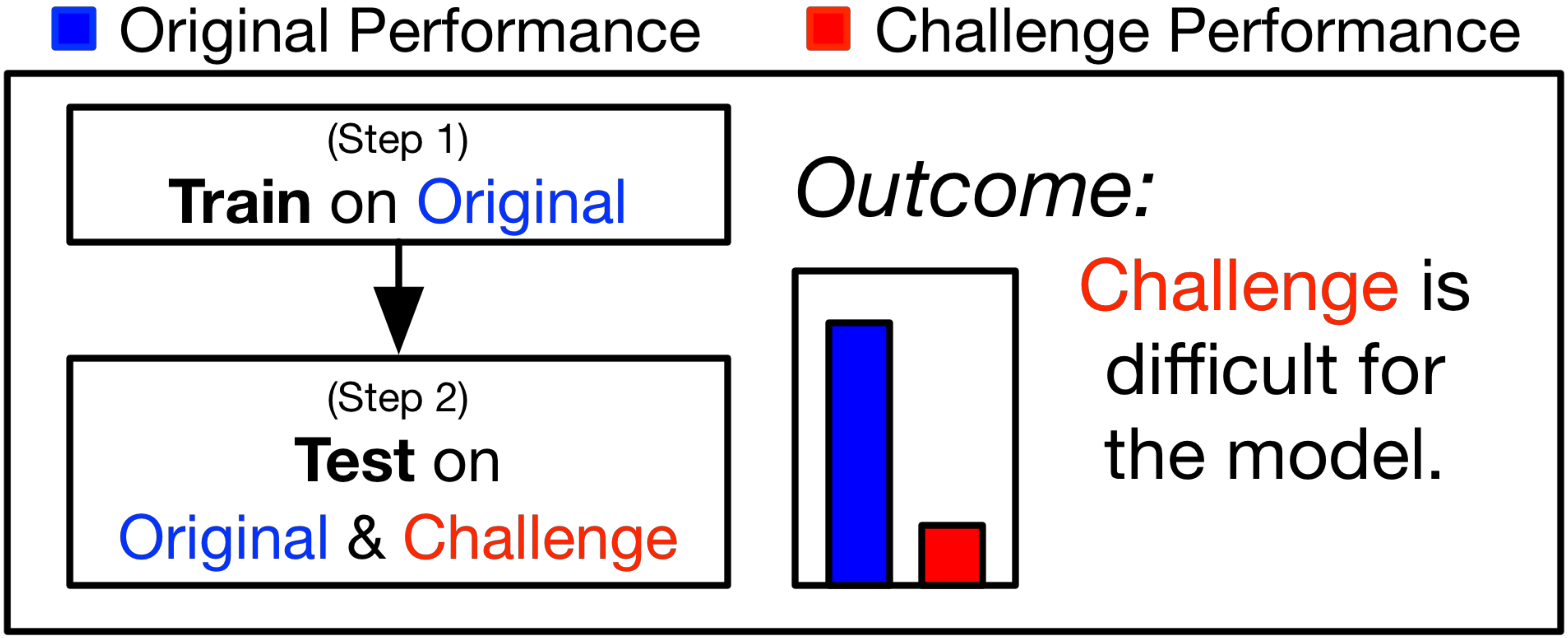
Challenge Datasets Break Models



(Step 1)
Train on Original

(Step 2)
Test on
Original & Challenge

NLP Systems Are Brittle



🟦 Original Performance 🟥 Challenge Performance

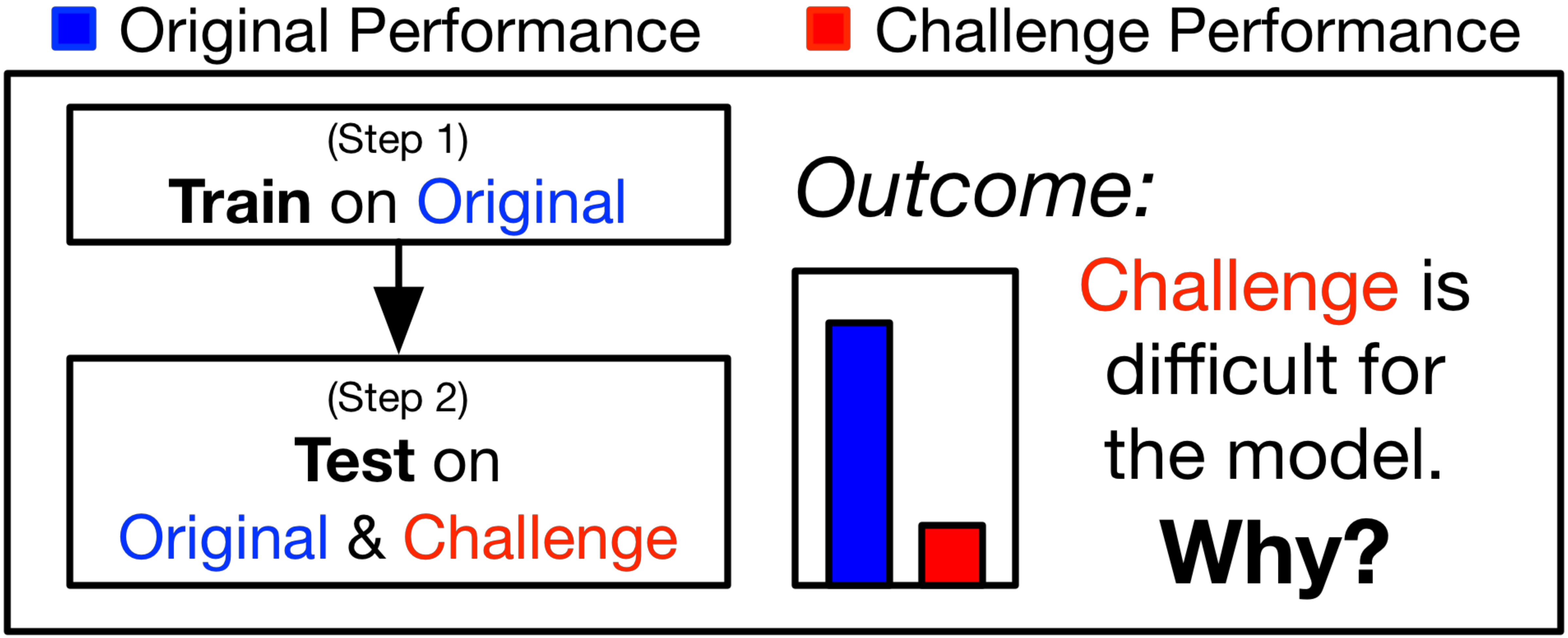
(Step 1)
Train on Original

(Step 2)
Test on
Original & Challenge

Outcome:

Challenge is difficult for the model.

NLP Systems Are Brittle



🟦 Original Performance 🟥 Challenge Performance

(Step 1)
Train on Original

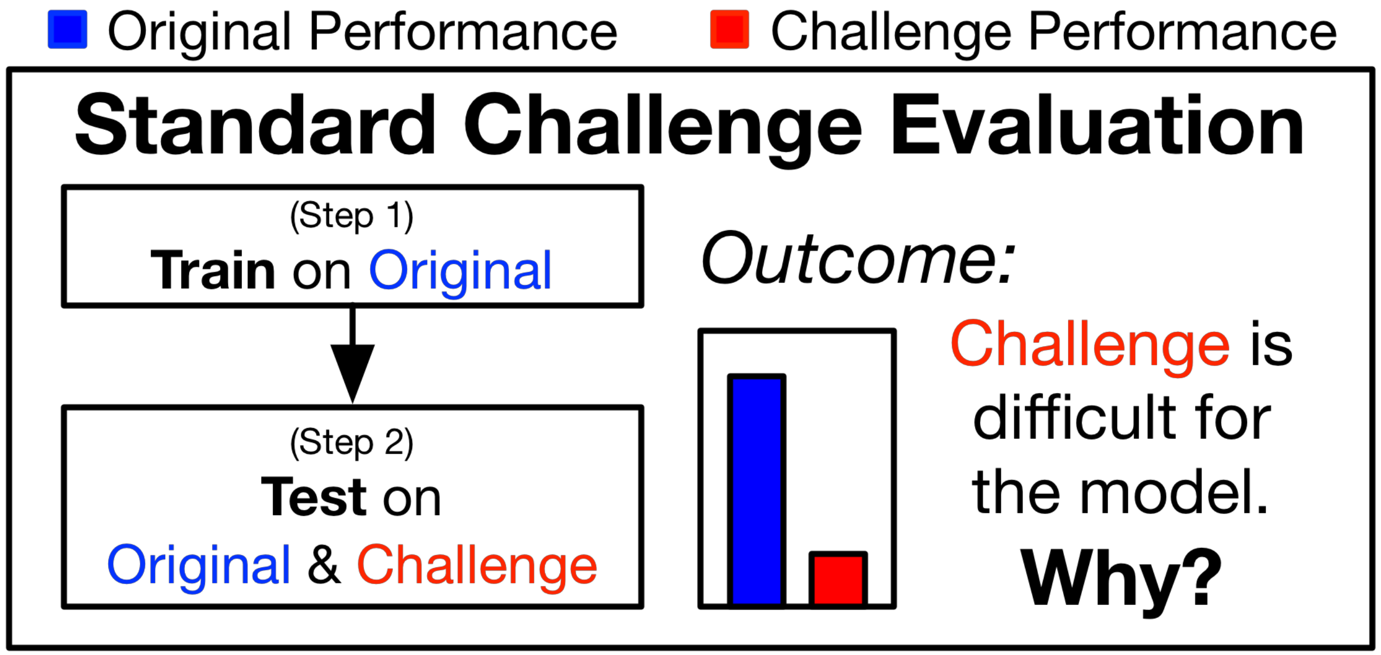
(Step 2)
Test on
Original & Challenge

Outcome:

Challenge is difficult for the model.

Why?

Inoculation by Fine-Tuning



🟦 Original Performance 🟥 Challenge Performance

Standard Challenge Evaluation

(Step 1)
Train on Original

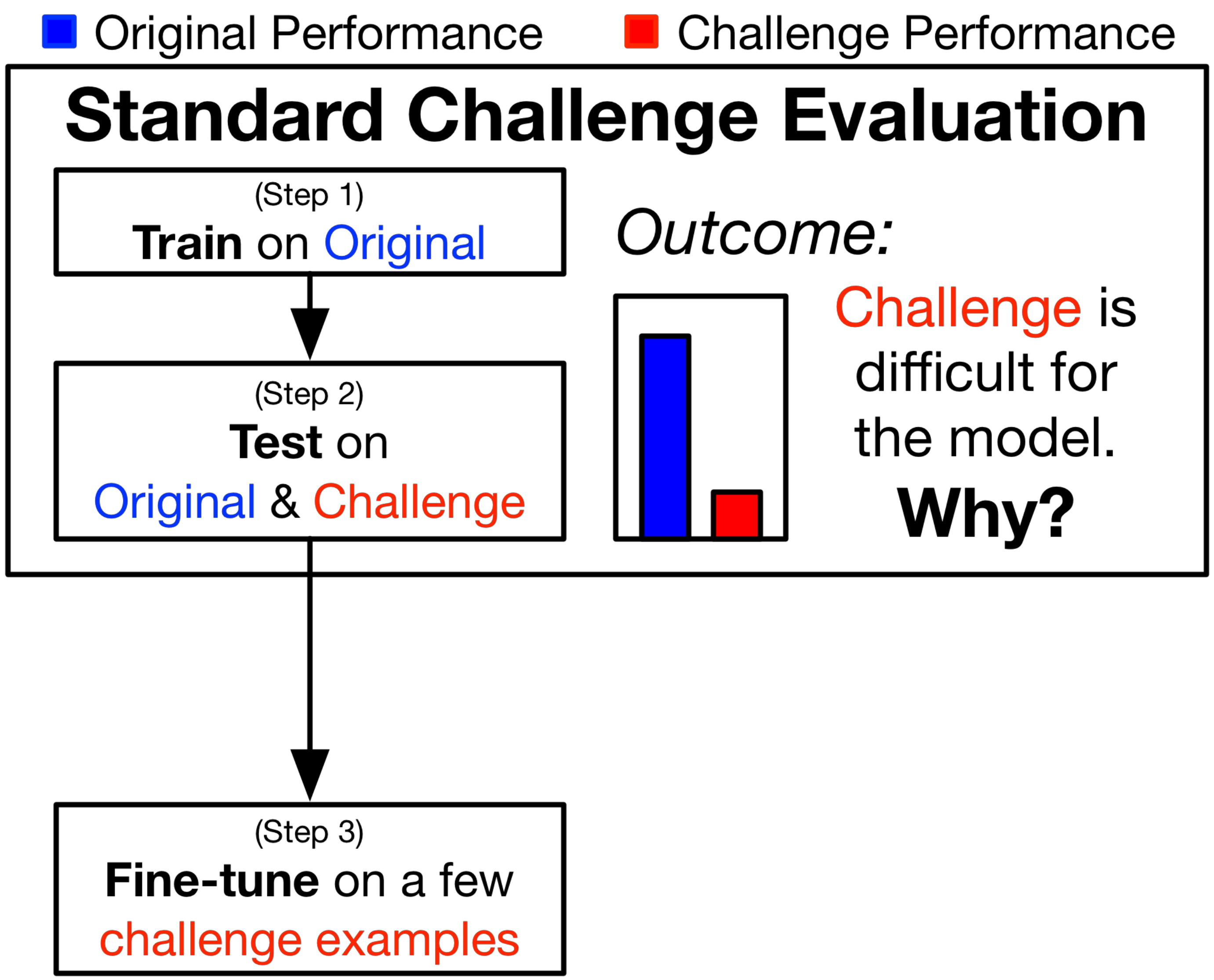
(Step 2)
Test on
Original & Challenge

Outcome:

Challenge is difficult for the model.

Why?

Inoculation by Fine-Tuning



🟦 Original Performance 🟥 Challenge Performance

Standard Challenge Evaluation

(Step 1)
Train on Original

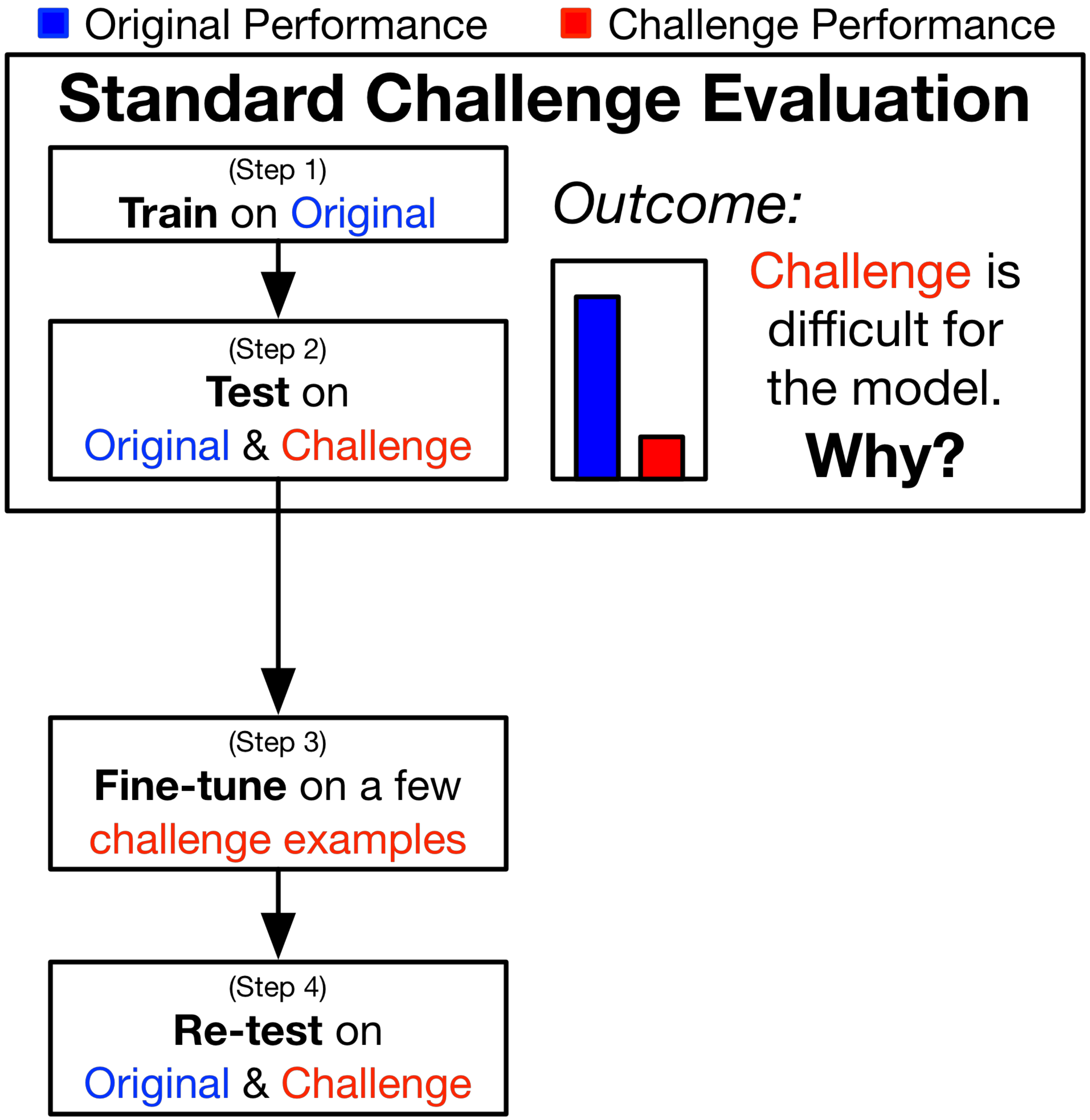
(Step 2)
Test on
Original & Challenge

Outcome:

Challenge is difficult for the model.
Why?

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

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.

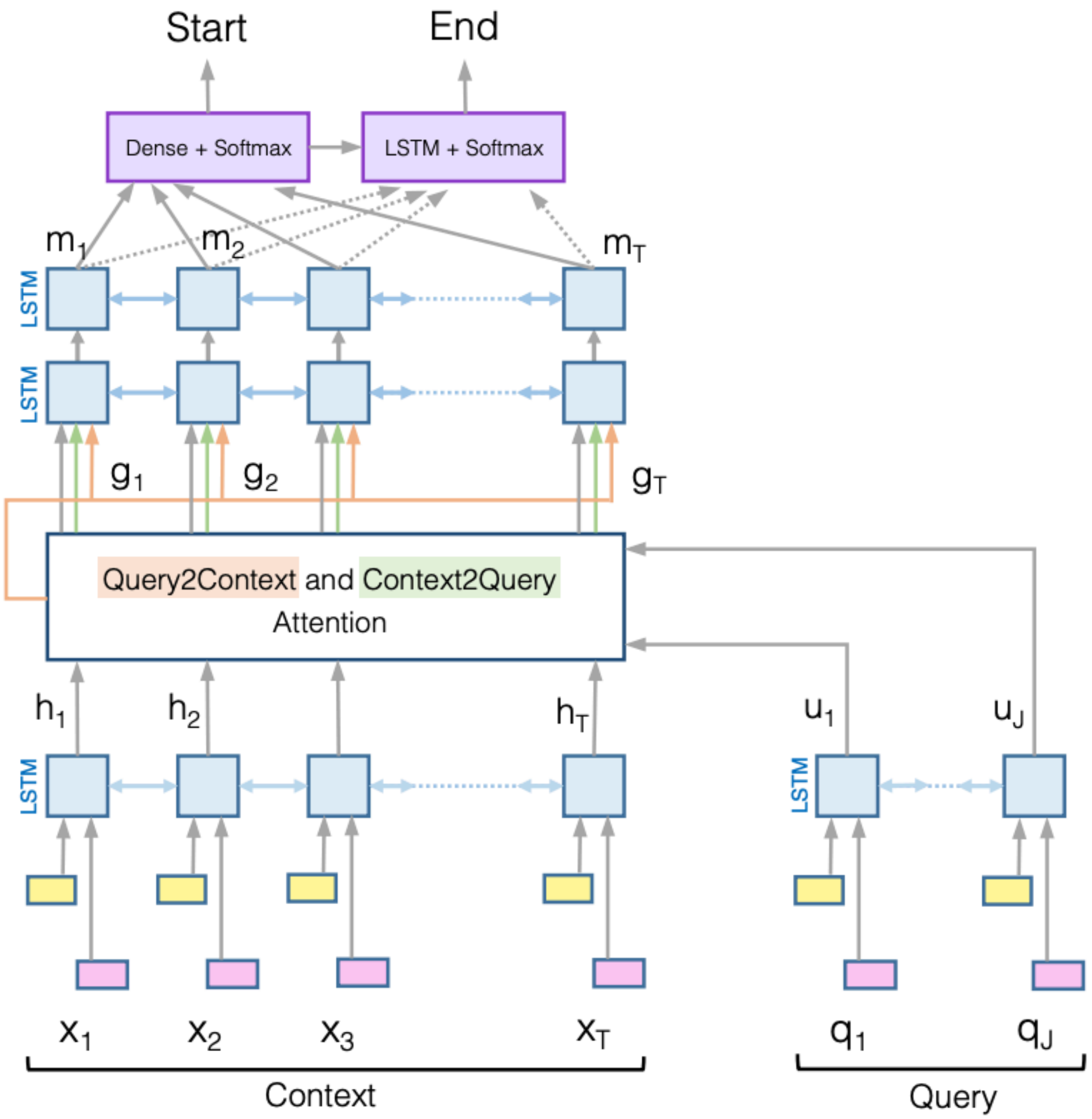
Why?

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

(Step 4)
Re-test on
Original & Challenge

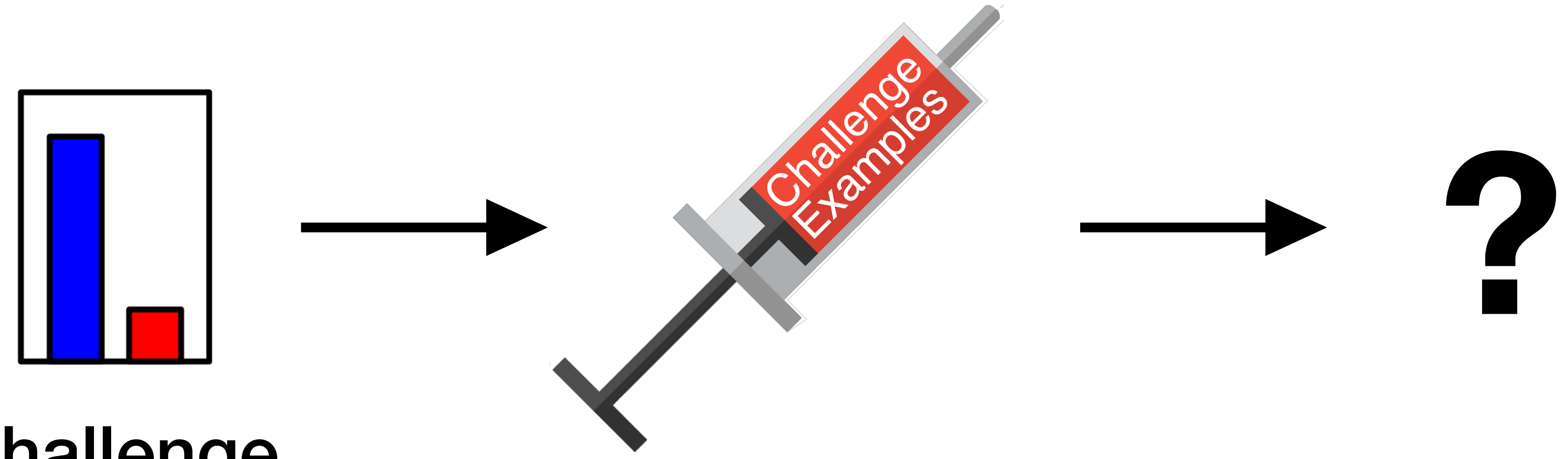
Inoculation

Inoculate Models to Better Understand Why They Fail



Three Clear Outcomes of Interest

🟦 Original Performance
🟥 Challenge Performance

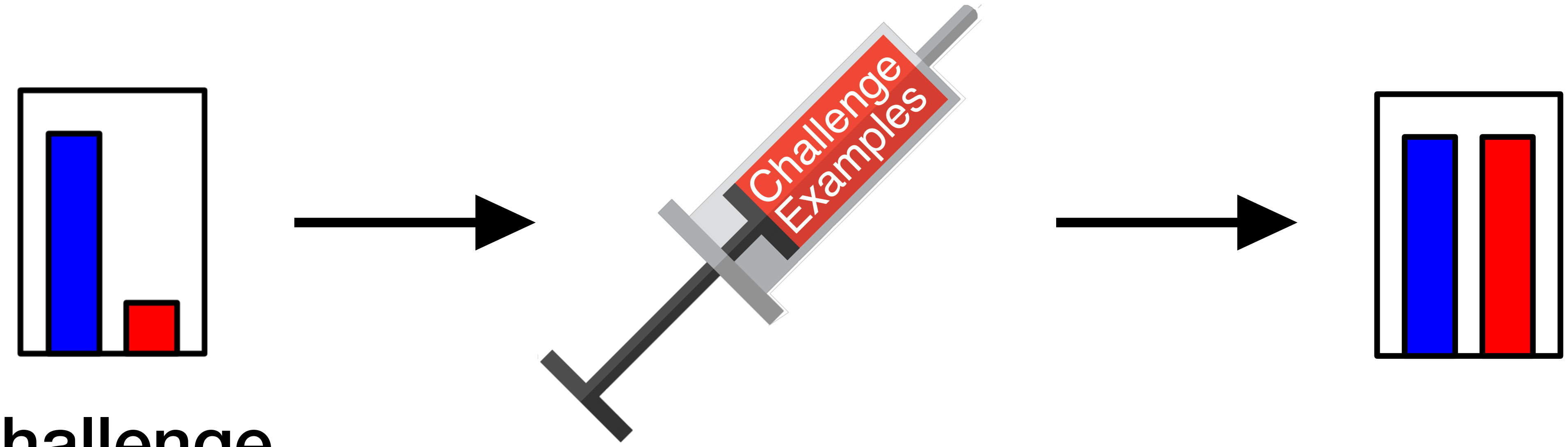


Challenge Evaluation Outcome

Inoculation

(1) Dataset Weakness

🟦 Original Performance
🟥 Challenge Performance

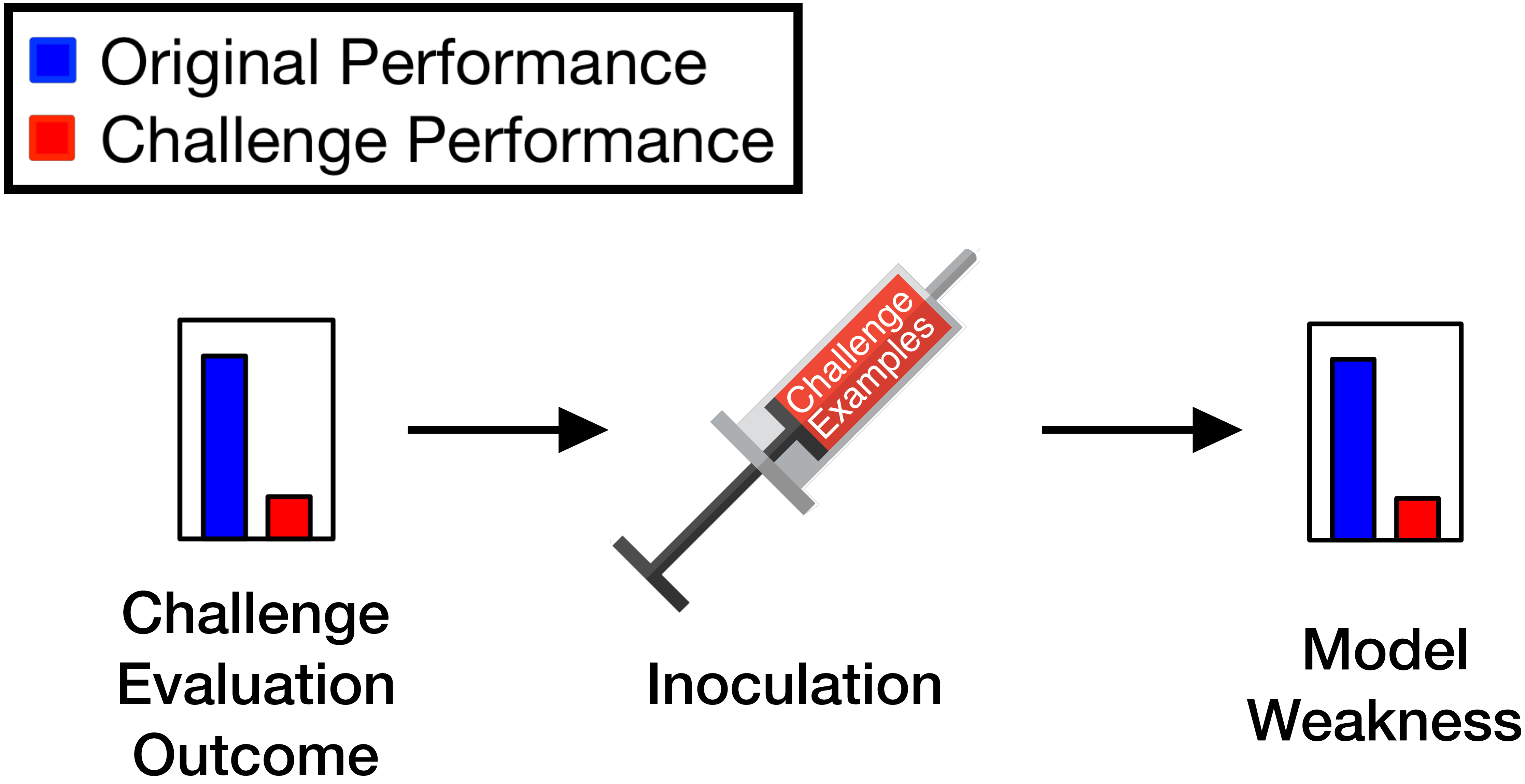


Challenge Evaluation Outcome

Inoculation

Dataset Weakness

(2) Model Weakness



🟦 Original Performance
🟥 Challenge Performance

Challenge Evaluation Outcome

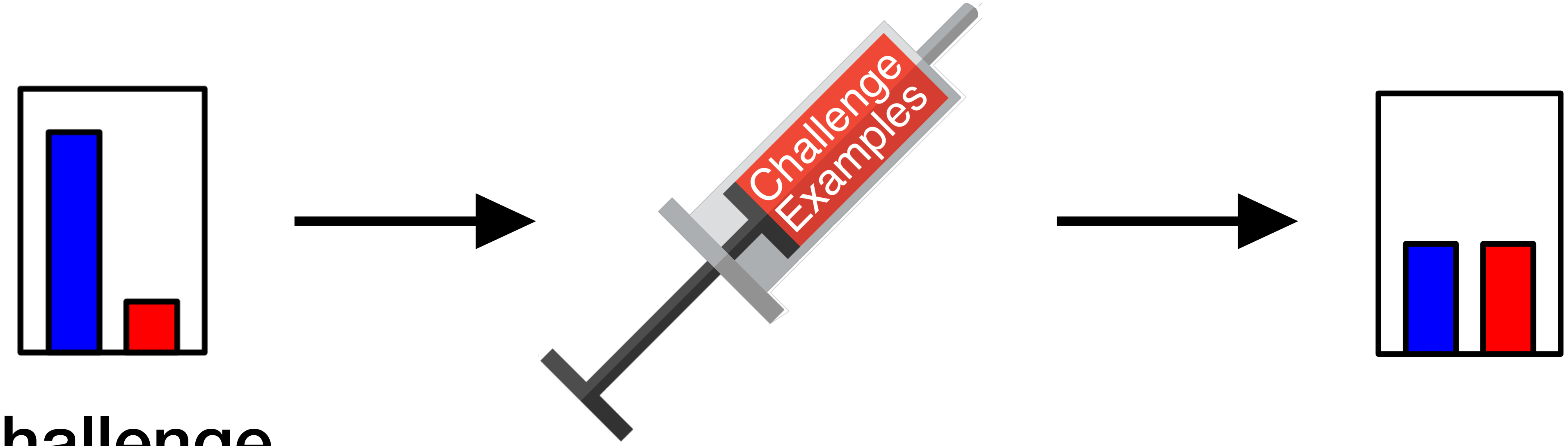
Inoculation

Model Weakness

(3) Predictive Artifacts / Other

🟦 Original Performance

🟥 Challenge Performance

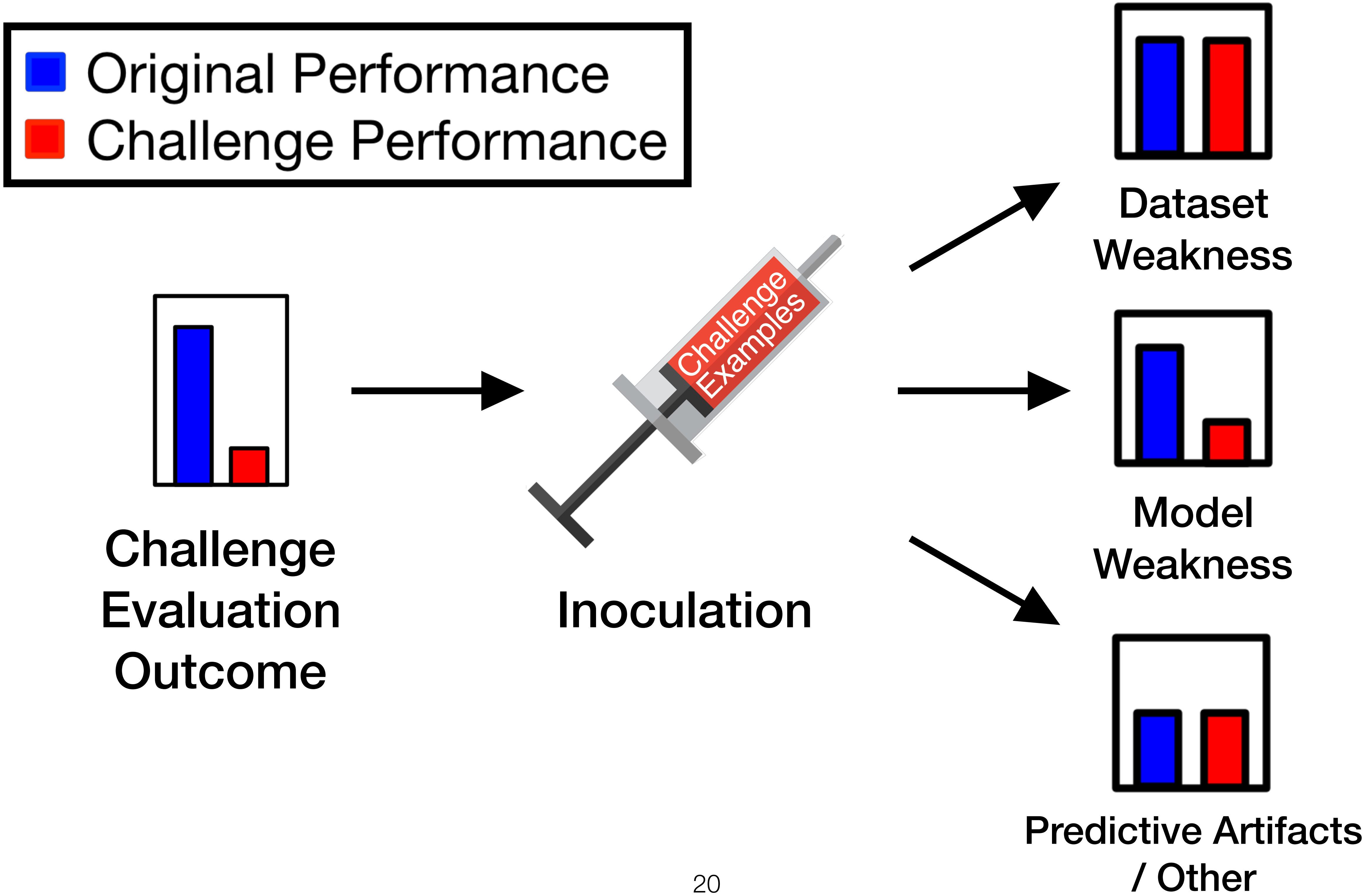


Challenge Evaluation Outcome

Inoculation

Predictive Artifacts / Other

Three Clear Outcomes of Interest



🟦 Original Performance

🟥 Challenge Performance

Challenge Evaluation Outcome

Inoculation

Dataset Weakness

Model Weakness

Predictive Artifacts / Other

Case Studies

- Inoculating natural language inference (NLI) models
- Inoculating 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

[Naik and Ravichander et al., 2018]

Two NLI Challenge Datasets

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

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

[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.***"

[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.*"

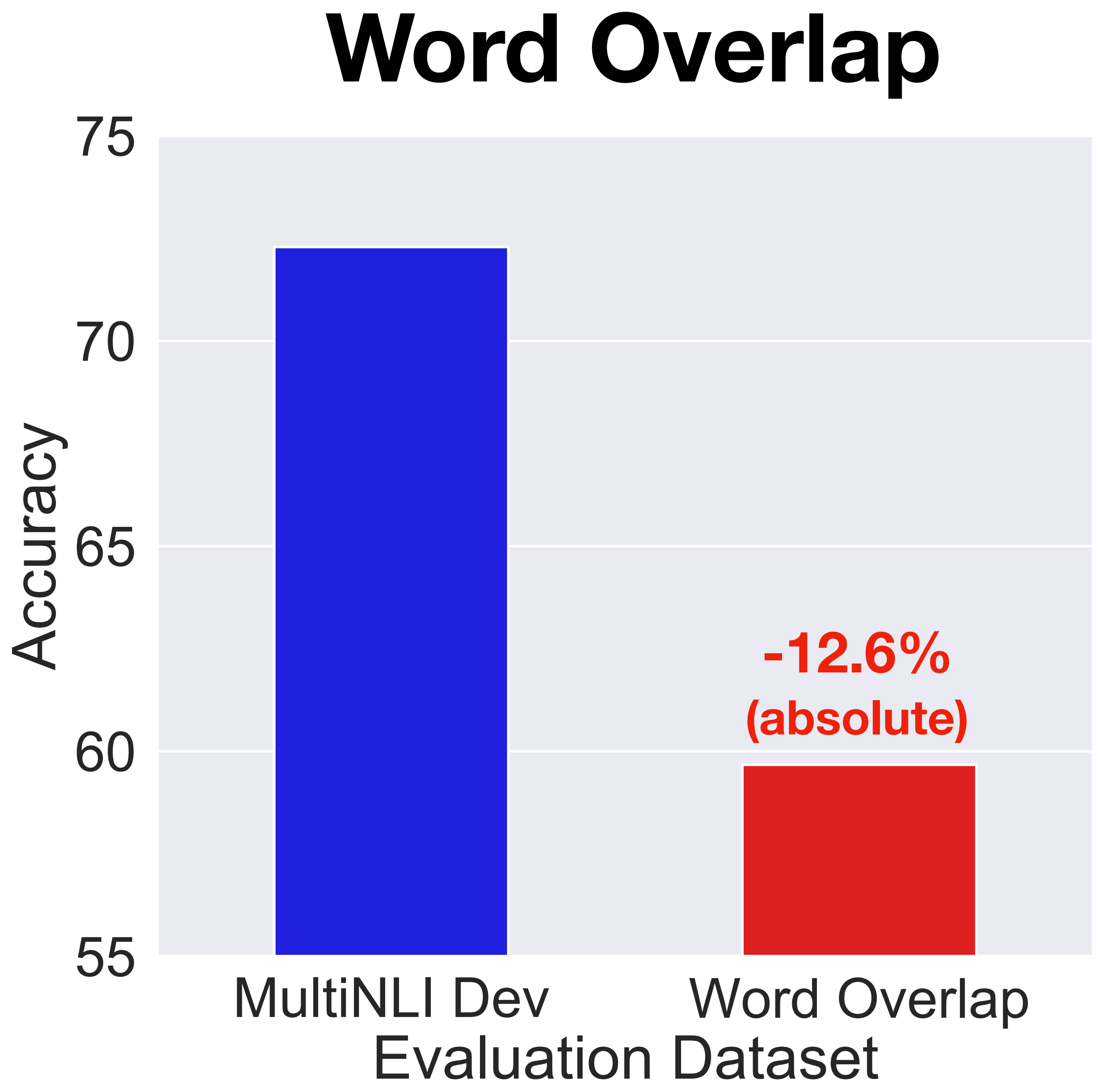
Hypothesis: "*I have disobeyed your orders* ***and true is true*****.**"

Spelling Errors Challenge Dataset

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

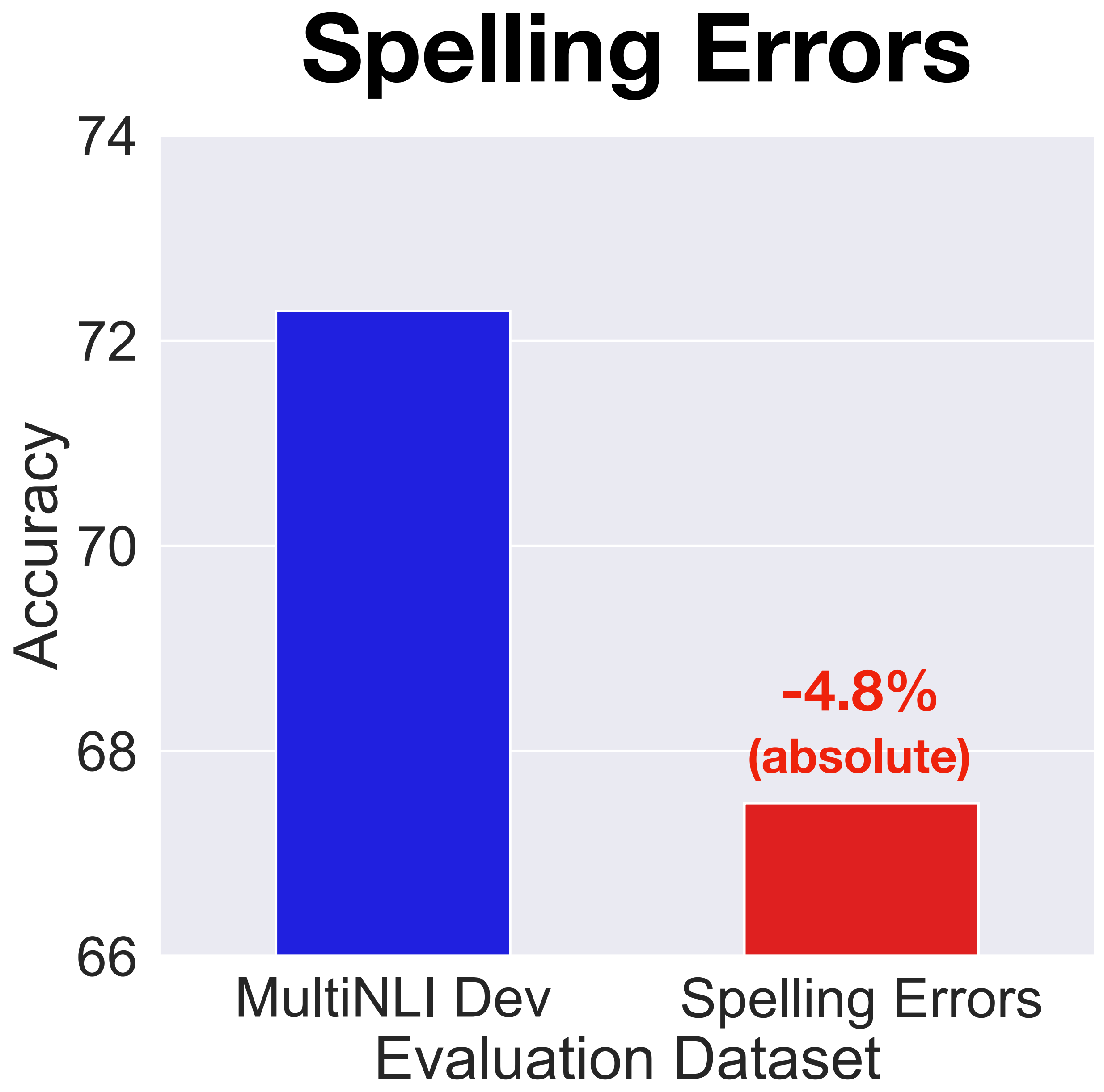
Hypothesis: "*I have disobeyed your* ***ordets*****.**"

Small Perturbations Break NLI Models



Word Overlap

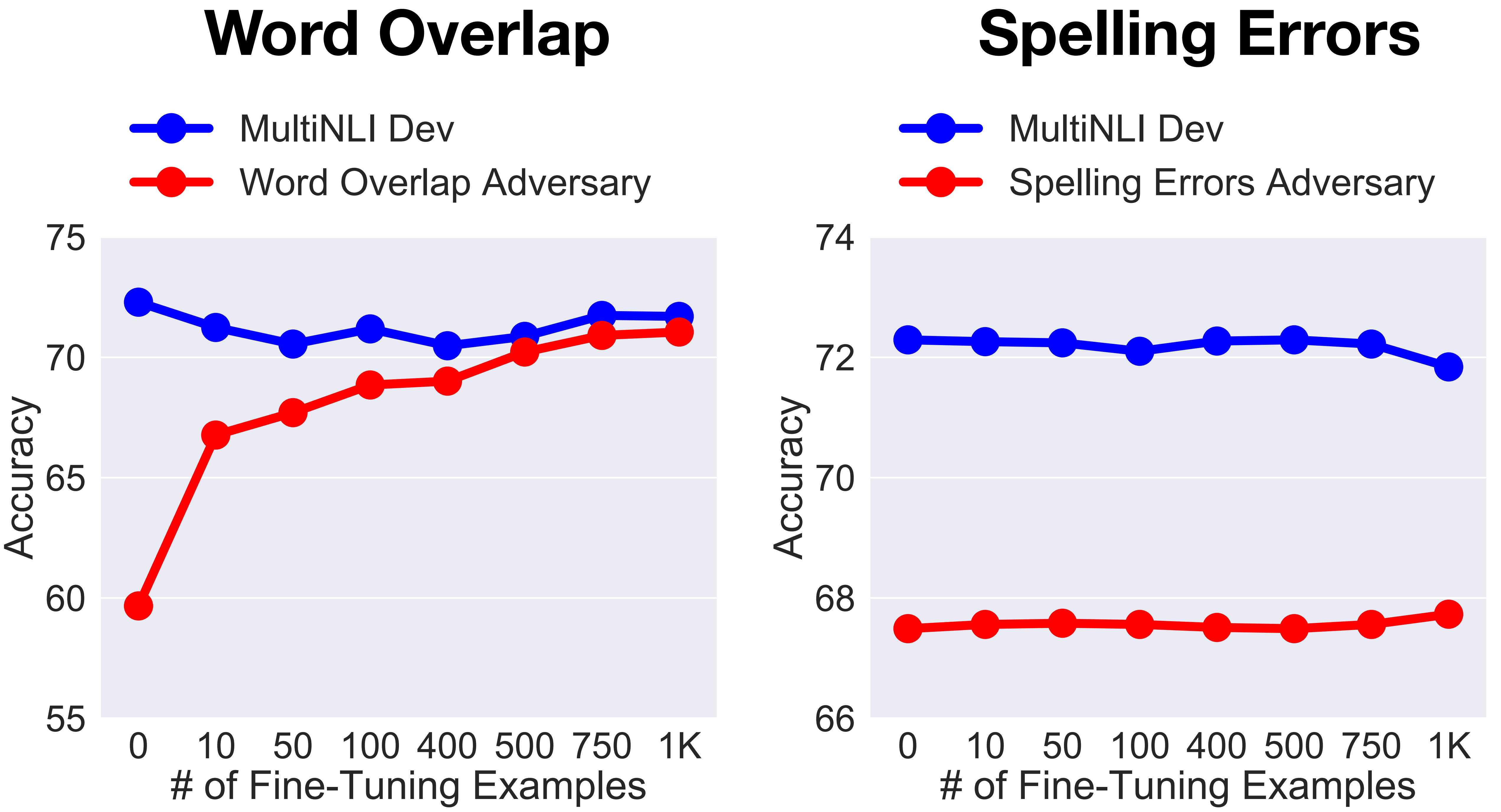
-12.6% (absolute)



Spelling Errors

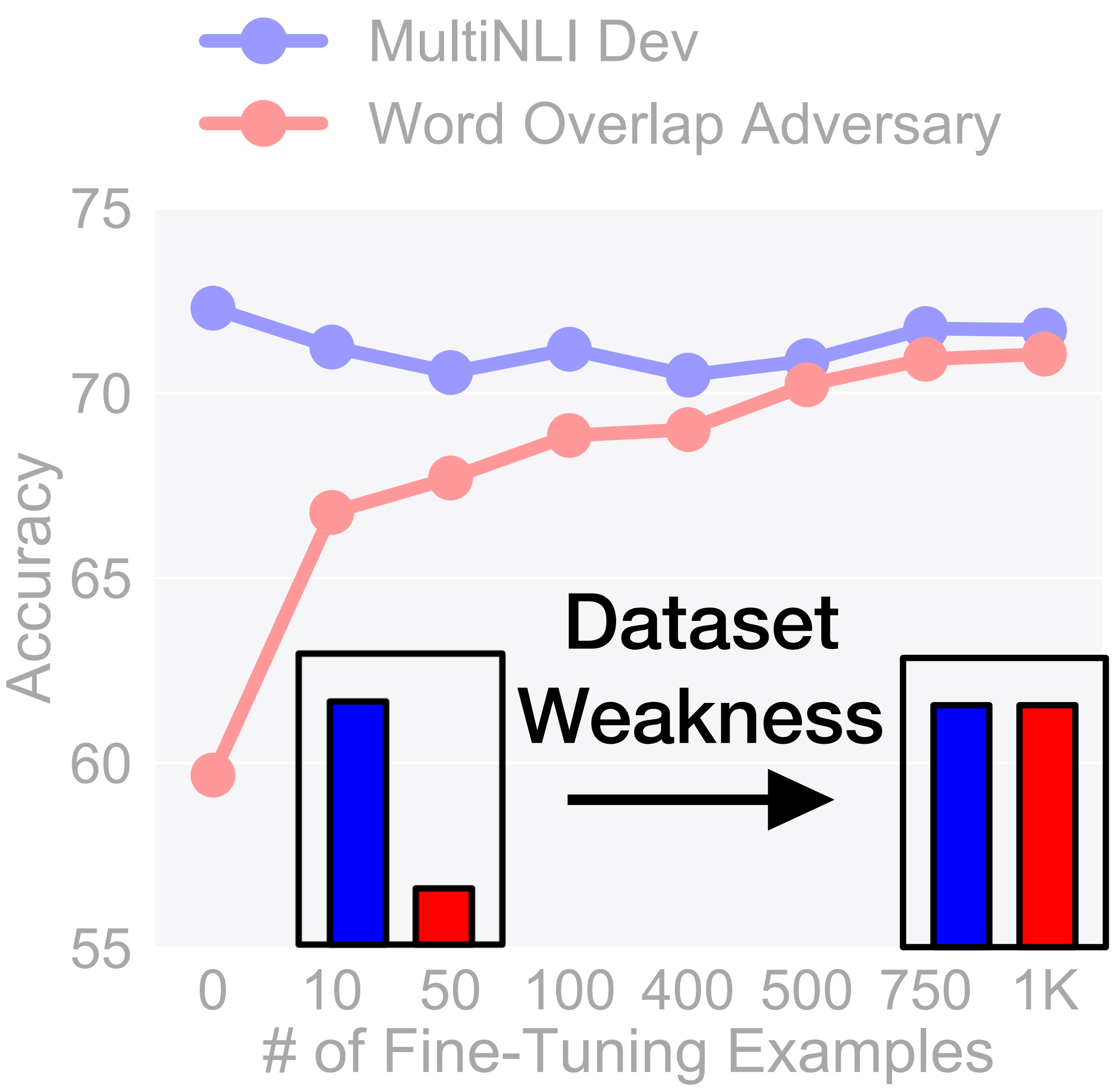
-4.8% (absolute)

Inoculating NLI models

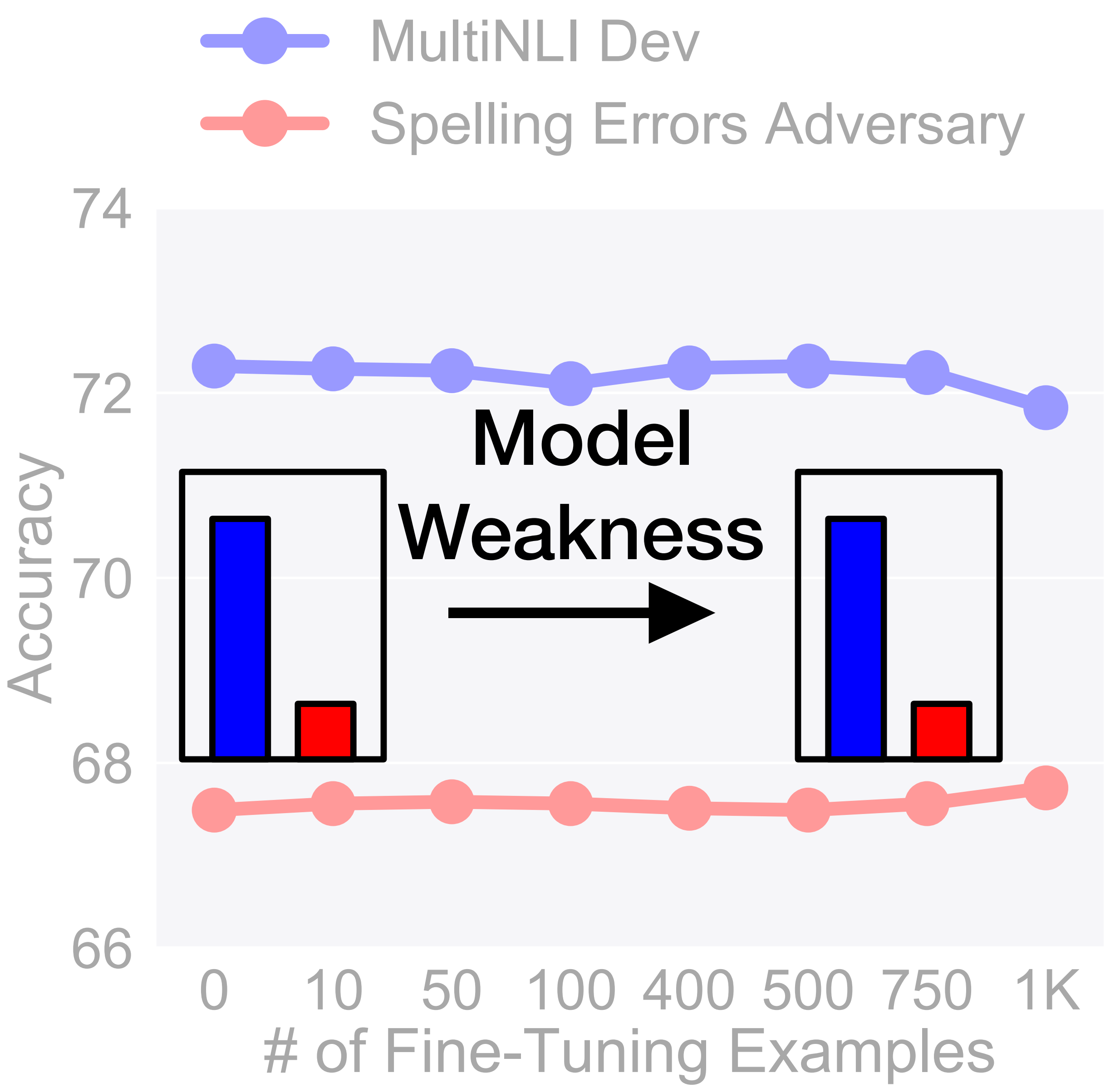


Inoculating NLI models

Word Overlap



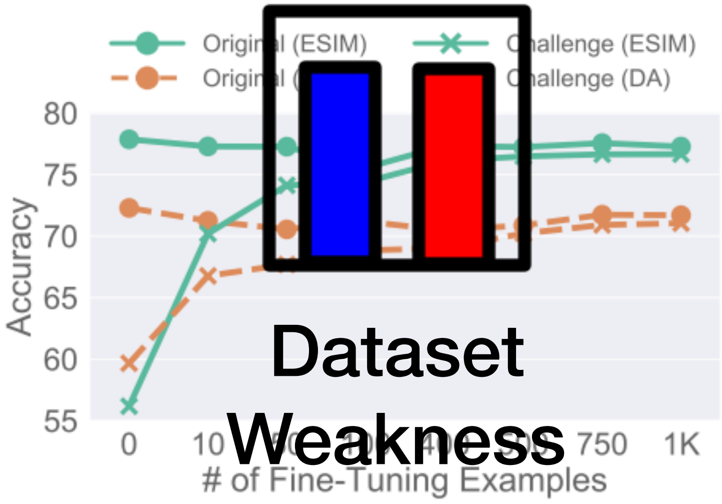
Spelling Errors



More Examples in the Paper!

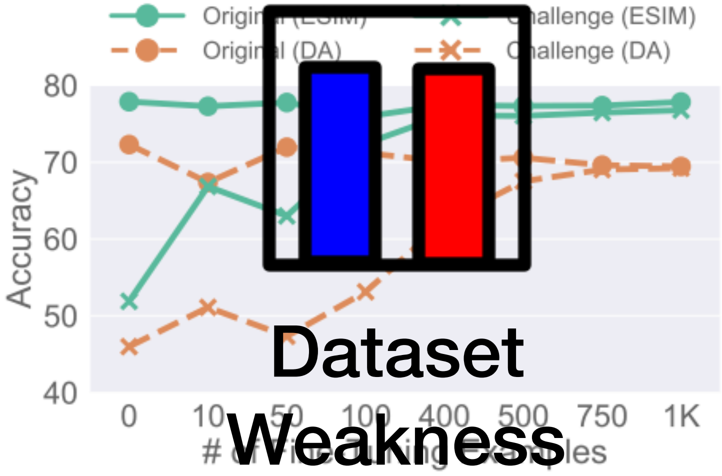
Outcome 1

(a) Word Overlap



Dataset Weakness

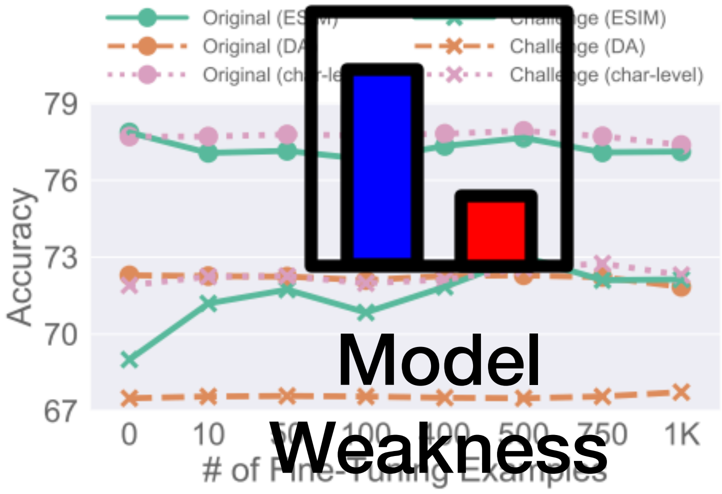
(b) Negation



Dataset Weakness

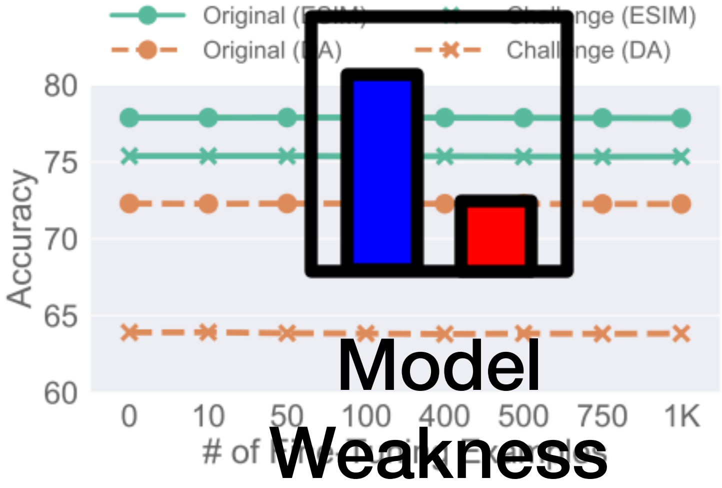
Outcome 2

(c) Spelling Errors



Model Weakness

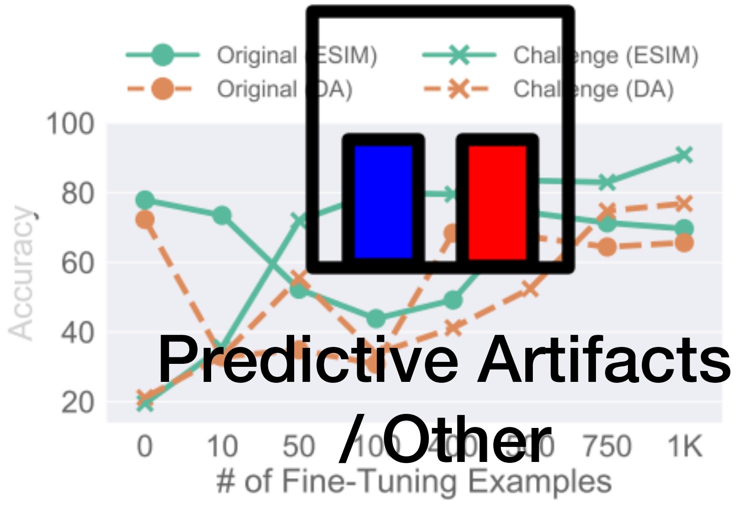
(d) Length Mismatch



Model Weakness

Outcome 3

(e) Numerical Reasoning



Predictive Artifacts / Other

[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***"

[Jia and Liang, 2017]
Example from Robin Jia

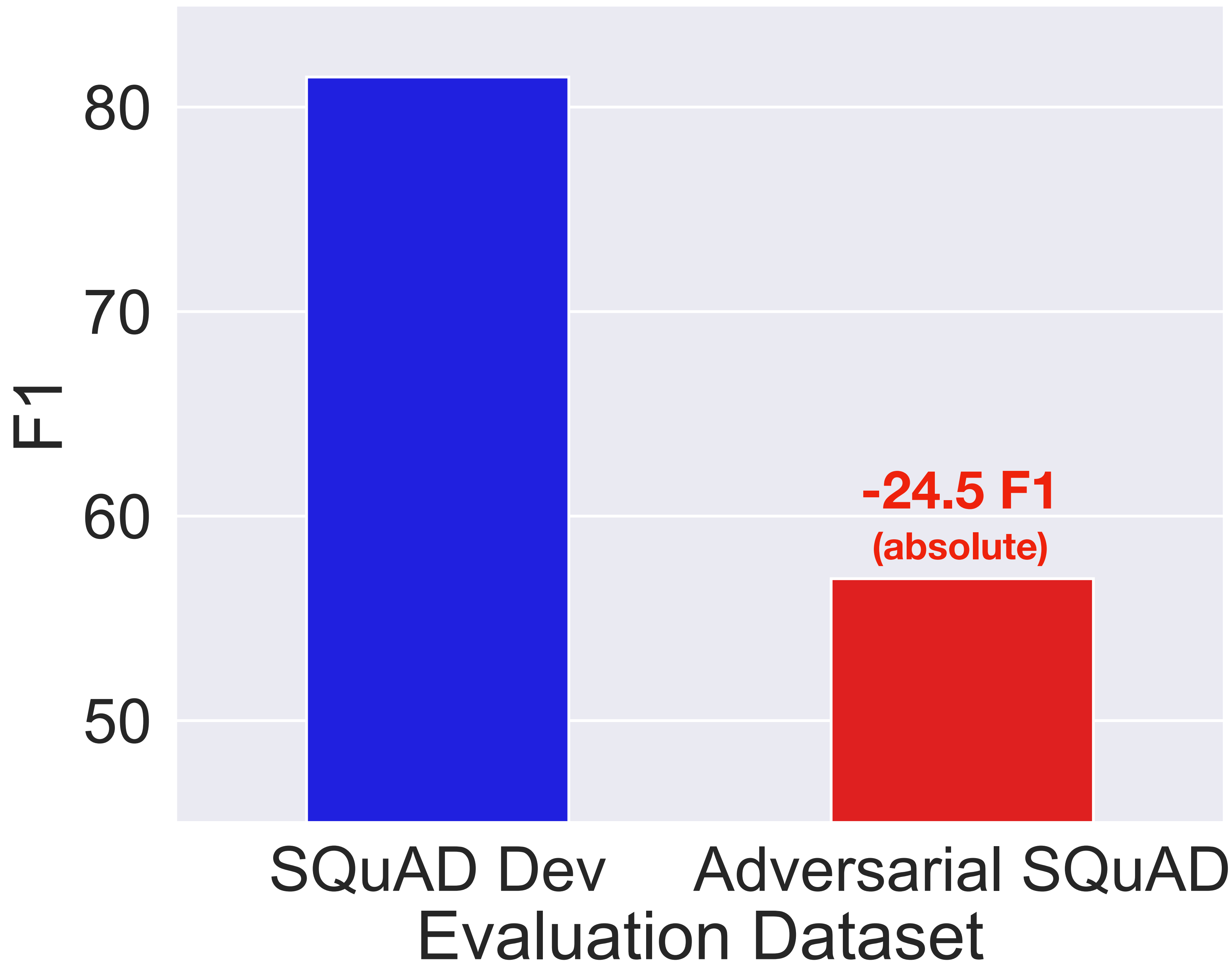
Adversarial SQuAD

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

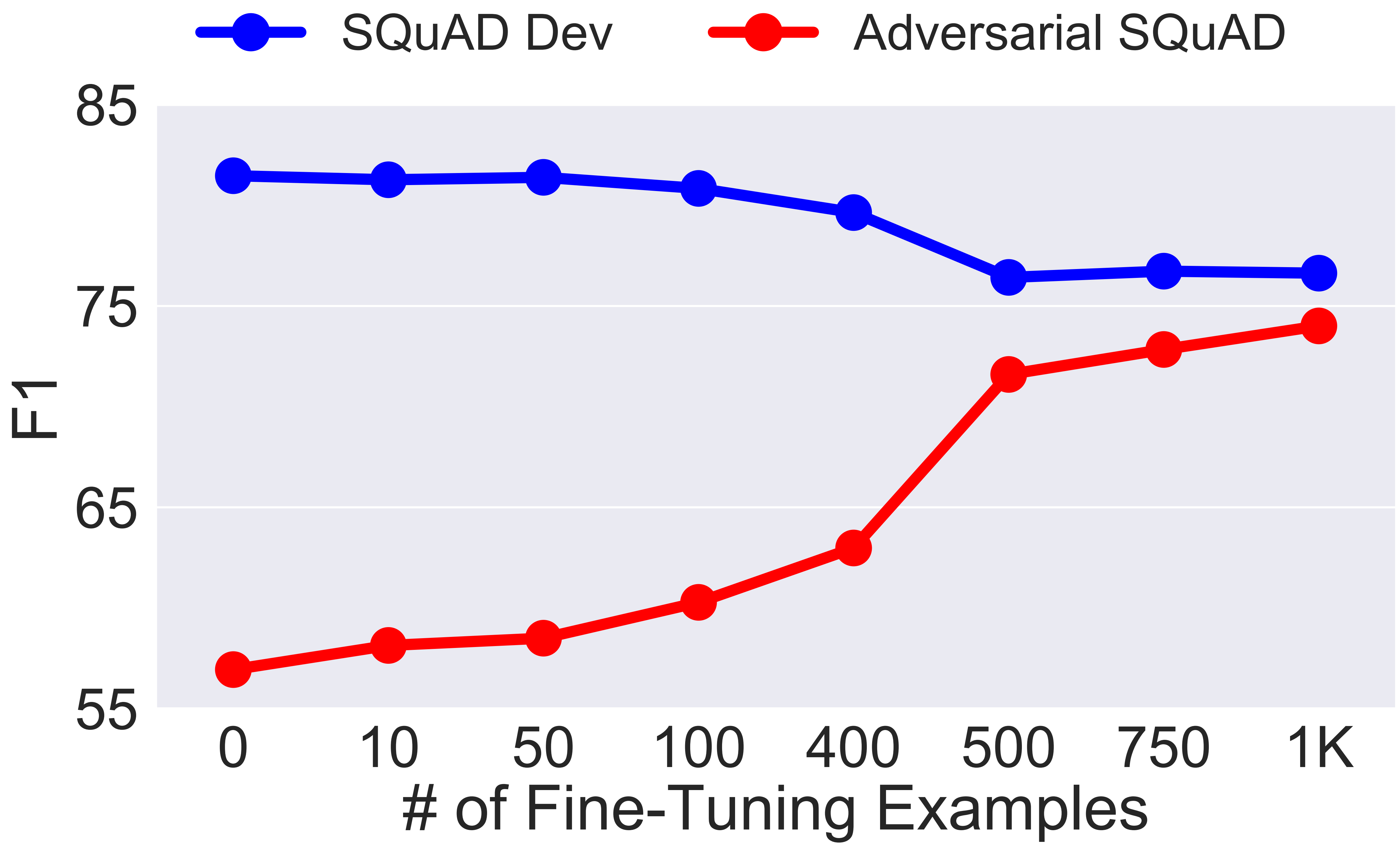
Passage: "*The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as* ***1700****; thereafter, the numbers declined. The number of old Acadian colonists declined after the year of* ***1675****.*"

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

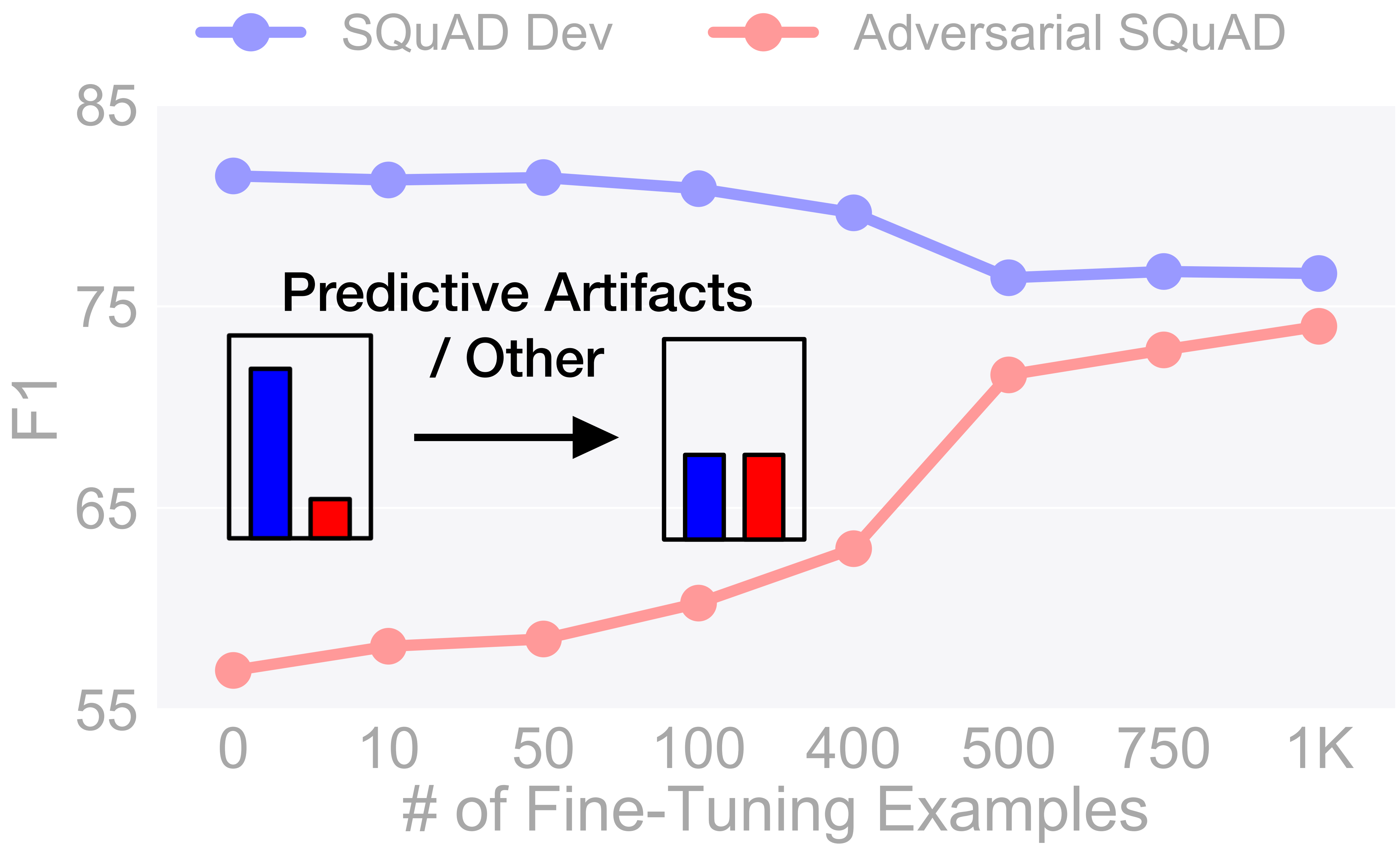
Small Perturbations Break SQuAD Models



Inoculating SQuAD models



Inoculating SQuAD models

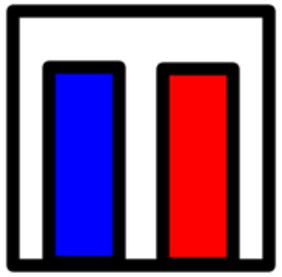


Predictive Artifacts / Other

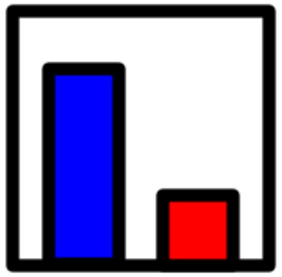
Takeaways

- Inoculation by Fine-Tuning helps us **understand why our models fail**.

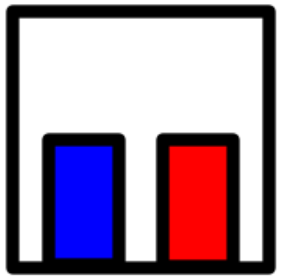
- While all challenge datasets break our models, **they stress them in different ways**.



Dataset Weakness



Model Weakness



Predictive Artifacts / Other

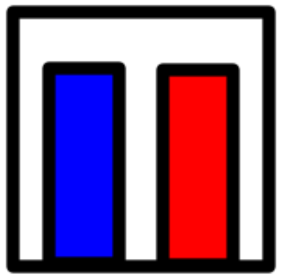
- Potentially many situations where inoculation can help clarify model results when transferring to other datasets.

Takeaways

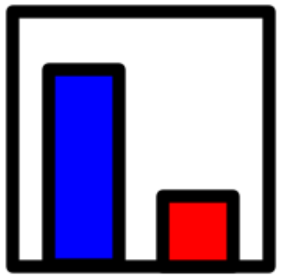
Thank You!
Questions?

- Inoculation by Fine-Tuning helps us **understand why our models fail**.

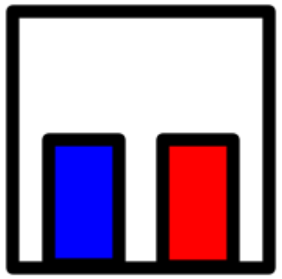
- While all challenge datasets break our models, **they stress them in different ways**.



Dataset
Weakness



Model
Weakness

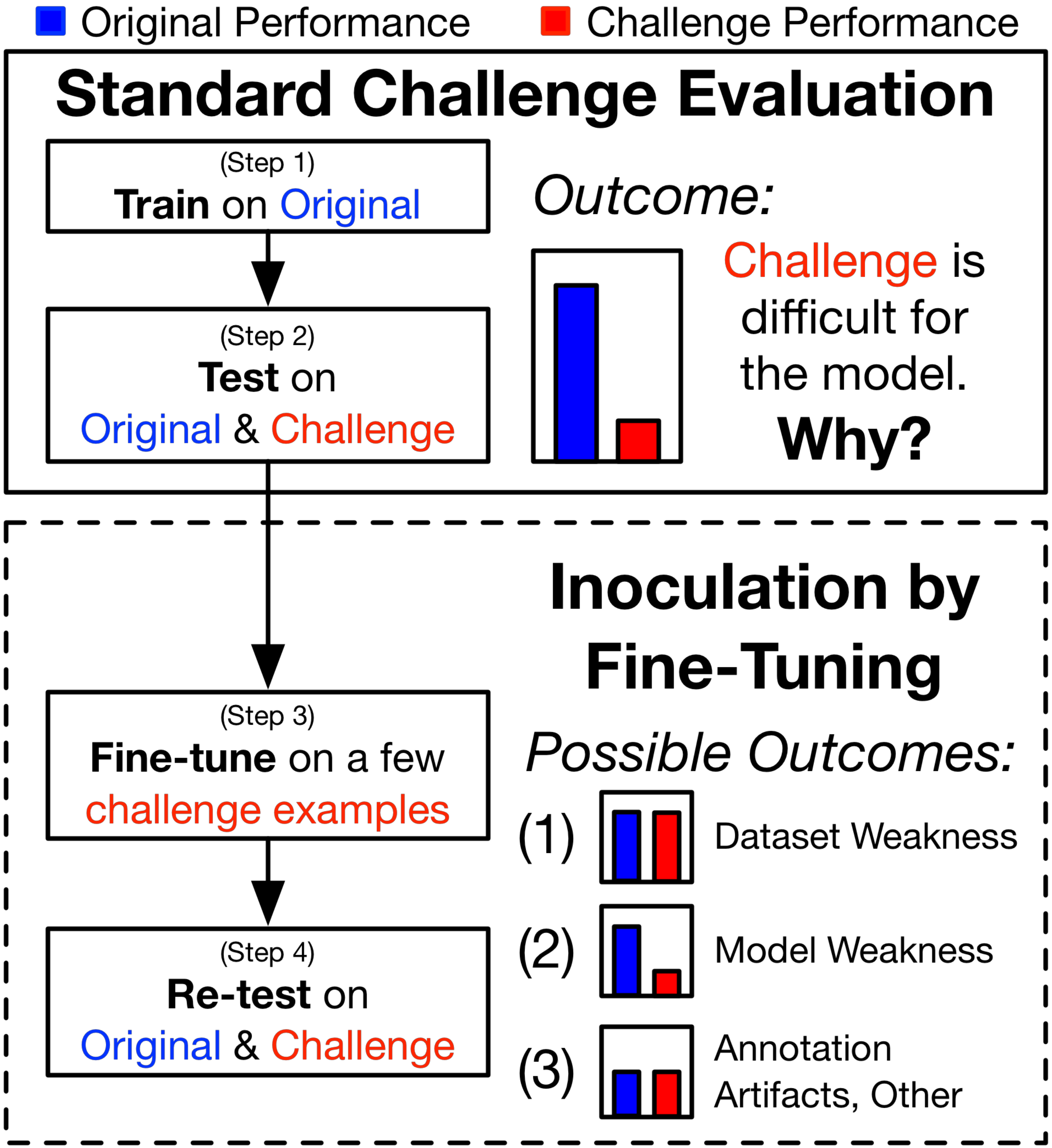


Predictive Artifacts
/ Other

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

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.



🟦 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

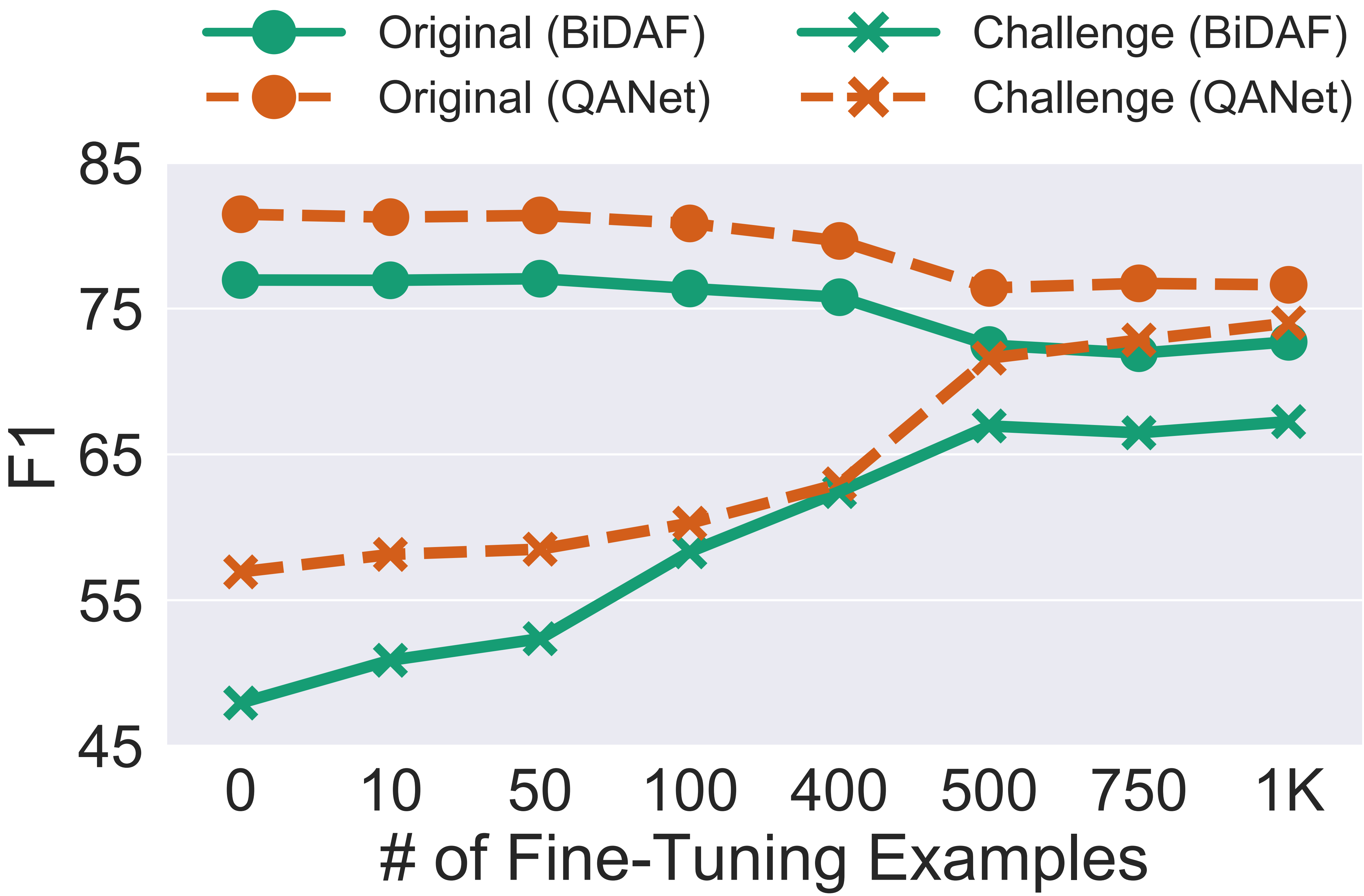
Possible Outcomes:

(1) Dataset Weakness

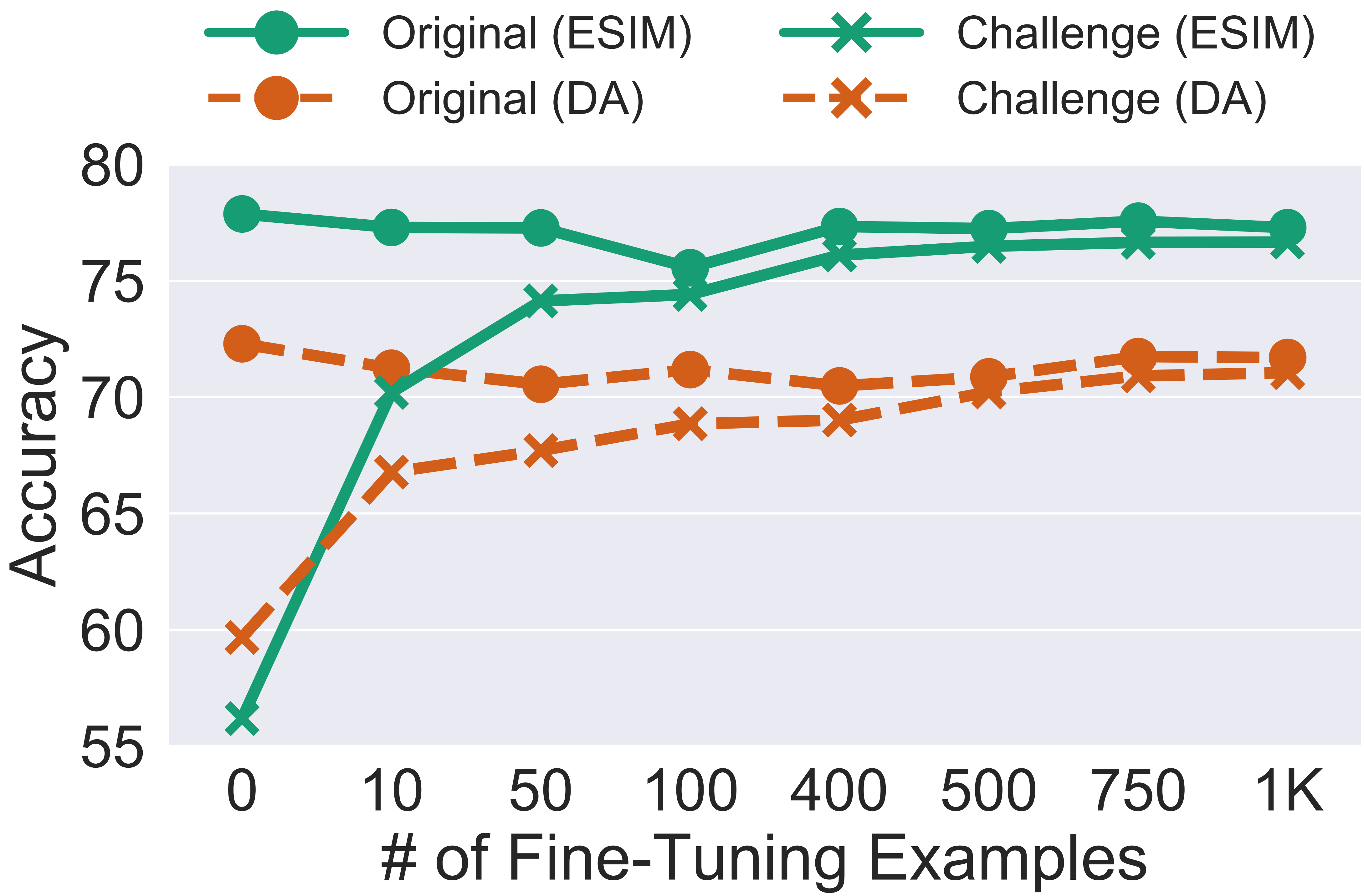
(2) Model Weakness

(3) Annotation Artifacts, Other

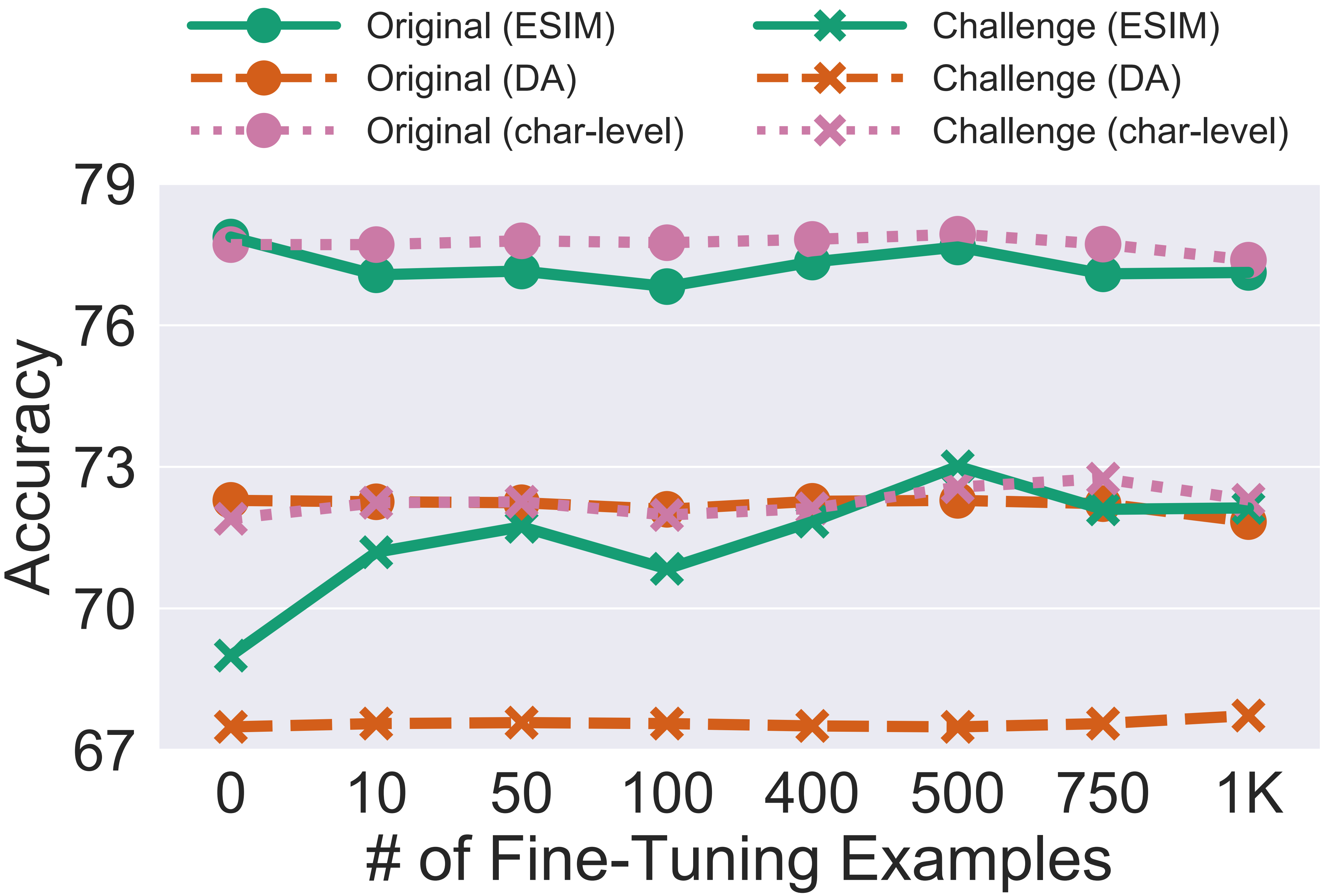
Inoculating Multiple SQuAD Reading Comprehension Models



Inoculating Multiple NLI Models Against Word Overlap Adversary



Inoculating Multiple NLI Models Against Spelling Errors


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