Adversarial NLI: A New Benchmark for Natural Language Understanding

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, Douwe Kiela

UNC Chapel Hill & Facebook AI Research

FACEBOOK AI

UNC NLP

Development of AI has been driven by benchmarks and datasets.

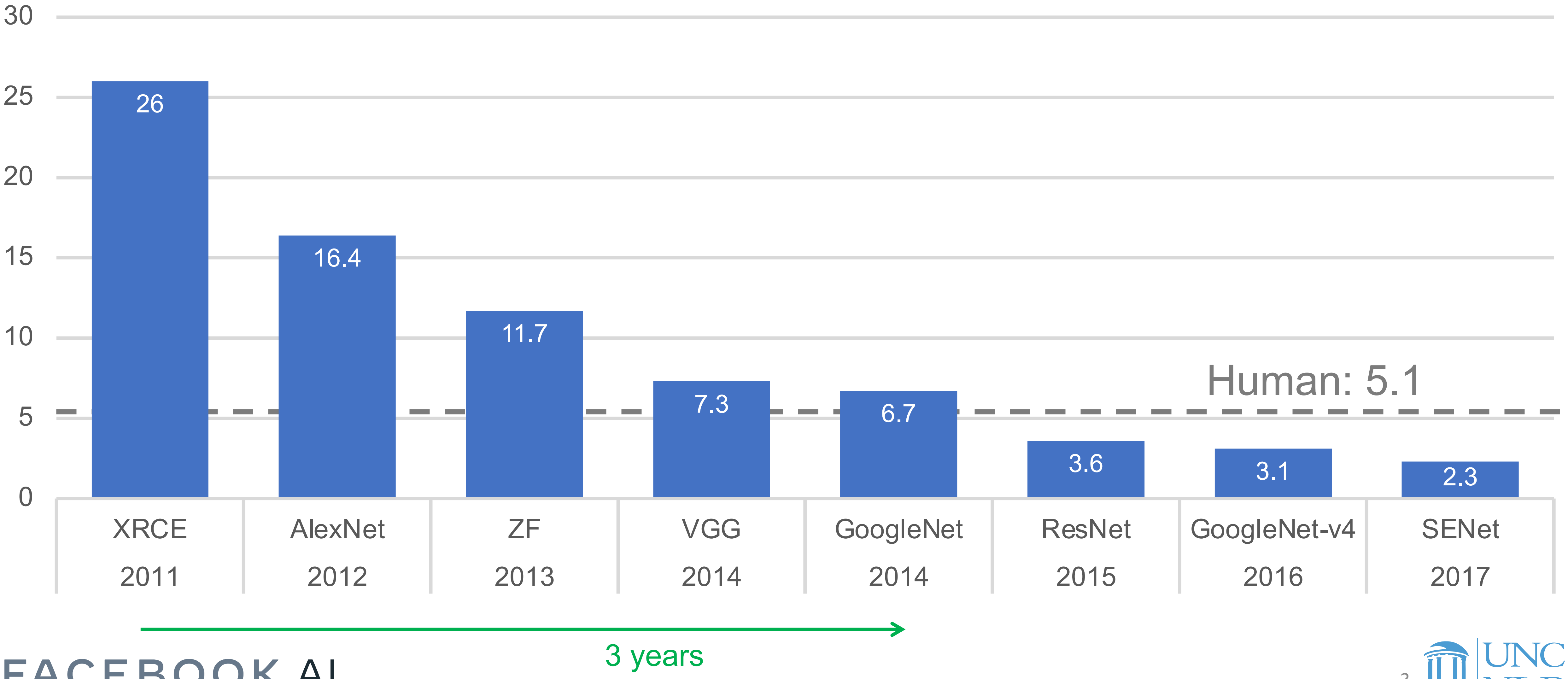
Computer Vision: IMAGENET (Russakovsky et al. 2015)

NLP: SQuAD (Rajpurkar et al. 2016),

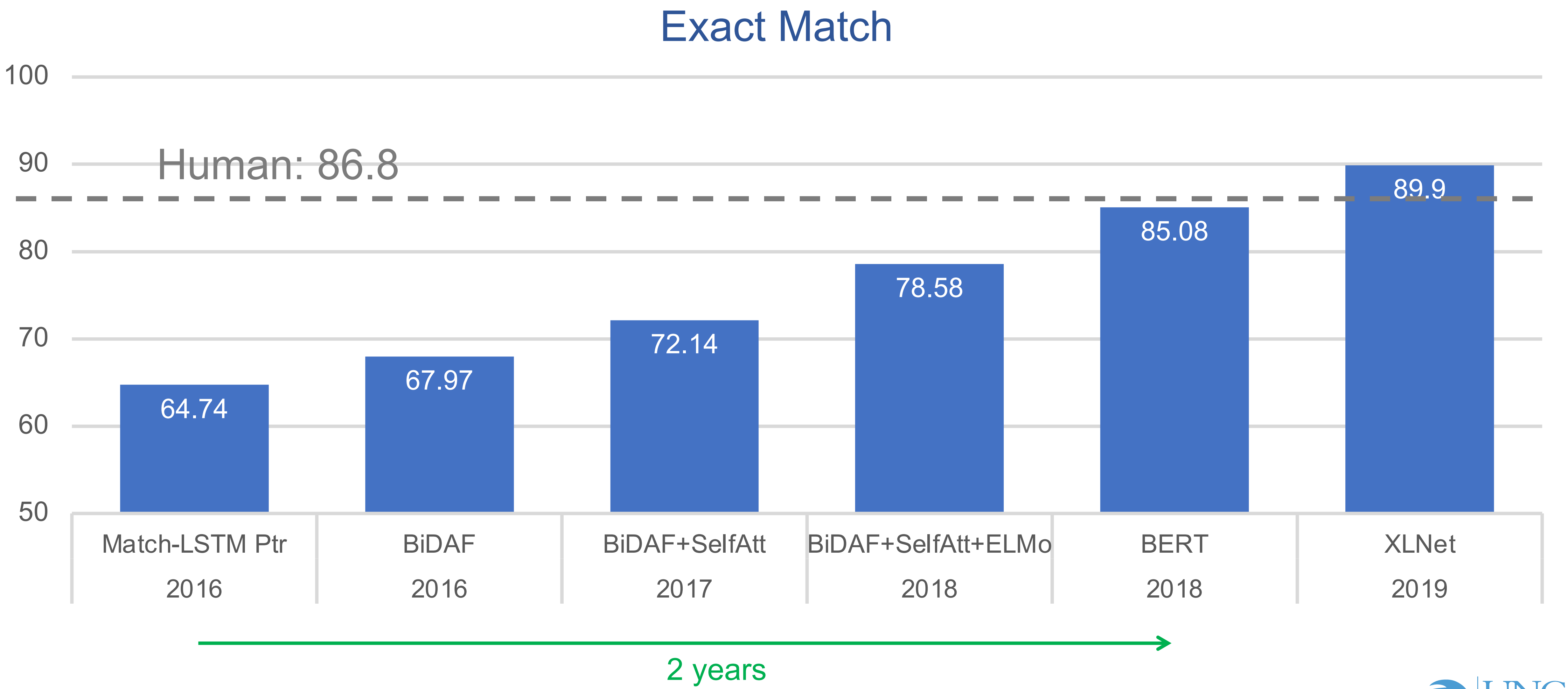
GLUE (Wang et al. 2018)

IMAGENET

Error Rate



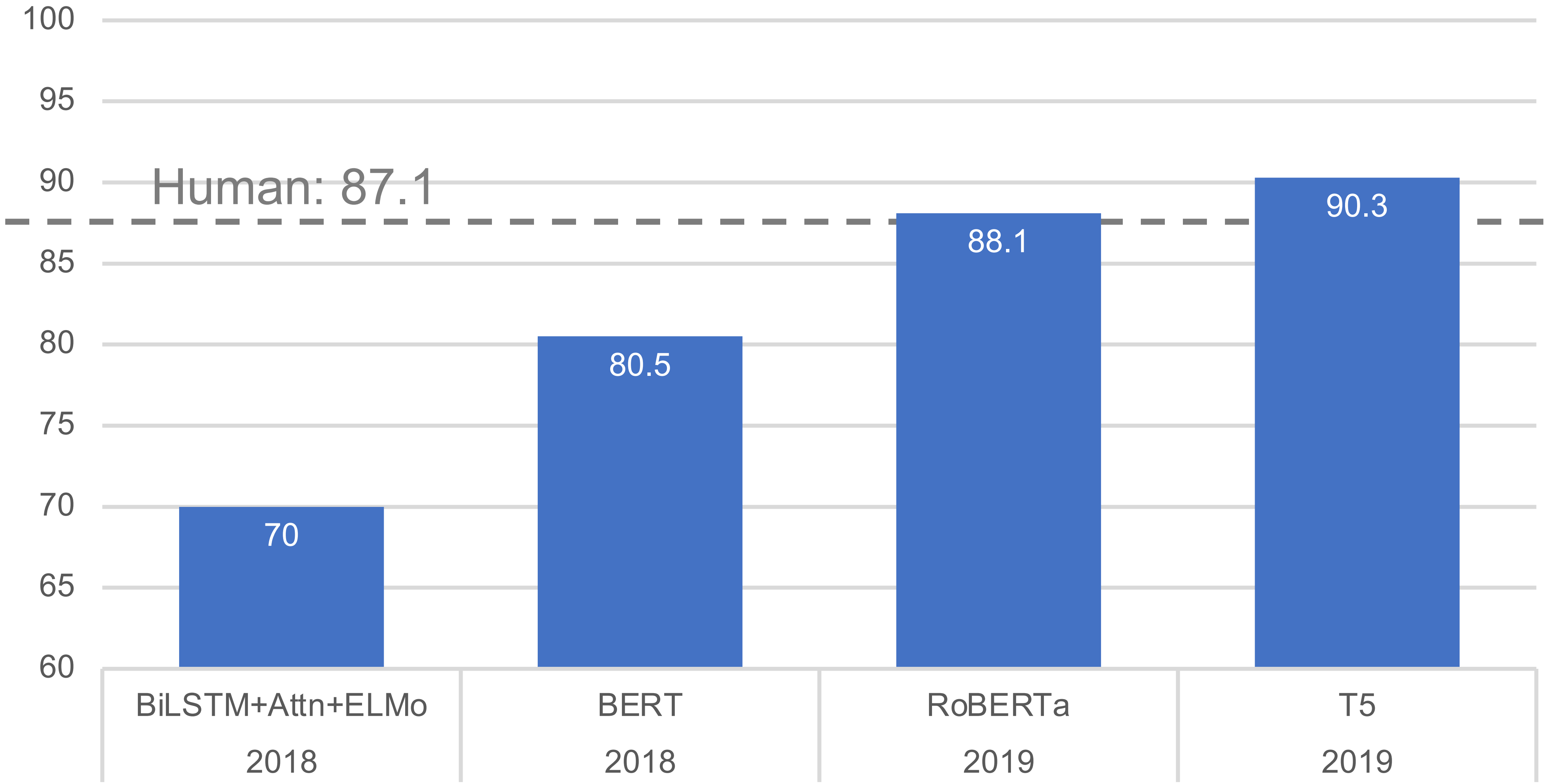
SQuAD



FACEBOOK AI

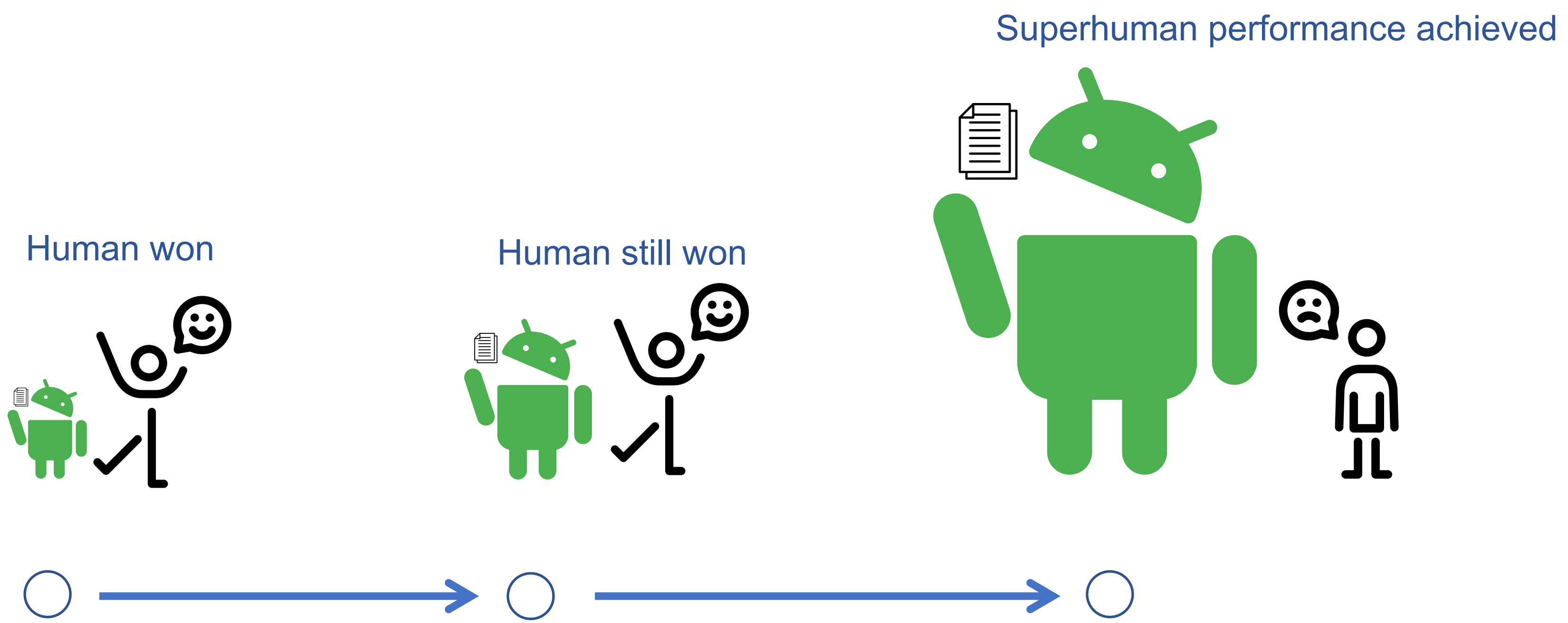
UNC NLP

GLUE
Score



1 year

Model vs. Human on Static Benchmarks



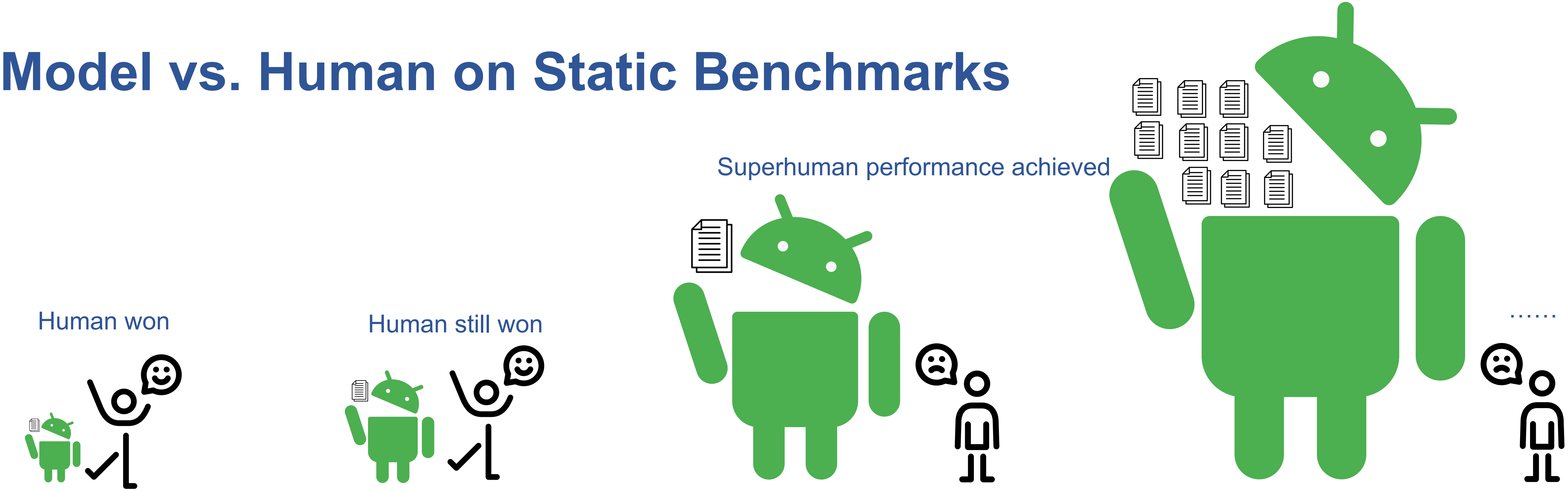
- Word2Vec
- Glove

- ELMo
- GPT-1

- BERT
- RoBERTa
- GPT-2

…

Model vs. Human on Static Benchmarks



- Word2Vec
- Glove

- ELMo
- GPT-1

- BERT
- RoBERTa
- GPT-2

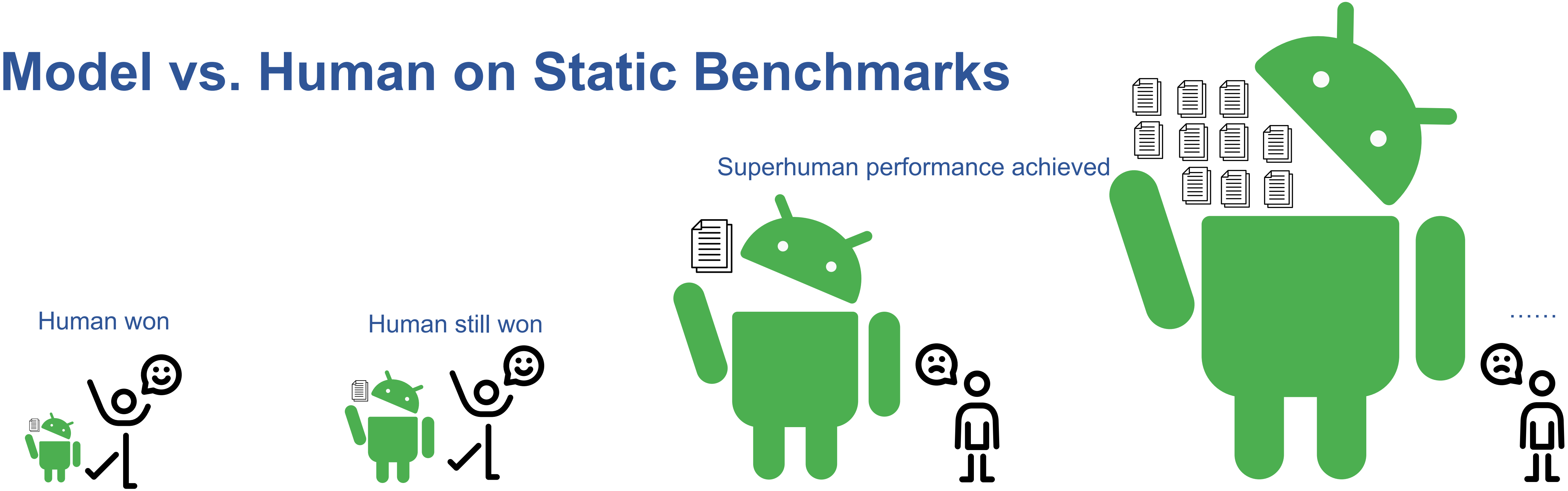
…

- T5
- GPT-3

FACEBOOK AI

UNC NLP

Model vs. Human on Static Benchmarks



- Word2Vec
- Glove

- ELMo
- GPT-1

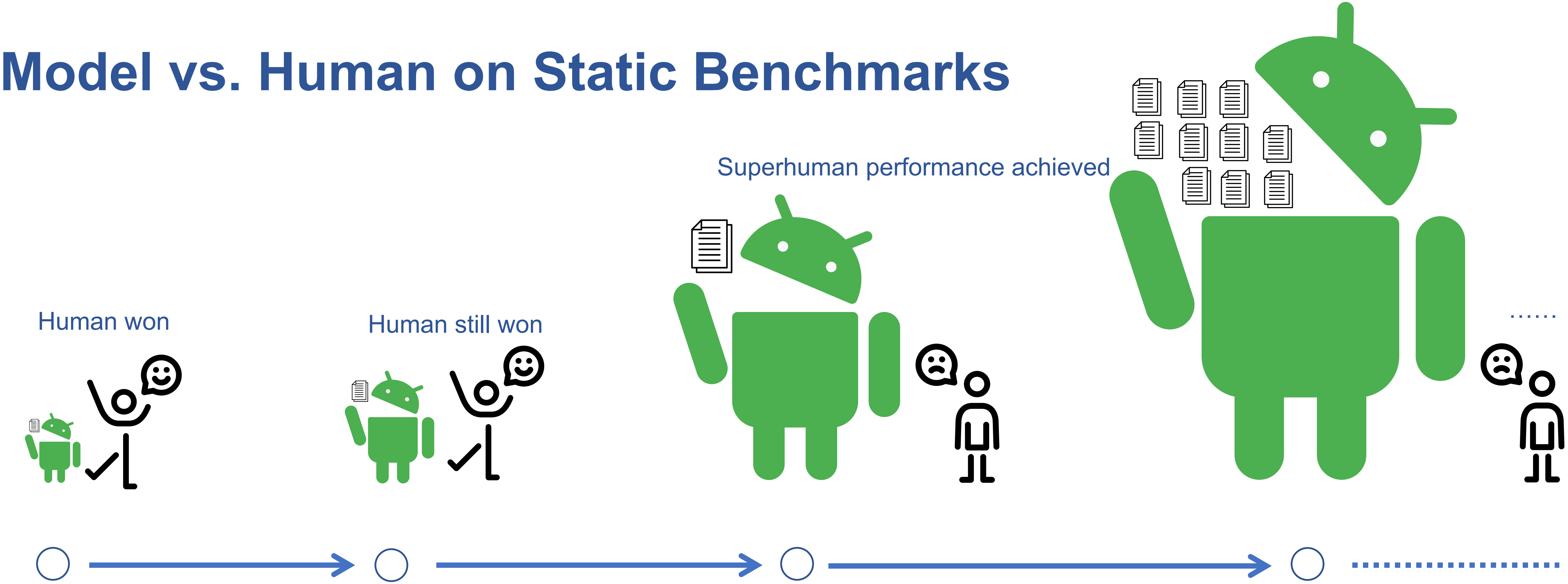
- BERT
- RoBERTa
- GPT-2

…

- T5
- GPT-3

Superhuman at NLU?

Model vs. Human on Static Benchmarks



- Word2Vec
- Glove

- ELMo
- GPT-1

- BERT
- RoBERTa
- GPT-2

…

- T5
- GPT-3

Are current NLU models genuinely as good as their high performance on static benchmark?

Overestimated NLU Ability

The state-of-the-art models learn to exploit spurious statistical patterns and are vulnerable to adversaries.

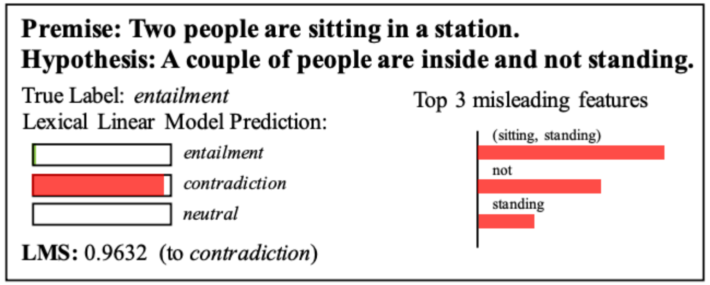
Adversary for reading comprehension

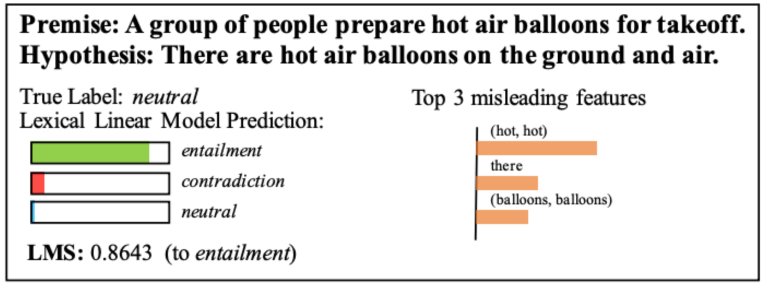
(Jia and Liang, 2017)

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Adversary for natural language inference

(Nie et al., 2018)





Overestimated NLU Ability

The state-of-the-art models learn to exploit spurious statistical patterns and are vulnerable to adversaries.

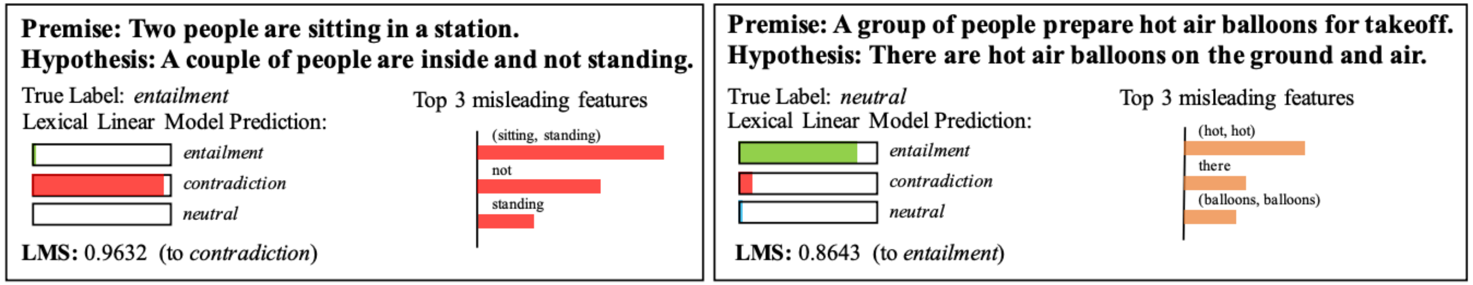
Adversary for reading comprehension

(Jia and Liang, 2017)

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Adversary for natural language inference

(Nie et al., 2018)



- Annotation artifacts (Gururangan et al., 2018, Poliak et al. 2018)
- Breaking NLI with lexical inference (Glockner et al., 2018)
- Pathologies of Neural Models (Feng et al., 2018)
- Modeling task or annotator? (Geva et al., 2019)
- Right for the wrong reason (McCoy et al., 2019)

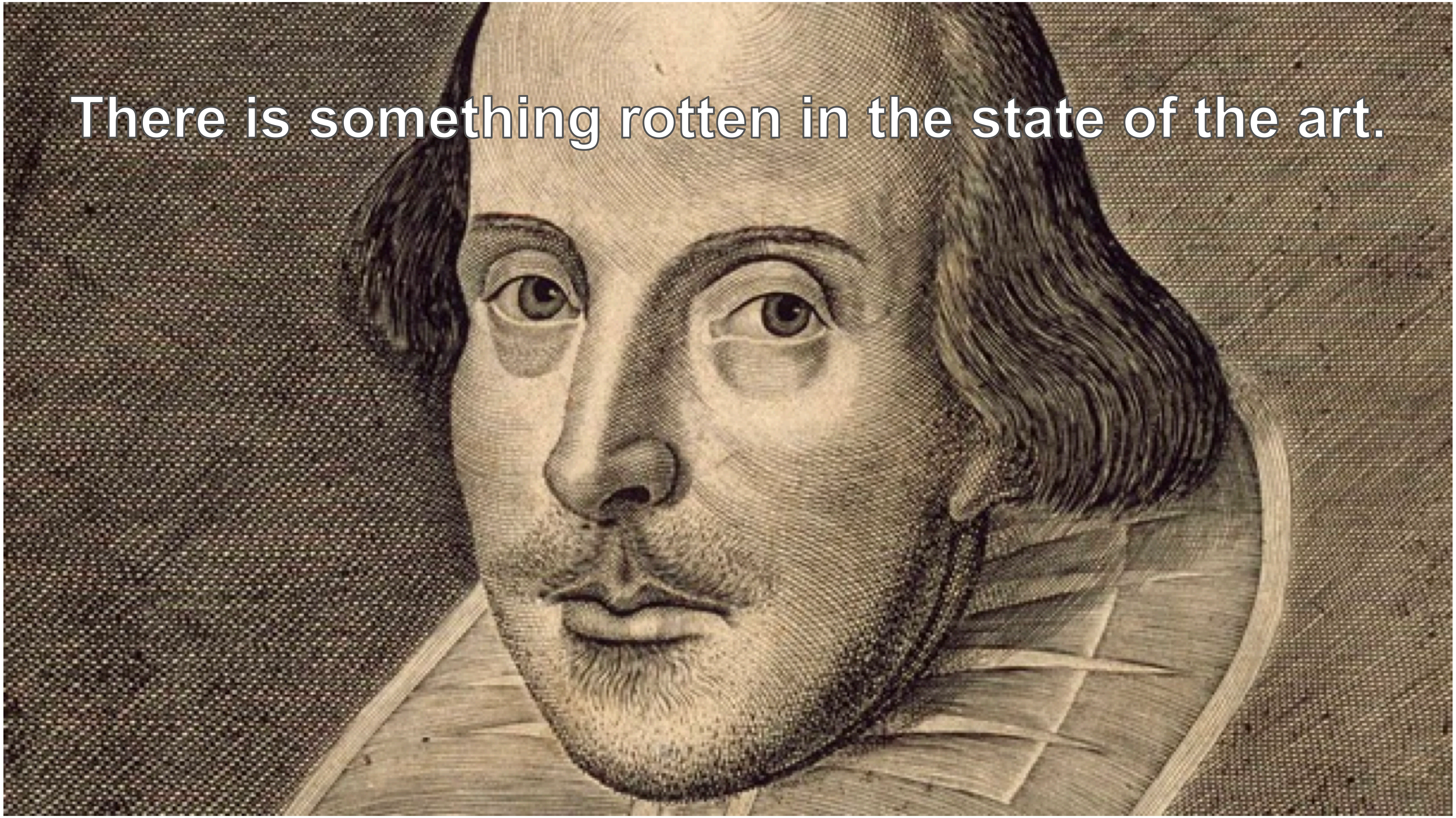
…

Performance is Overestimated

Model brittleness can be exposed by researchers or non-experts.

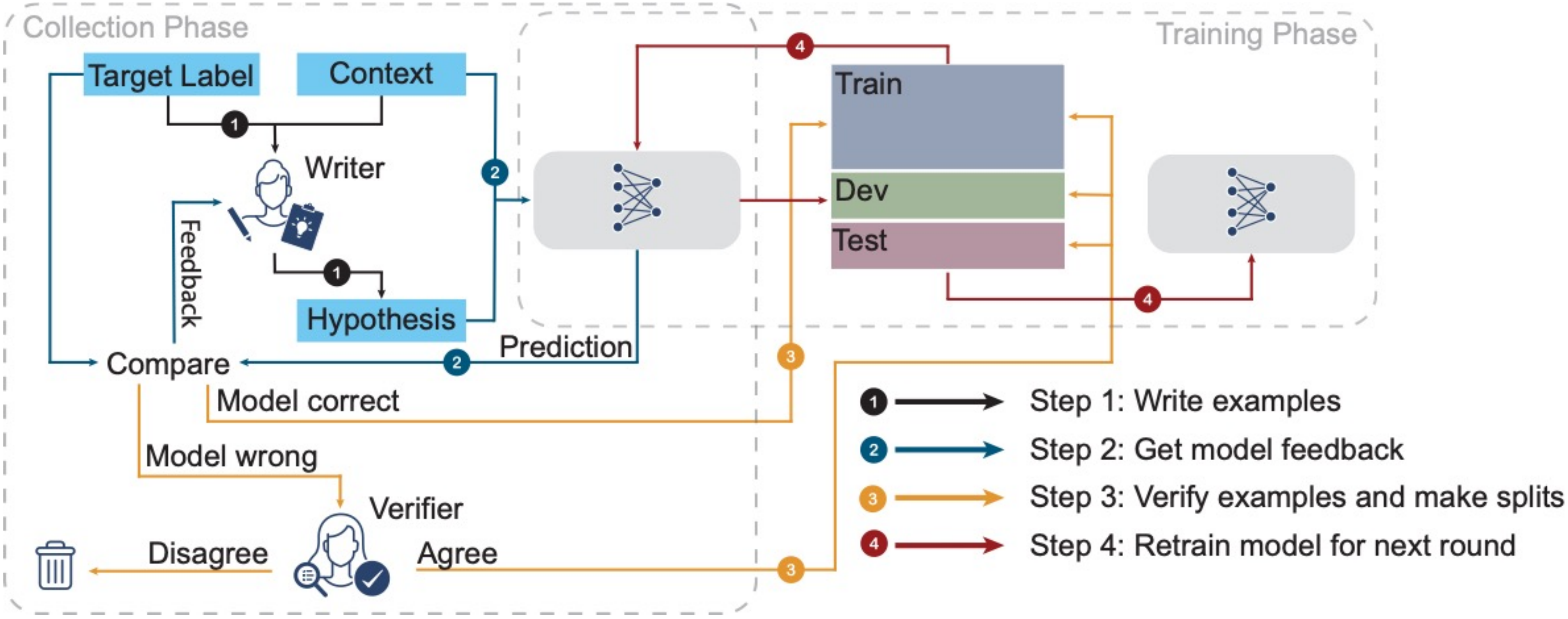
General NLU is still far from achieved despite the high performance.

How to solve the benchmark **fast-saturation** and **robustness** issues?


There is something rotten in the state of the art.

HAMLET

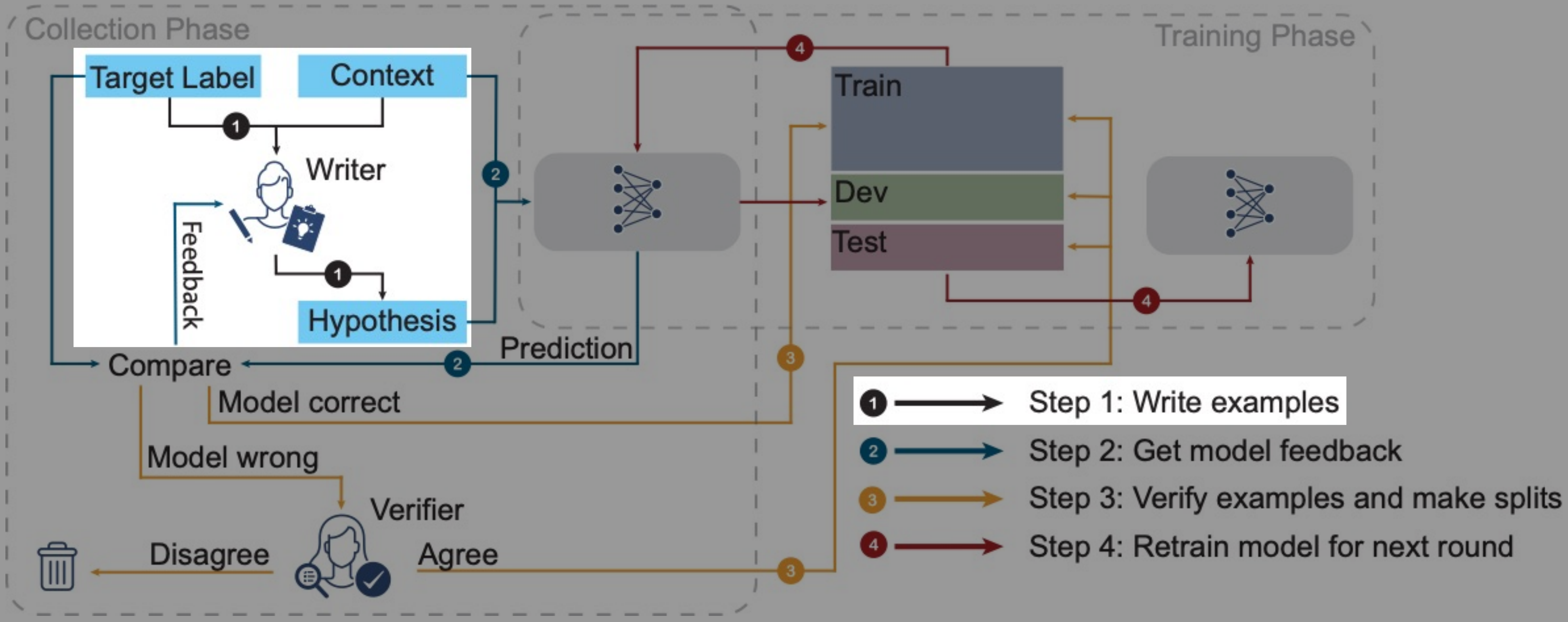
Human-And-Model-in-the-Loop Enabled Training



Context is also premise according to NLI terminology.

HAMLET

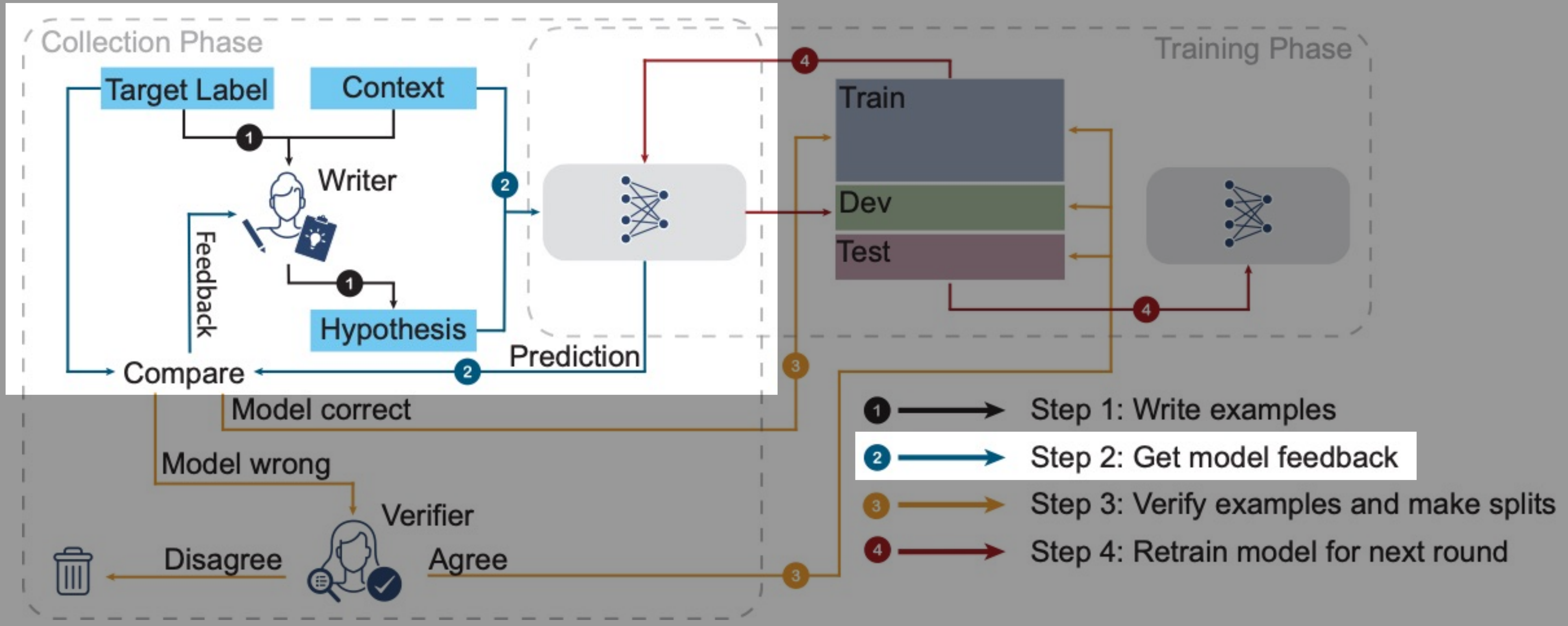
Human-And-Model-in-the-Loop Enabled Training



Context is also premise according to NLI terminology.

HAMLET

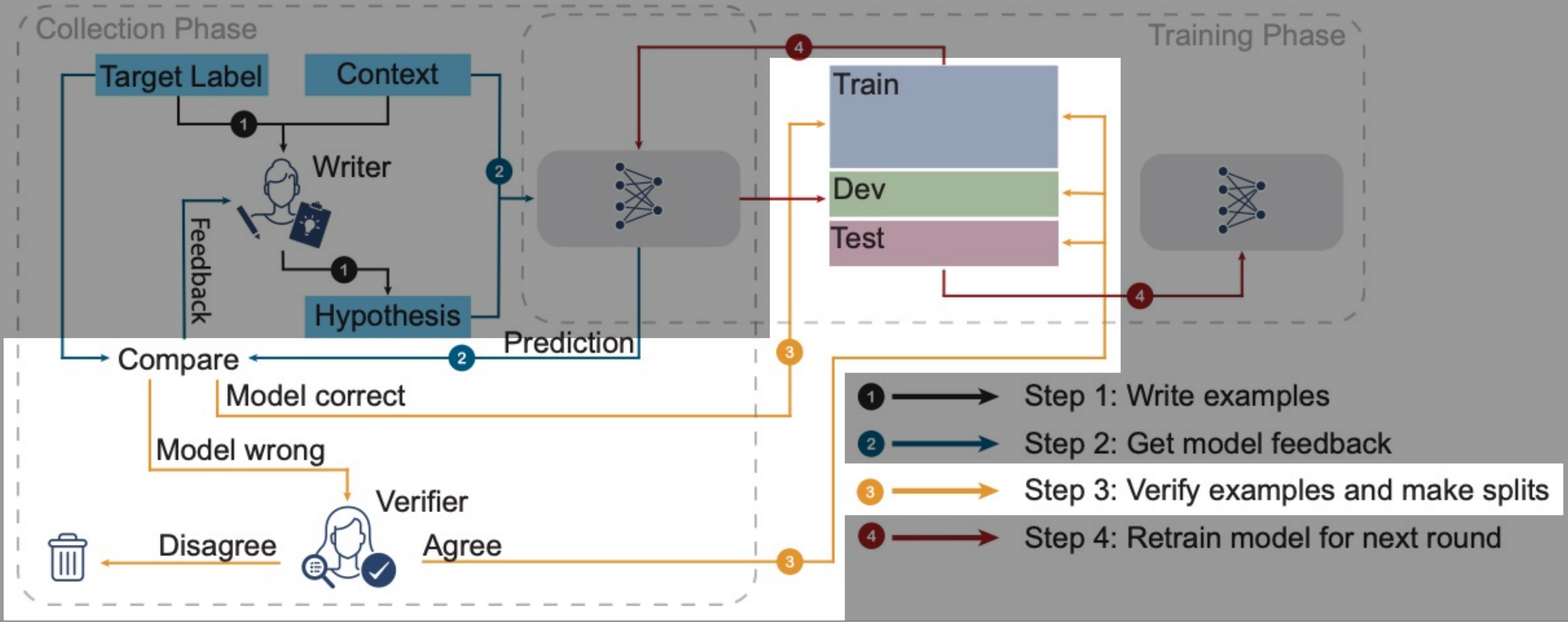
Human-And-Model-in-the-Loop Enabled Training



Context is also premise according to NLI terminology.

HAMLET

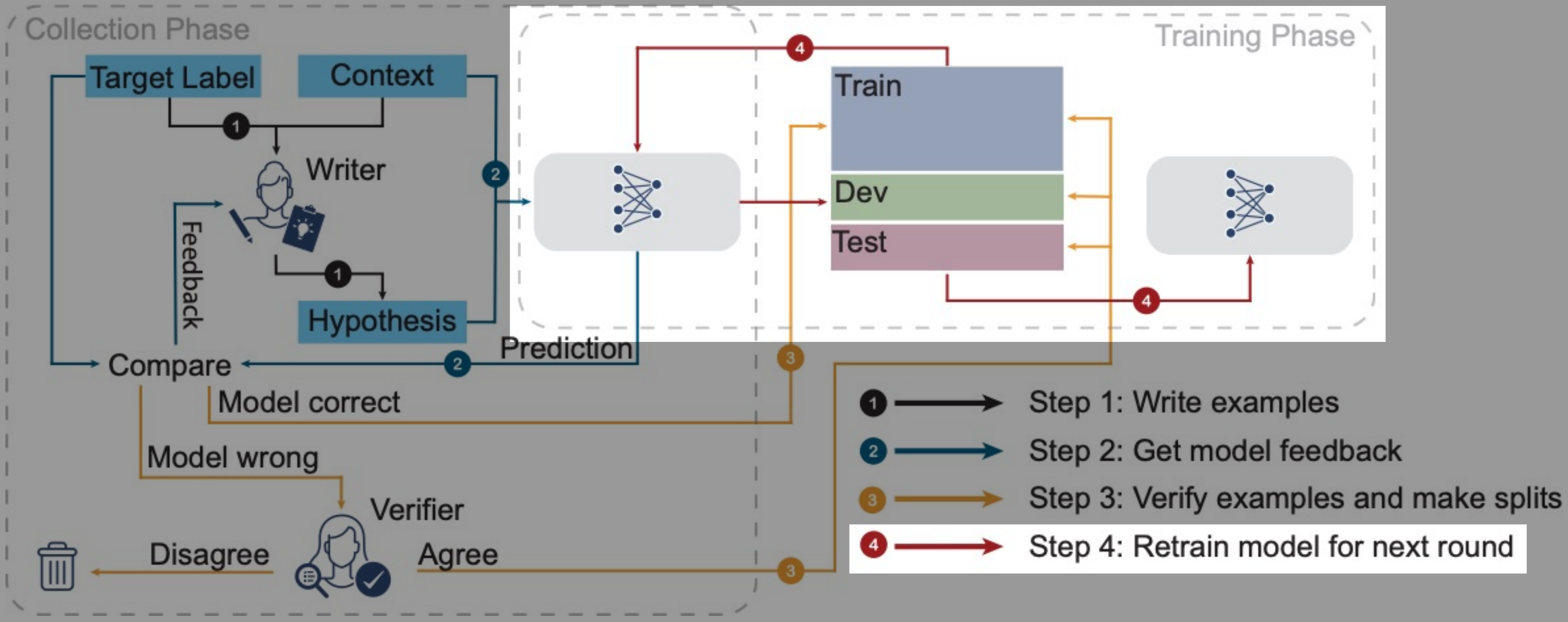
Human-And-Model-in-the-Loop Enabled Training



Context is also premise according to NLI terminology.

HAMLET

Human-And-Model-in-the-Loop Enabled Training

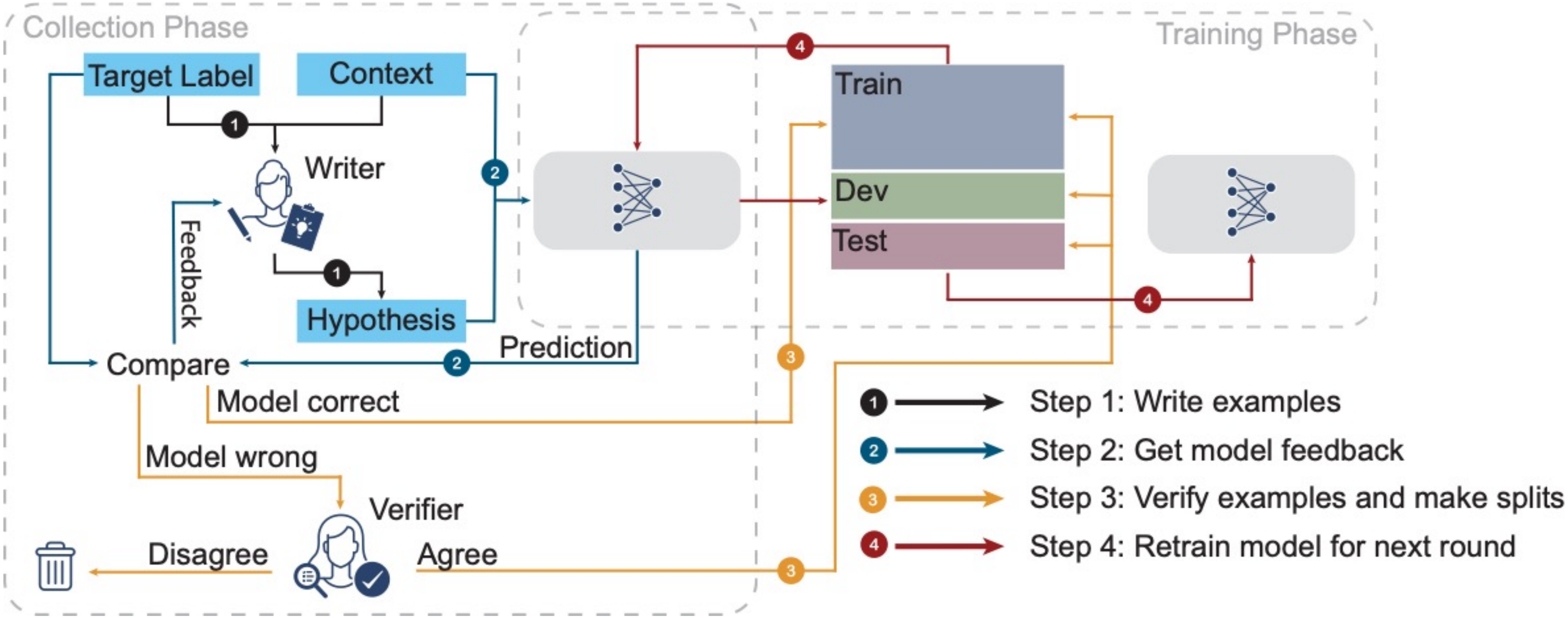


FACEBOOK AI

Context is also premise according to NLI terminology.

HAMLET

Human-And-Model-in-the-Loop Enabled Training



FACEBOOK AI

Context is also premise according to NLI terminology.

UNC NLP

Related work

Adversarial & Human-in-the-Loop

Build It, Break It, Fix It: Contesting Secure Development

James Parker, Michael Hicks, Andrew Ruef, Michelle L. Mazurek, Dave Levin, Daniel Votipka, Piotr Mardziel, Kelsey R. Fulton

Universal Adversarial Triggers for Attacking and Analyzing NLP

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, Sameer Singh

Towards Linguistically Generalizable NLP Systems: A Workshop and Shared Task

Allyson Ettinger, Sudha Rao, Hal Daumé III, Emily M. Bender

Mastering the Dungeon: Grounded Language Learning by Mechanical Turker Descent

Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H. Miller, Arthur Szlam, Douwe Kiela, Jason Weston

Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack

Emily Dinan, Samuel Humeau, Bharath Chintagunta, Jason Weston

Adversarial Filters of Dataset Biases

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, Yejin Choi

SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference

Rowan Zellers, Yonatan Bisk, Roy Schwartz, Yejin Choi

Trick Me If You Can: Human-in-the-loop Generation of Adversarial Examples for Question Answering

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, Jordan Boyd-Graber

Adversarial attacks against Fact Extraction and VERification

James Thorne, Andreas Vlachos

CODAH: An Adversarially Authored Question-Answer Dataset for Common Sense

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, Doug Downey

Adversarial NLI (ANLI)

Analogy: white-hat hackers finding vulnerabilities in models, which we then patch for the next round.

Three rounds of data collection.

- **Round 1**
  Model: BERT (Trained on **SNLI+MNLI**)
  Domain: Wikipedia

- **Round 2**
  Model: RoBERTa ensemble (Trained on **SNLI+MNLI+FEVER+A1**)
  Domain: Wikipedia

- **Round 3**
  Model: RoBERTa ensemble (Trained on **SNLI+MNLI+FEVER+A1+A2**)
  Domains: Wikipedia, News, Fiction, Spoken, WikiHow, RTE5

Adversarial NLI (ANLI)

Analogy: white-hat hackers finding vulnerabilities in models, which we then patch for the next round.

Three rounds of data collection.

- **Round 1 (A1)**
  Model: BERT (Trained on **SNLI+MNLI**)
  Domain: Wikipedia

- **Round 2 (A2)**
  Model: RoBERTa ensemble (Trained on **SNLI+MNLI+FEVER+A1**)
  Domain: Wikipedia

- **Round 3 (A3)**
  Model: RoBERTa ensemble (Trained on **SNLI+MNLI+FEVER+A1+A2**)
  Domains: Wikipedia, News, Fiction, Spoken, WikiHow, RTE5

Dataset	Genre	Context	Train / Dev / Test
A1	Wiki	2,080	16,946 / 1,000 / 1,000
A2	Wiki	2,694	45,460 / 1,000 / 1,000
A3	Various (Wiki subset)	6,002 1,000	100,459 / 1,200 / 1,200 19,920 / 200 / 200
ANLI	Various	10,776	162,865 / 3,200 / 3,200

SNLI: 570K
MNLI: 433K

ANLI: 163K

Adversarial NLI (ANLI)

Analogy: white-hat hackers finding vulnerabilities in models, which we then patch for the next round.

Three rounds of data collection.

- **Round 1 (A1)**
  Model: BERT (Trained on **SNLI+MNLI**)
  Domain: Wikipedia

- **Round 2 (A2)**
  Model: RoBERTa ensemble (Trained on **SNLI+MNLI+FEVER+A1**)
  Domain: Wikipedia

- **Round 3 (A3)**
  Model: RoBERTa ensemble (Trained on **SNLI+MNLI+FEVER+A1+A2**)
  Domains: Wikipedia, News, Fiction, Spoken, WikiHow, RTE5

Dataset	Genre	Context	Train / Dev / Test
A1	Wiki	2,080	16,946 / 1,000 / 1,000
A2	Wiki	2,694	45,460 / 1,000 / 1,000
A3	Various (Wiki subset)	6,002 1,000	100,459 / 1,200 / 1,200 19,920 / 200 / 200
ANLI	Various	10,776	162,865 / 3,200 / 3,200

SNLI: 570K
MNLI: 433K

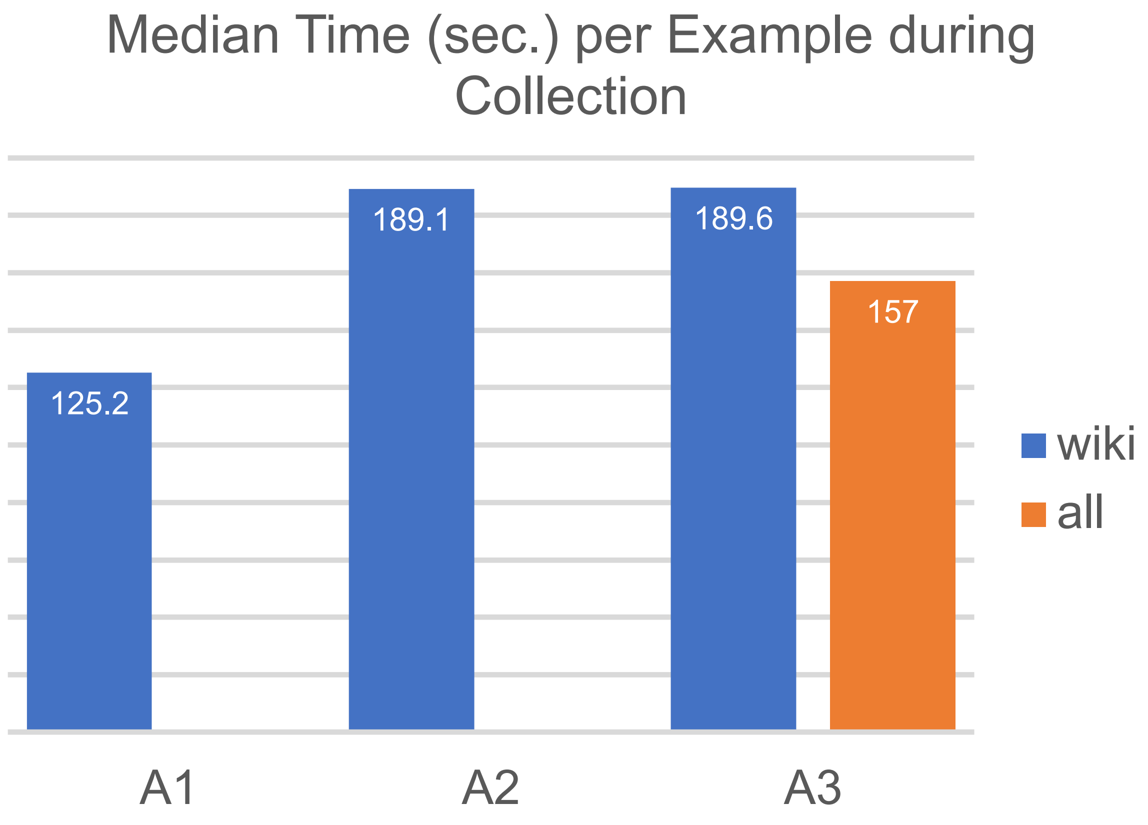
ANLI: 163K

- Adversarially collected
- More data-efficient in training

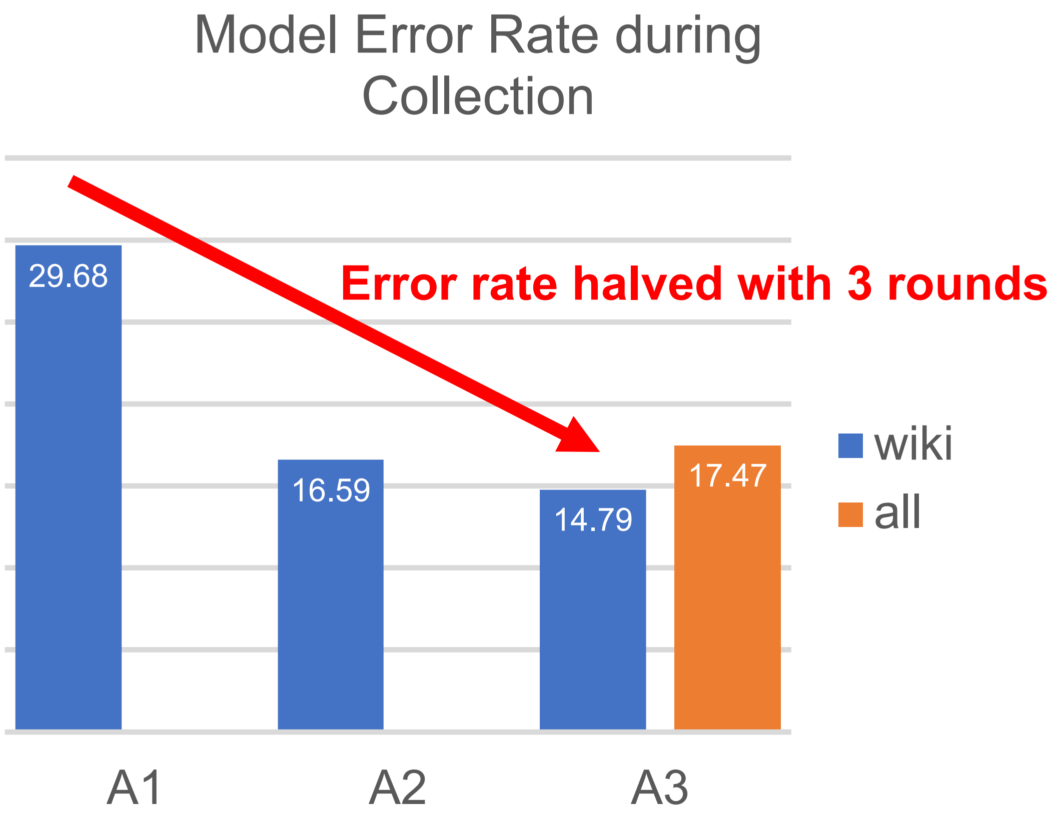
FACEBOOK AI

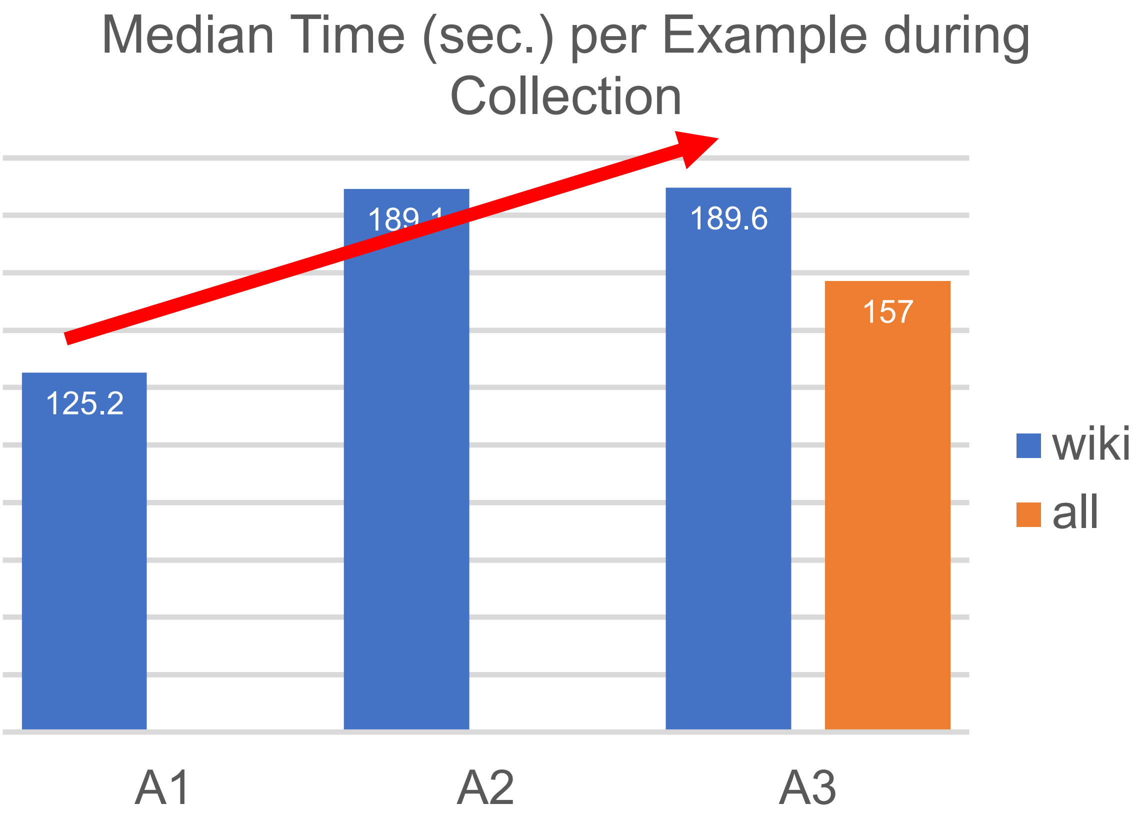
UNC NLP

Collection Statistics



Collection Statistics





Room for improvement on NLI still exists

FACEBOOK AI

UNC NLP

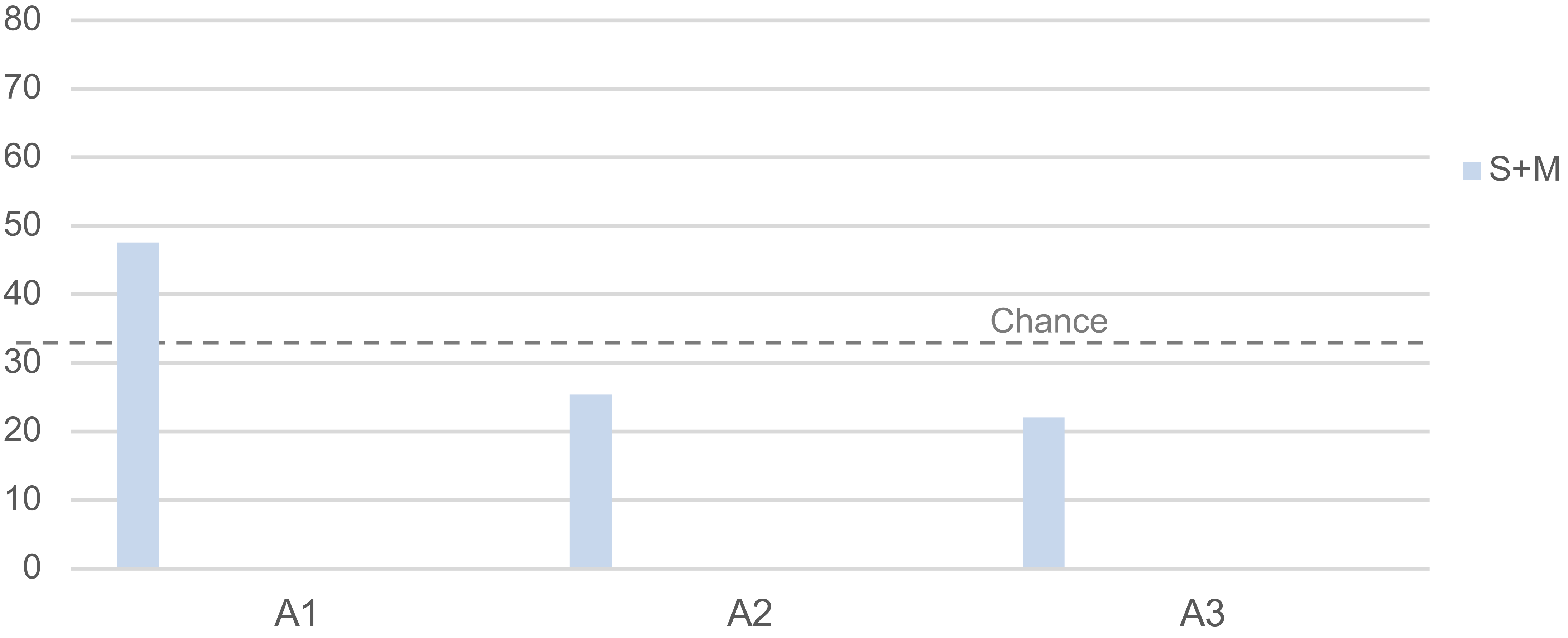
Findings

Base model (backend model in the collection) performance is low

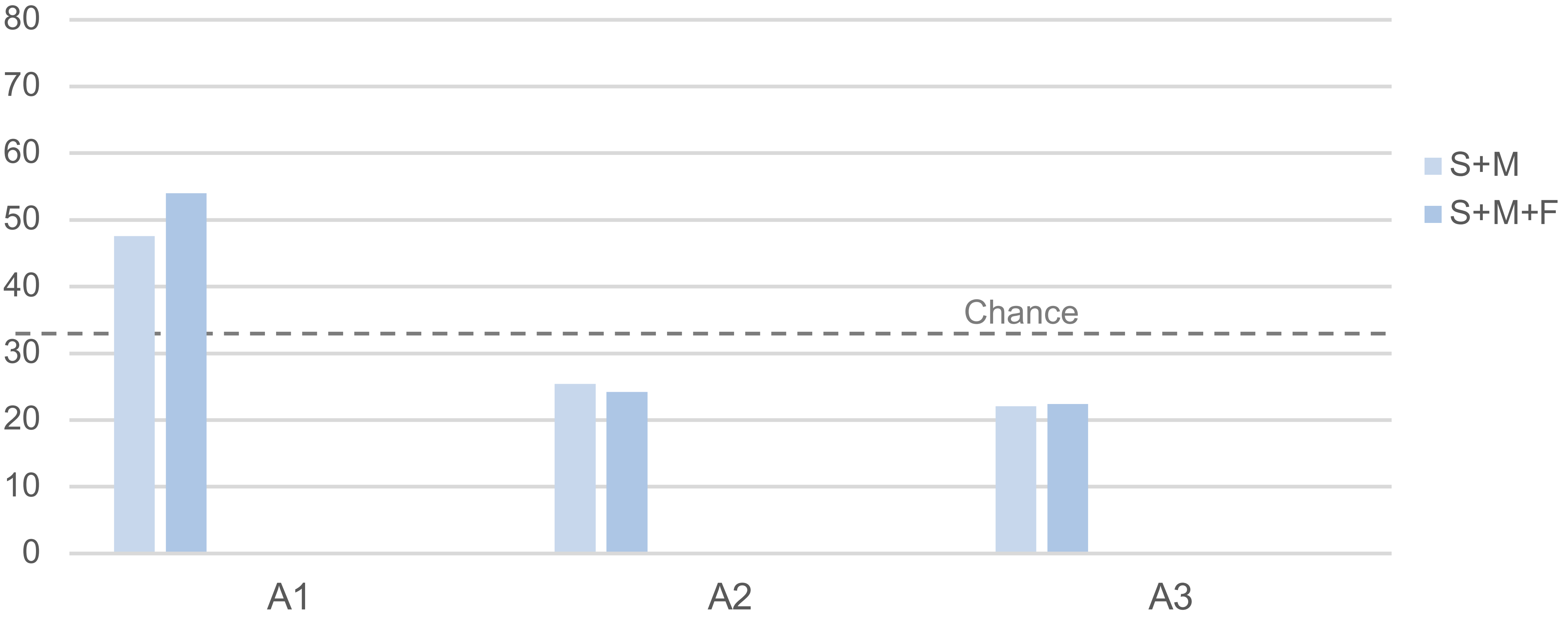
Model	Training Data	A1	A2	A3	ANLI	ANLI-E	SNLI	MNLI-m/-mm
BERT	S,M$^{\star 1}$	00.0	28.9	28.8	19.8	19.9	91.3	86.7 / 86.4
	+A1	44.2	32.6	29.3	35.0	34.2	91.3	86.3 / 86.5
	+A1+A2	57.3	45.2	33.4	44.6	43.2	90.9	86.3 / 86.3
	+A1+A2+A3	57.2	49.0	46.1	50.5	46.3	90.9	85.6 / 85.4
	S,M,F,ANLI	57.4	48.3	43.5	49.3	44.2	90.4	86.0 / 85.8
XLNet	S,M,F,ANLI	67.6	50.7	48.3	55.1	52.0	91.8	89.6 / 89.4
RoBERTa	S,M	47.6	25.4	22.1	31.1	31.4	92.6	90.8 / 90.6
	+F	54.0	24.2	22.4	32.8	33.7	92.7	90.6 / 90.5
	+F+A1$^{\star 2}$	68.7	19.3	22.0	35.8	36.8	92.8	90.9 / 90.7
	+F+A1+A2$^{\star 3}$	71.2	44.3	20.4	43.7	41.4	92.9	91.0 / 90.7
	S,M,F,ANLI	73.8	48.9	44.4	53.7	49.7	92.6	91.0 / 90.6

Table 3: Model Performance. 'S' refers to SNLI, 'M' to MNLI dev (-m=matched, -mm=mismatched), and 'F' to FEVER; 'A1–A3' refer to the rounds respectively and 'ANLI' refers to A1+A2+A3, '-E' refers to test set examples written by annotators exclusive to the test set. Datasets marked '$\star n$' were used to train the base model for round n, and their performance on that round is underlined (A2 and A3 used ensembles, and hence have non-zero scores).

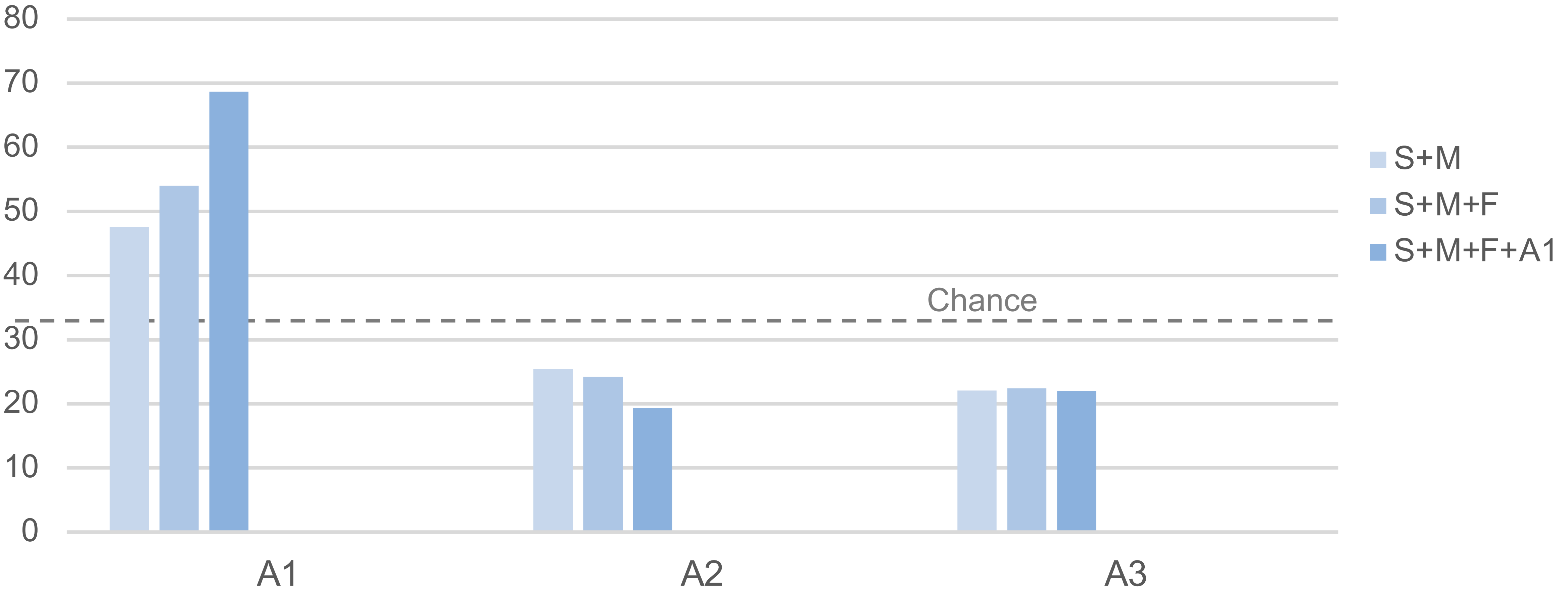
RoBERTa performance on different rounds as we accumulatively combine training data (S=SNLI, M=MNLI, F=FEVER)



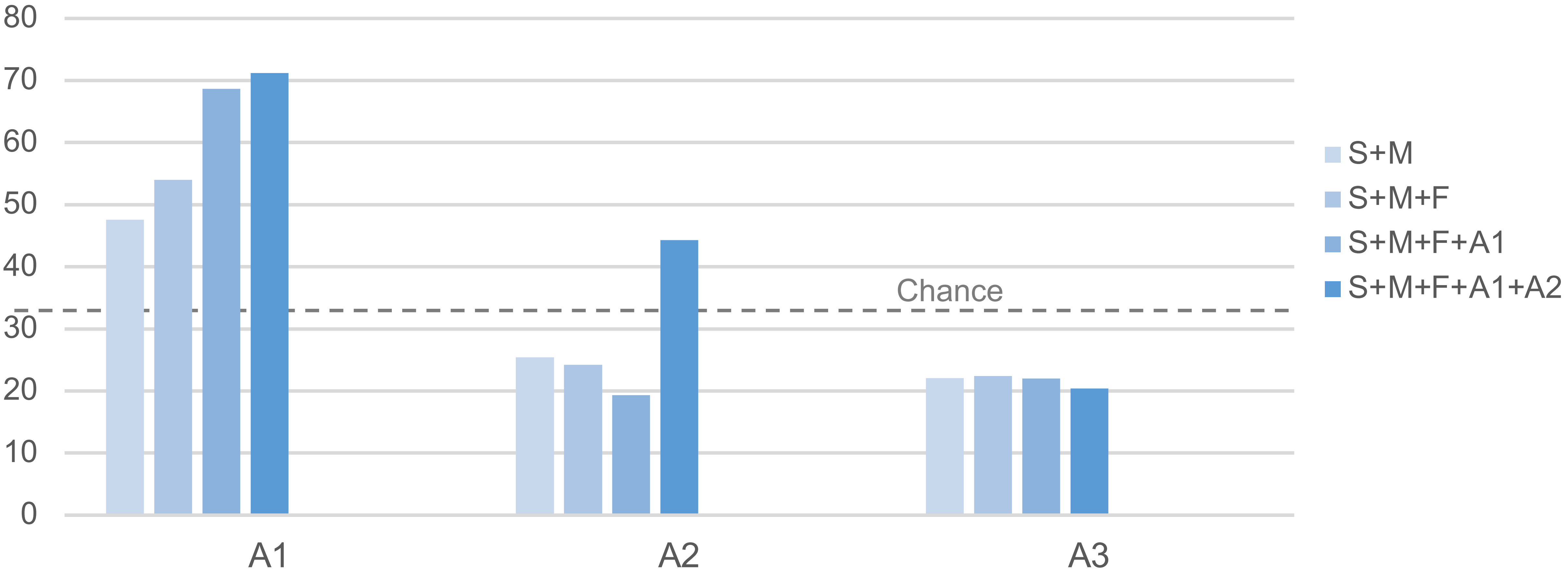
RoBERTa performance on different rounds as we accumulatively combine training data (S=SNLI, M=MNLI, F=FEVER)



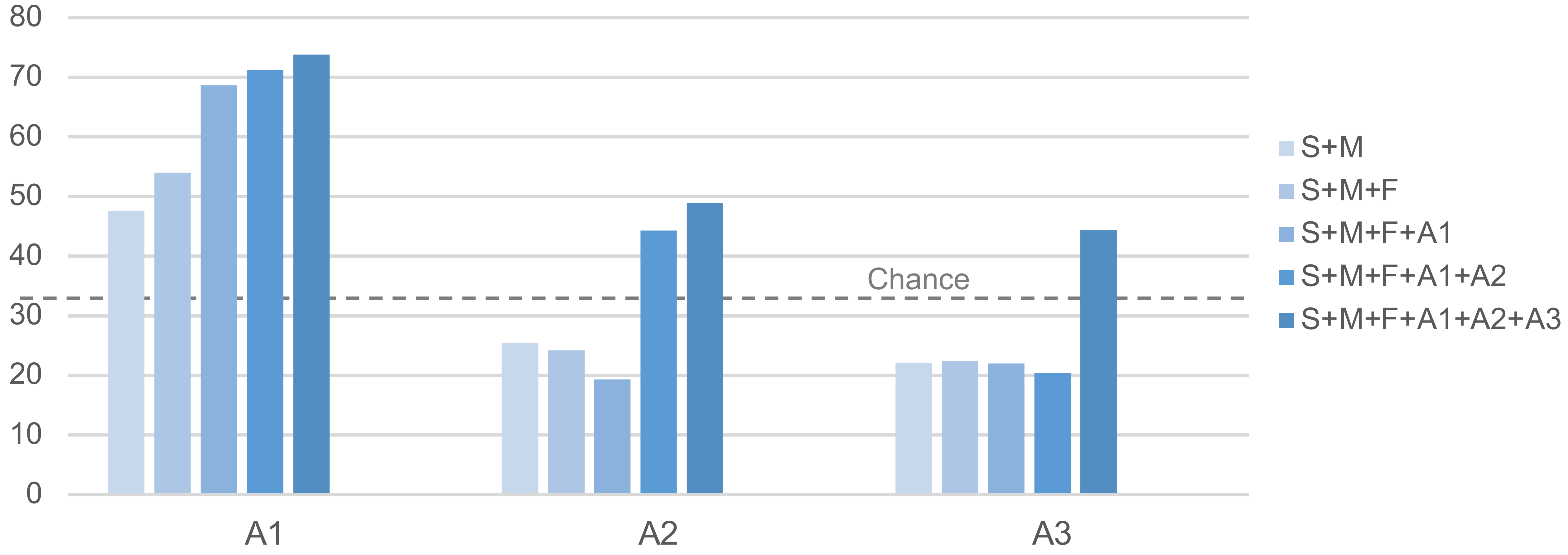
RoBERTa performance on different rounds as we accumulatively combine training data (S=SNLI, M=MNLI, F=FEVER)



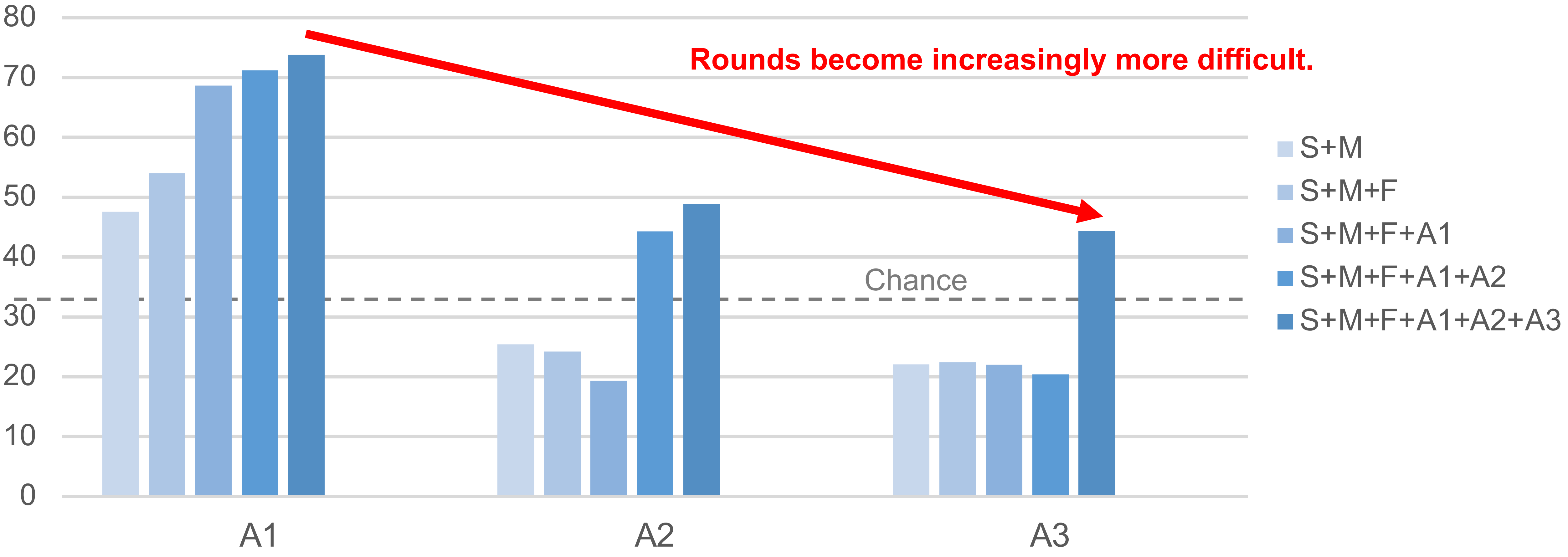
RoBERTa performance on different rounds as we accumulatively combine training data (S=SNLI, M=MNLI, F=FEVER)



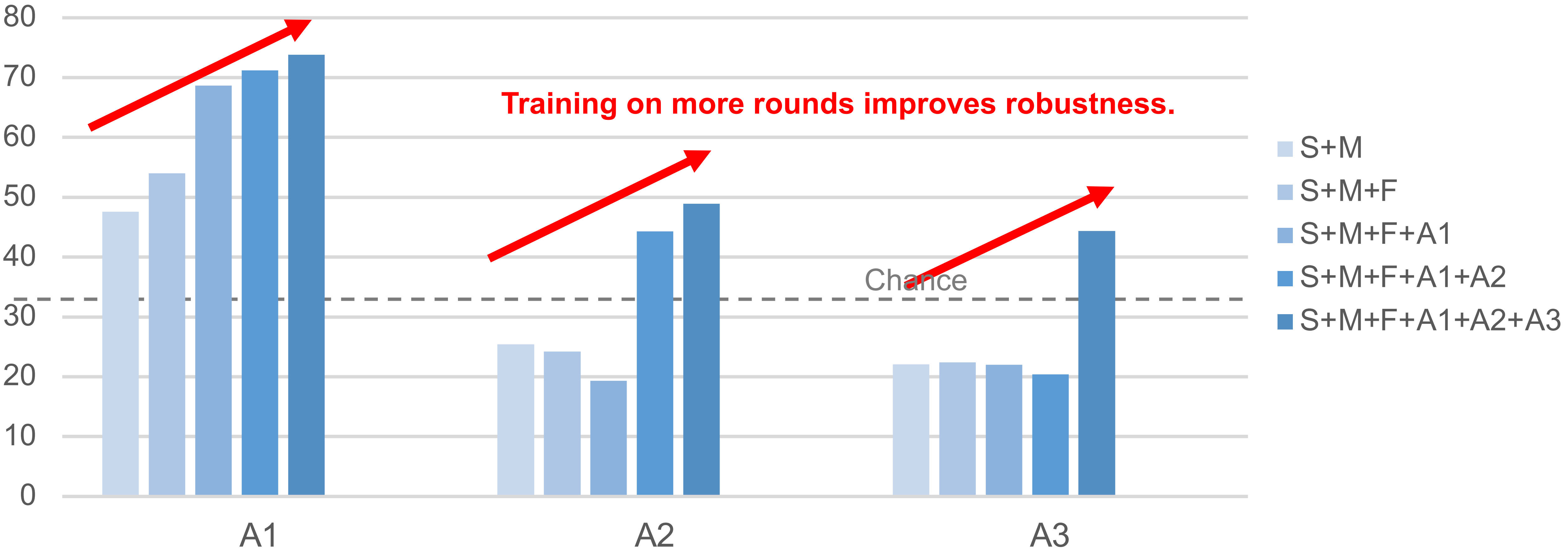
RoBERTa performance on different rounds as we accumulatively combine training data (S=SNLI, M=MNLI, F=FEVER)



RoBERTa performance on different rounds as we accumulatively combine training data (S=SNLI, M=MNLI, F=FEVER)



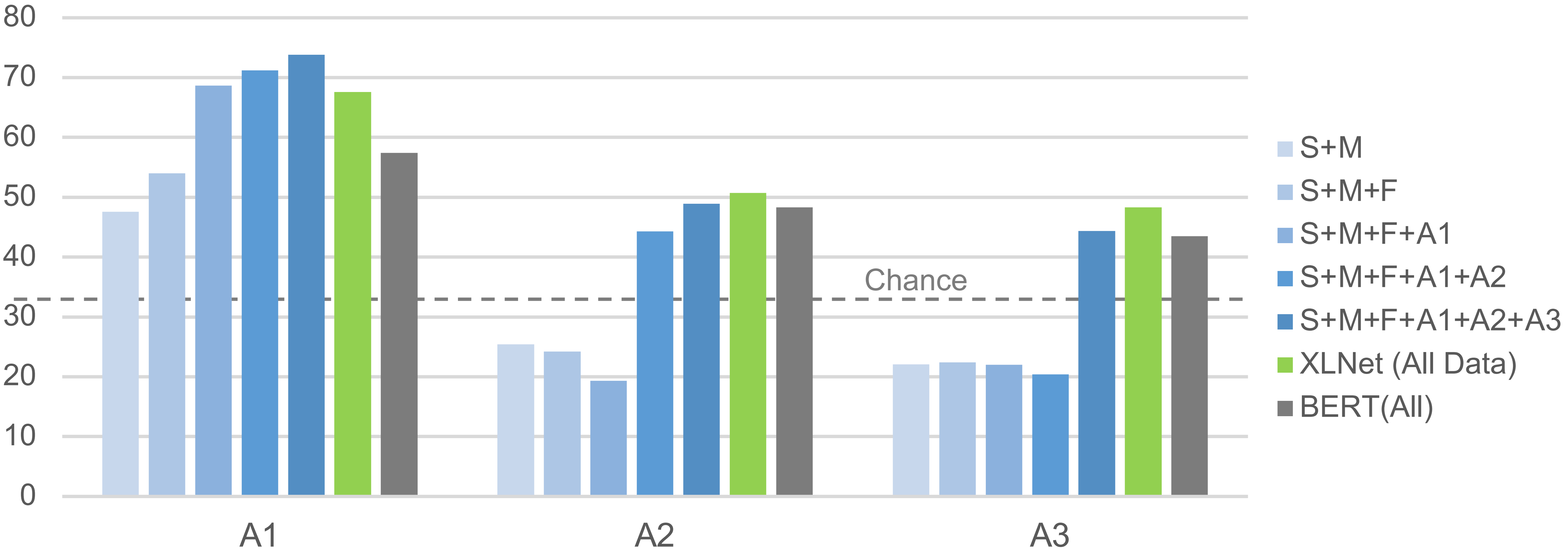
RoBERTa performance on different rounds as we accumulatively combine training data (S=SNLI, M=MNLI, F=FEVER)



FACEBOOK AI

UNC NLP

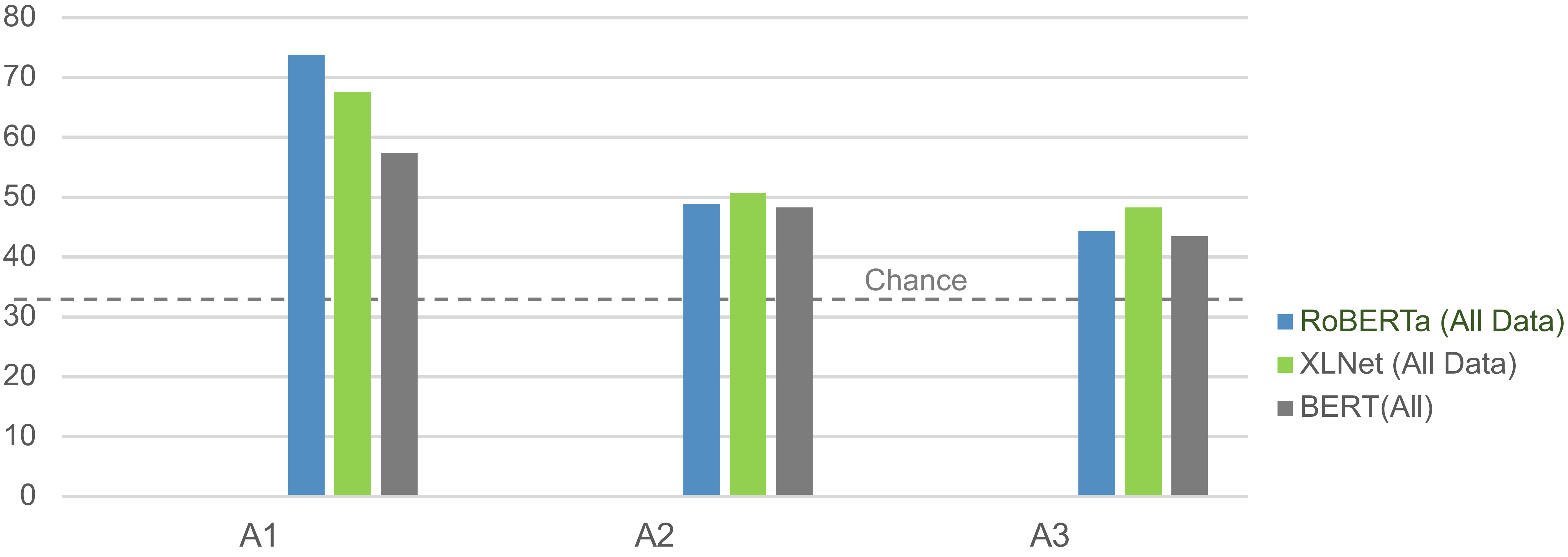
RoBERTa performance on different rounds as we accumulatively combine training data (S=SNLI, M=MNLI, F=FEVER)



FACEBOOK AI

UNC NLP

RoBERTa (All Data) vs. XLNet (All Data) vs. BERT (All Data)

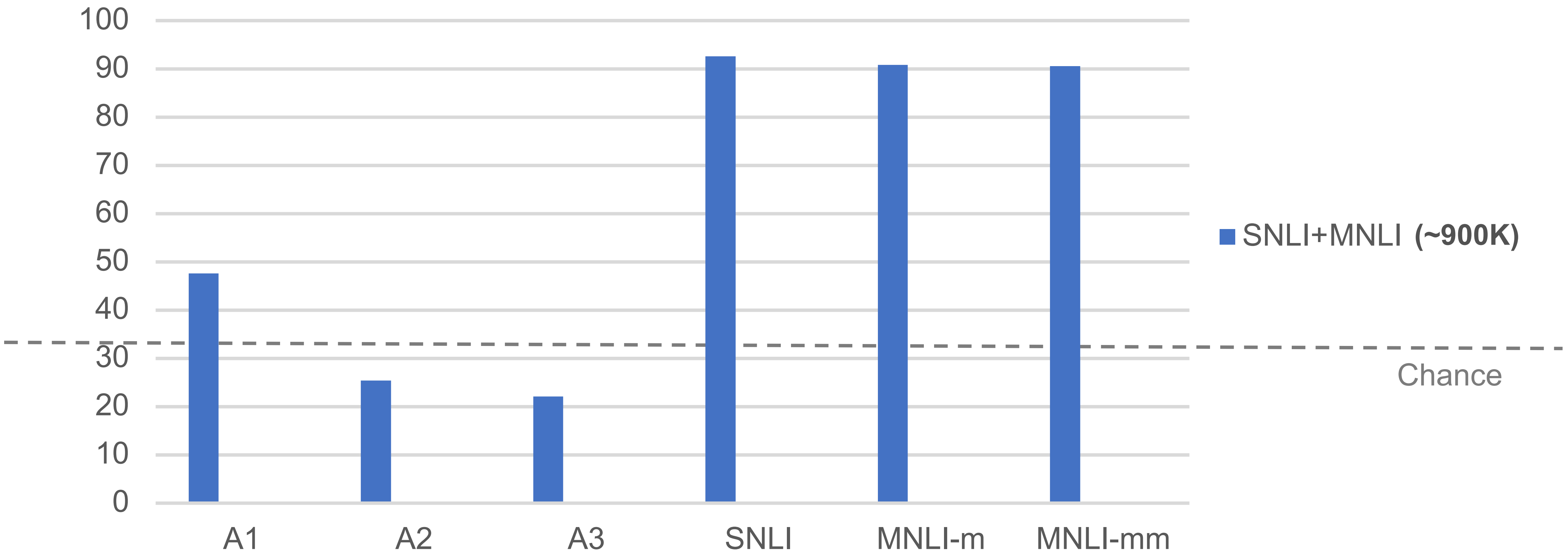


Different models have different weakness

FACEBOOK AI

UNC NLP

RoBERTa performance with different training data

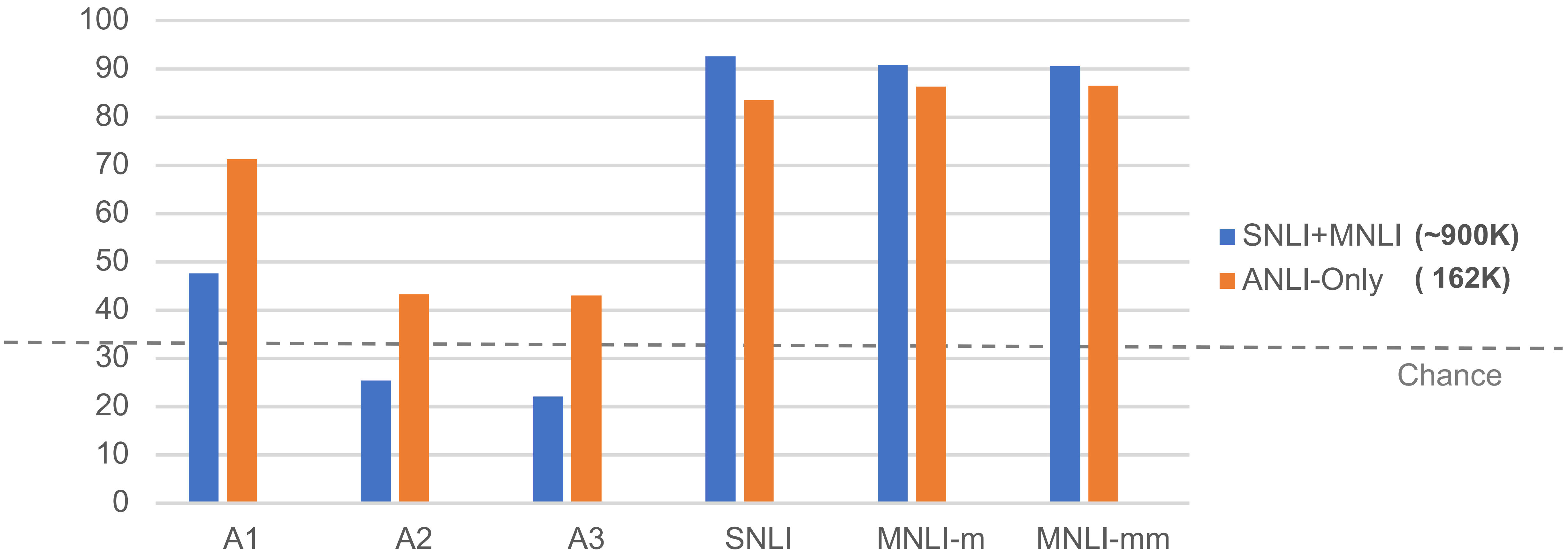


Model trained only on SNLI and MNLI (statically collected) is not good at ANLI

FACEBOOK AI

UNC NLP

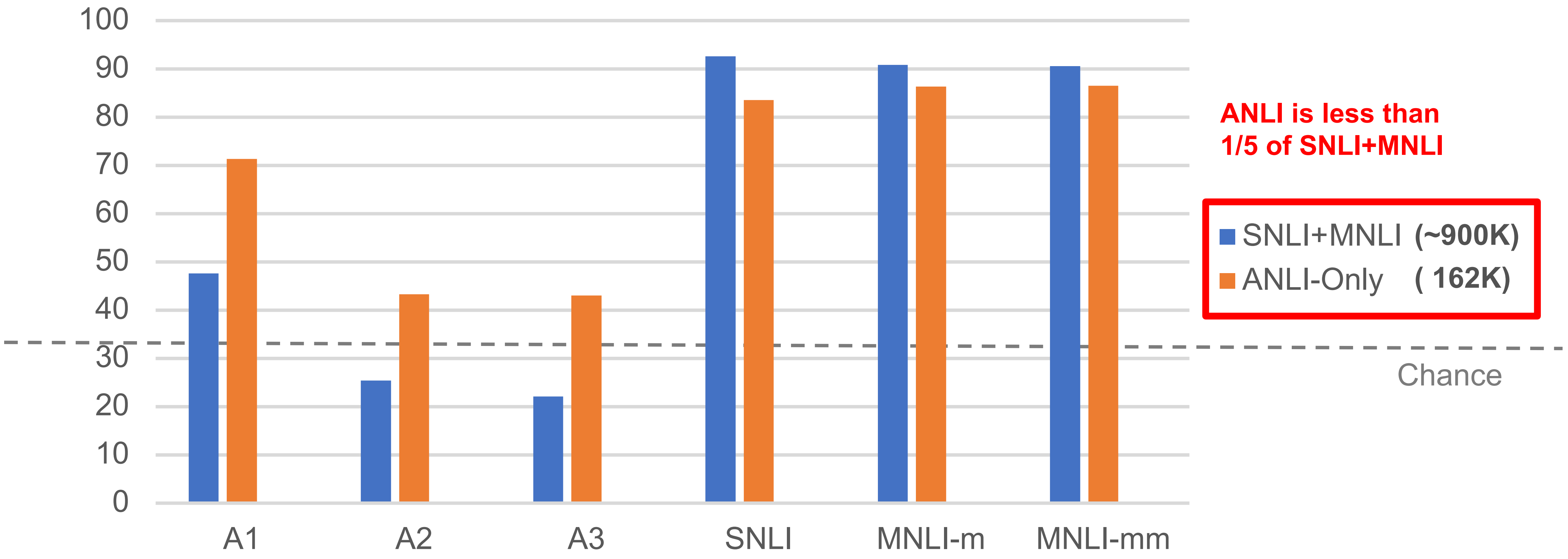
RoBERTa performance with different training data



Model trained only on SNLI and MNLI (statically collected) is not good at ANLI

But Model trained only on ANLI (adversarially collected) is reasonably good at SNLI and MNLI

RoBERTa performance with different training data



ANLI is less than 1/5 of SNLI+MNLI

SNLI+MNLI **(~900K)**
ANLI-Only **(162K)**

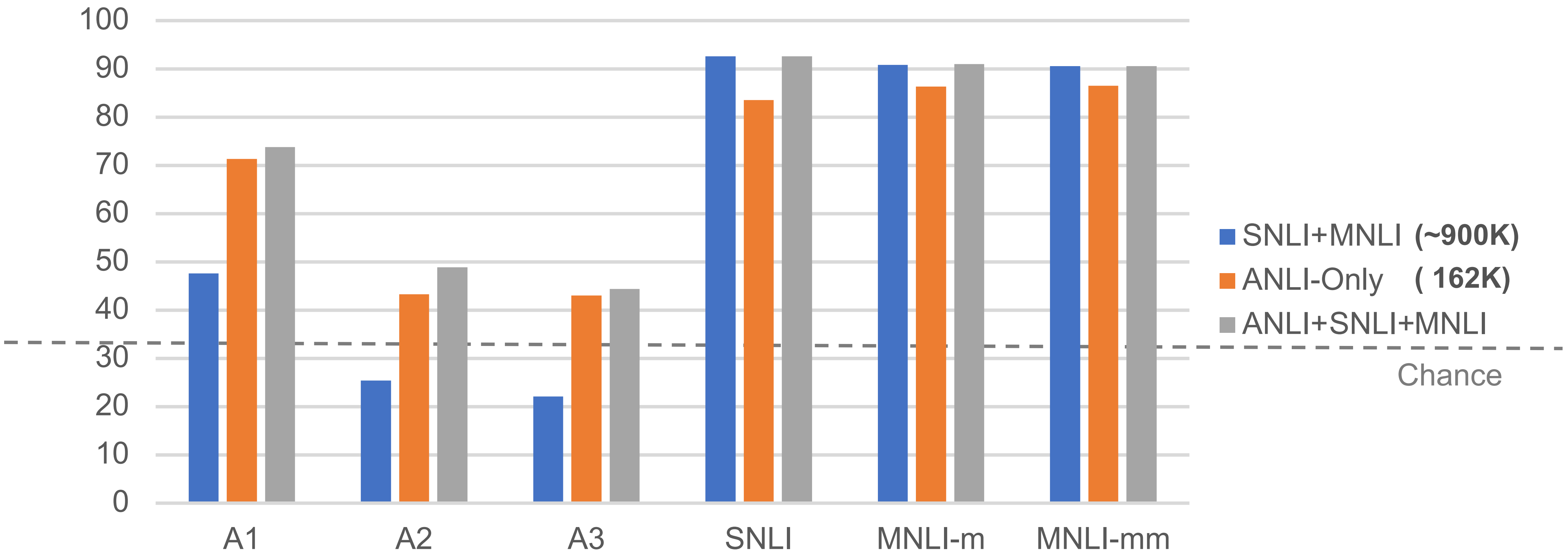
Model trained only on SNLI and MNLI (statically collected) is not good at ANLI

But Model trained only on ANLI (adversarially collected) is reasonably good at SNLI and MNLI

FACEBOOK AI

UNC NLP

RoBERTa performance with different training data



Model trained only on SNLI and MNLI (statically collected) is not good at ANLI

But model trained only on ANLI (adversarially collected) is reasonably good at SNLI and MNLI

Combining them together helps

NLI Stress Test

Model	SNLI-Hard	NLI Stress Tests					
		AT (m/mm)	NR	LN (m/mm)	NG (m/mm)	WO (m/mm)	SE (m/mm)
Previous models	72.7	14.4 / 10.2	28.8	58.7 / 59.4	48.8 / 46.6	50.0 / 50.2	58.3 / 59.4
BERT (All)	82.3	75.0 / 72.9	65.8	84.2 / 84.6	64.9 / 64.4	61.6 / 60.6	78.3 / 78.3
XLNet (All)	83.5	88.2 / 87.1	85.4	87.5 / 87.5	59.9 / 60.0	68.7 / 66.1	84.3 / 84.4
RoBERTa (S+M+F)	84.5	81.6 / 77.2	62.1	88.0 / 88.5	61.9 / 61.9	67.9 / 66.2	86.2 / 86.5
RoBERTa (All)	84.7	85.9 / 82.1	80.6	88.4 / 88.5	62.2 / 61.9	67.4 / 65.6	86.3 / 86.7

All=S+M+F+ANLI;
AT=Antonym; NR=Numerical Reasoning; LN=Length; NG=Negation; WO=Word Overlap SE=Spell Error

Training on ANLI is useful for the Antonym, Numerical Reasoning, and Negation.

Analysis

What kind of vulnerabilities do annotators find?

Round	Numerical & Quant.	Reference & Names	Standard	Lexical	Tricky	Reasoning & Facts	Quality
A1	38%	13%	18%	13%	22%	53%	4%
A2	32%	20%	21%	21%	20%	59%	3%
A3	10%	18%	27%	27%	27%	63%	3%
Average	27%	17%	22%	22%	23%	58%	3%

Type of inference in the data changed, and so are the model weaknesses.

Examples

Premise	Hypothesis	Reason	Model Prediction	Human Label	Linguistic Annotation
Kota Ramakrishna Karanth (born May 1, 1894) was an Indian lawyer and politician who served as the Minister of Land Revenue for the Madras Presidency from March 1, 1946 to March 23, 1947. He was the elder brother of noted Kannada novelist K. Shivarama Karanth.	Kota Ramakrishna Karanth has a brother who was a novelist and a politician.	Although Kota Ramakrishna Karanth's brother is a novelist, we do not know if the brother is also a politician	Entailment	Neutral	Standard Conjunction, Reasoning Plausibility Likely, Tricky Syntactic

Discussion

Discussion:

- HAMLET is model-agnostic. (Ensemble different backend models)
- It can be easily applied to any classification tasks.

What is underexplored?:

- How to extend the framework to generation tasks.
- Cost and time trade-off between adversarial and static data collection.

Summary

- NLU is far from solved;
- HAMLET (Human-And-Model-in-the-Loop-Enabled-Training);
- We applied it to NLI and collect ANLI;
- The procedure can provide more difficult and iterative benchmarks.

"… all of our models smaller than GPT-3 perform at almost exactly random chance on ANLI, even in the few-shot setting (~33%), whereas GPT-3 itself shows signs of life on Round 3."

GPT-3 performance on ANLI(A1/A2/A3): 36.8/34.0/40.2

Ideally, in its limit, HAMLET can help converge towards "real NLU"
Adversarial collecting & training help improve robustness

Thank you

Demo: https://adversarialnli.com/

GitHub: https://github.com/facebookresearch/anli/