

EMNLP 2021 Tutorial

Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection

Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar,
Sam Bowman, and Yoav Artzi

Introduction

Presented by Yoav Artzi and Sam Bowman



ML²



Penn
UNIVERSITY OF PENNSYLVANIA

Tutorial Goal

- Crowdsourcing is a fundamental tool in data-driven NLP research and practice
- Although critical, crowdsourcing often receives limited attention in papers and teaching materials
- Partially because general principles are elusive
- Instead use of crowdsourcing is guided by common practices and personal experience

This tutorial aims to be an educational resource through the discussion of a diverse set of case studies

Structure

- Brief background
- Five case studies
- Seven presenters
- Segmented videos



Alane



Nikita



Clara



Mark



Sam



Yoav



Maarten

Hybrid Presentation

- All videos are available online, and will be made public following EMNLP
- During the tutorial slot in EMNLP: a live clinic in person and via Zoom
- **So: best to watch the videos in advance!**
- The Zoom link is available via Underline

The EMNLP Crowdsourcing Clinic

- Taking place live in person and on Zoom during the tutorial slot in EMNLP → Wednesday, November 10, 8-11:30am EST (9:00-12:30pm conference time)
- We will start with about 30min of introduction and background
- The rest will be dedicated to question answering, discussion, and (best-effort) advice
- We are happy to discuss the case studies and **your own crowdsourcing scenarios**

Case Studies

Case studies display high diversity of task setups, reasoning, and data scale

- Inputs: single sentence, text interaction, image, and situated interaction
- Outputs: classification labels, span predictions, and generated sequences
- Sizes: 24k—570k examples

Case Study I: NLI

- Task: textual entailment recognition
- Text-only reasoning task
- Variants: single or multi-domains

Premise

someone else noticed it and i said
well i guess that's true and it was
somewhat melodious in other words
it wasn't just you know it was really
funny

Switchboard

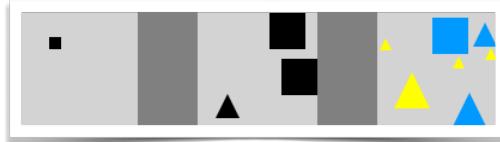
Hypothesis

No one noticed and it wasn't funny
at all.

Contradiction

Case Study II: NLVR

- Task: classify statement truth value with regard to a pair of images
- Multimodal: text and images
- Two variants: synthetic images or Internet photos



there are exactly three squares not touching any edge

FALSE



All dogs are corgis with upright ears, and one image contains at least twice as many real corgis as the other image.

TRUE

Case Study III: CerealBar

- Task: instruction execution/generation
- Situated collaborative interaction
- Game-like environment



Case Study IV: QUAc

- Task: span-based question answering in interaction context
- Text-only teacher-student chat

Section: Daffy Duck, Origin & History

STUDENT: **What is the origin of Daffy Duck?**
TEACHER: → first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: → assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ↛ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: → Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: → Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: → Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: → One often-repeated “official” story is that it was modeled after producer Leon Schlesinger’s tendency to lisp.
STUDENT: **Is there an “unofficial” story?**
TEACHER: → Yes, Mel Blanc (...) contradicts that conventional belief

...

Case Study V: SocialIQA

- Task: multiple choice question answering
- Domain: commonsense reasoning about social situations

REASONING ABOUT MOTIVATION

Tracy had accidentally pressed upon Austin in the small elevator and it was awkward.

Q Why did Tracy do this?

- A** (a) get very close to Austin
(b) squeeze into the elevator ✓
(c) get flirty with Austin

REASONING ABOUT WHAT HAPPENS NEXT

Alex spilled the food she just prepared all over the floor and it made a huge mess.

Q What will Alex want to do next?

- A** (a) taste the food
(b) mop up ✓
(c) run around in the mess

REASONING ABOUT EMOTIONAL REACTIONS

In the school play, Robin played a hero in the struggle to the death with the angry villain.

Q How would others feel afterwards?

- A** (a) sorry for the villain
(b) hopeful that Robin will succeed ✓
(c) like Robin should lose

See you in EMNLP!

 Punta Cana, Dominican Republic AST (UTC -4)	Wed, Nov 10, 2021	9:00 am
 New York, NY, USA EST (UTC -5)	Wed, Nov 10, 2021	8:00 am
 Seattle, WA, USA PST (UTC -8)	Wed, Nov 10, 2021	5:00 am
 Philadelphia, PA, USA EST (UTC -5)	Wed, Nov 10, 2021	8:00 am
 London, United Kingdom GMT (UTC +0)	Wed, Nov 10, 2021	1:00 pm
 Shanghai, China CST (UTC +8)	Wed, Nov 10, 2021	9:00 pm



*Scan to get
your time!*

EMNLP 2021 Tutorial

Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection

Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar,
Sam Bowman, and Yoav Artzi

Background

Sam Bowman



Platforms

Crowdsourcing Platforms

Amazon Mechanical Turk:

- Largest, oldest marketplace.
- Flexible—supports arbitrary custom code.
- Oriented toward 1–10m microtasks.
- Most workers in US or India, part-time, college educated.



Crowdsourcing Platforms

Upwork:

- Requesters hire workers individually and specifically.
- Oriented around longer gigs or hiring specialists.
- Higher typical pay—mostly >\$25 USD/h.
- Need data annotated by doctors?



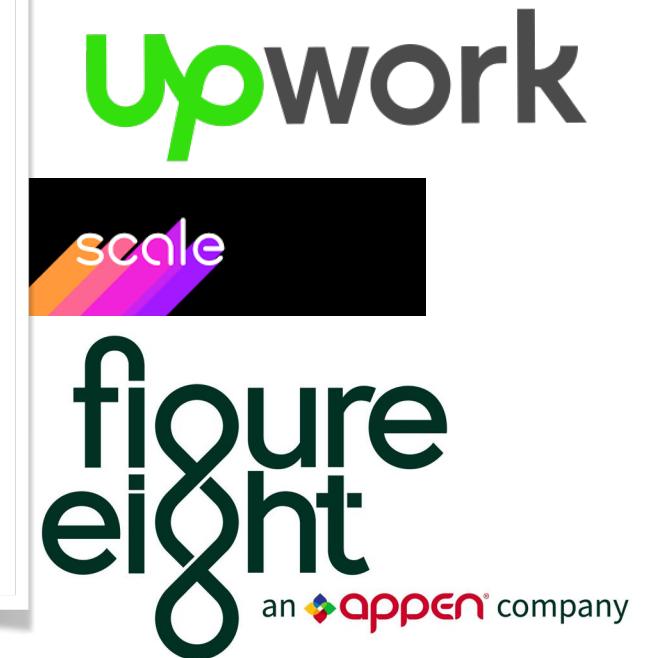
Crowdsourcing Platforms

Many others available!



Crowdsourcing Platforms

Many others available!



Mechanical Turk Basics

Mechanical Turk Basics

- Workers and requesters (i.e., researchers) join the platform. No training or experience required on either side.
- A requester designs a simple UI (often an HTML form) to collect data.
- The requester posts a batch of *human intelligence tasks* (HITs) using that UI, each representing individual small jobs that pay a fixed amount (\$1?), and deposits money.
- Over the following hours/days, workers choose HITs and complete them one-by-one.
- Requesters quickly review submitted work and approve it (at their sole discretion), releasing payment.

This doesn't quite work.

"Market for Lemons" (Akerlof 1970)

the quality of goods traded in a market can degrade in the presence of **information asymmetry** between buyers and sellers



“Market for Lemons”

What happens when a buyer cannot accurately judge the quality of an individual product prior to committing to its purchase?

The buyer will **average the quality of all similar products** in their decision as to how much to pay.



“Market for Lemons”

What happens when a buyer cannot accurately judge the quality of an individual product prior to committing to its purchase?

The buyer will average the quality of all similar products in their decision as to how much to pay.

Sellers will then be incentivized to lower the quality of their goods, since they will be paid an average price in any case, and thus **they can benefit more from each transaction if the payment they receive is greater than the value of what they gave in return**.



“Market for Lemons”

What happens when a buyer cannot accurately judge the quality of an individual product prior to committing to its purchase?

The buyer will average the quality of all similar products in their decision as to how much to pay.

Sellers will then be incentivized to lower the quality of their goods, since they will be paid an average price in any case, and thus they can benefit more from each transaction if the payment they receive is greater than the value of what they gave in return.

Good workers then tend to leave the market, because they get paid less than their actual value.



“Market for Lemons”

What happens when a buyer cannot accurately judge the quality of an individual product prior to committing to its purchase?

The buyer will average the quality of all similar products in their decision as to how much to pay.

Sellers will then be incentivized to lower the quality of their goods, since they will be paid an average price in any case, and thus they can benefit more from each transaction if the payment they receive is greater than the value of what they gave in return.

Good workers then tend to leave the market, because they get paid less than their actual value.



End result: market decreases in quality

Major Issues

AMT median hourly wage is **only ~\$2/hr** (current lowest US minimum wage **\$7.25/hr**)

- and only 4% earned more than \$7.25/hr
- average requester pays > \$11/hr
- lower-paying requesters post much more work

A Data-Driven Analysis of Workers' Earnings on AMT (2018)

Hara, Adams, Milland, Savage, Callison-Burch, Bigham

Mechanical Turk Tips

Mechanical Turk Basics

Amazon has largely given up on maintaining Mechanical Turk, but it's still an extremely active marketplace and standard in NLP.

Workarounds are often needed.

Recruiting Trustworthy Workers

- Amazon lets you filter by experience level: Common to limit HITs to experienced workers (>5,000 HITs completed) with low rejection rates (<2%).
 - Be careful about needlessly high HIT counts: They push newer good workers into underpaid work (Kummerfeld '21 ACL).
- Amazon also lets you recruit its promoted 'Master' workers. This is meaningless.

Qualifications

- You can assign manual *qualifications* to workers. Common setup:
 - Post a public training/practice HIT that workers can only do once.
 - Manually review work on that HIT, and use it to grant qualifications to work on the rest of the HITs.
 - Periodically monitor work, and revoke qualifications if major problems arise.
- *Don't* reject work unless it's very clearly spam/fraud. This revokes payment for work that has already been done.

Quality Control

- Use multiple HITs to ensure reasonable quality in test/validation data:
 - When collecting test data for classification and annotation tasks, have several workers annotate each example.
 - Fancy statistical methods can aggregate multiple annotations better than majority vote.
 - When multiple parallel annotations can't be combined, consider building a second validation HIT to double check each data point.

Building a UI

- For simple tasks, Amazon has simple HTML form templates you can edit, and it will let you upload/download CSVs with data.
 - You can use simple javascript snippets to validate responses and add other simple interactive features.
- For more complex/interactive tasks, there are more powerful tools that integrate with MTurk for things like dialog, QA example creation.

Reputation

- Ambitious/reliable workers use forums (esp. TurkerNation) and plugins (TurkOpticon/Crowd-Workers) to find HITs.
- These are probably the workers you want to hire, so your reputation there matters.

Reputation

- To maintain a good reputation:
 - Pay well.
 - Have clear, fair criteria for bonuses and rejection (typically in an FAQ doc).
 - Respond to worker questions quickly—daily at least.
 - Design your HITs to be usable and efficient.
 - Identify yourself clearly.
 - Give clear instructions, especially for how to handle weird/broken prompts. (Link to an FAQ.)
 - Make sure HITs that pay the same rate take roughly the same amount of time.

Hourly Wage Estimation

- Amazon's hourly wage estimate tool isn't trustworthy.
- To start:
 - Do the work yourself for an hour and see how far you get.
- Once your HIT is live:
 - Tools like Crowd-Workers and TurkOpticon let you see better estimates of actual time elapsed. (And let you see your reputation!)

Research

- In papers on data collection, it's increasingly standard to list the effective rate that you paid. Make sure this isn't embarrassing or illegal.
- If you're *studying* your workers (more common in Linguistics-style projects), you're based at a university, and you try to publish your work, you'll have to show an approved study protocol number from your university's institutional review board.

EMNLP 2021 Tutorial

Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection

Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar,
Sam Bowman, and Yoav Artzi

Case Study I: NLI

Presented by Nikita Nangia, Clara Vania, and Sam Bowman



ML²



Penn
UNIVERSITY OF PENNSYLVANIA

Natural Language Inference aka Recognizing Textual Entailment

Premise: *I'm watching an EMNLP talk.*

Hypothesis: *I'm having loads of fun!*

Label: {entailment, contradiction, neutral}

Why NLI?

NLU benchmarking and (previously) transfer learning.

- It lets you test sentence understanding comprehensively *without* grounding or semantic formalisms.
- It caught on as a benchmark task, and played a significant role in the development of self-attention and pretraining.
- It's also been useful as a *pretraining* task: Fine-tuning BERT/RoBERTa/T5/etc. on NLI data makes it easier for that model to adapt to future tasks.
 - Less clear with the latest large models.

Stanford NLI & Multi-Genre NLI

Sam Bowman

Initial Efforts: The SNLI Data Collection Prompt

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.

Definitely correct Example: For the caption "*Two dogs are running through a field.*" you could write "*There are animals outdoors.*"

Write a sentence that follows from the given caption.

Maybe correct Example: For the caption "*Two dogs are running through a field.*" you could write "*Some puppies are running to catch a stick.*"

Write a sentence which may be true given the caption, and may not be.

Definitely incorrect Example: For the caption "*Two dogs are running through a field.*" you could write "*The pets are sitting on a couch.*" This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

Problems (optional) If something is wrong, have a look at the [FAQ](#), do your best above, and let us know here.

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.

Definitely correct Example: For the caption "*Two dogs are running through a field.*" you could write "*There are animals outdoors.*"

Write a sentence that follows from the given caption.

Entailment

Maybe correct Example: For the caption "*Two dogs are running through a field.*" you could write "*Some puppies are running to catch a stick.*"

Write a sentence which may be true given the caption, and may not be.

Definitely incorrect Example: For the caption "*Two dogs are running through a field.*" you could write "*The pets are sitting on a couch.*" This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

Problems (optional) If something is wrong, have a look at the [FAQ](#), do your best above, and let us know here.

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.

Definitely correct Example: For the caption "*Two dogs are running through a field.*" you could write "*There are animals outdoors.*"

Write a sentence that follows from the given caption.

Entailment

Maybe correct Example: For the caption "*Two dogs are running through a field.*" you could write "*Some puppies are running to catch a stick.*"

Write a sentence which may be true given the caption, and may not be.

Neutral

Definitely incorrect Example: For the caption "*Two dogs are running through a field.*" you could write "*The pets are sitting on a couch.*" This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

Problems (optional) If something is wrong, have a look at the [FAQ](#), do your best above, and let us know here.

Instructions

The [Stanford University NLP Group](#) is collecting data for use in research on computer understanding of English. We appreciate your help! We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.

Definitely correct Example: For the caption "*Two dogs are running through a field.*" you could write "*There are animals outdoors.*"

Write a sentence that follows from the given caption.

Entailment

Maybe correct Example: For the caption "*Two dogs are running through a field.*" you could write "*Some puppies are running to catch a stick.*"

Write a sentence which may be true given the caption, and may not be.

Neutral

Definitely incorrect Example: For the caption "*Two dogs are running through a field.*" you could write "*The pets are sitting on a couch.*" This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

Contradiction

Problems (optional) If something is wrong, have a look at the [FAQ](#), do your best above, and let us know here.

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

Your job is to figure out, based on the correct caption for a photo, if another caption is also correct:

- Choose **definitely correct** if any photo that was captioned with the caption on the left would also fit the caption on the right.
Example: "*A kitten with spots is playing with yarn.*" / "*A cat is playing.*"
- Choose **maybe correct** if the second caption could describe photos that fit the first caption, but could also describe sentences that don't fit the first caption. Example: "*A kitten with spots is playing with yarn.*" / "*A kitten is playing with yarn on a sofa.*"
- Choose **definitely incorrect** if any photo that could possibly be captioned with the caption on the left would not fit the caption on the right. Example: "*A kitten with spots is playing with yarn.*" / "*A puppy is playing with yarn.*"

We have already labeled one out of every 250 HITs. Completing one of these HITs yields a bonus of \$1 for each response that matches our label for up to \$5. More questions? See the [FAQ](#).

Correct caption	Candidate caption	Def. correct	Maybe correct	Def. incorrect
\${caption1}	\${sentence1}	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
\${caption2}	\${sentence2}	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
\${caption3}	\${sentence3}	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
\${caption4}	\${sentence4}	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
\${caption5}	\${sentence5}	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Problems (optional) If something is wrong with a caption that makes it hard to understand (more than just a typo), do your best above and let us know here.

Subset of examples validated by four annotators.
The final label is the majority vote of the five.

Crowdwork Setting

- Amazon Mechanical Turk
- Qualification: 5000 HIT \times 98% acceptance rate
- No other qualification process, spot-checking for disqualifications
- Average HIT ~\$0.16, based on our initial time estimates
 - Looking at clearer timing information from more recent studies, this was way too low: Likely under \$6/hr for slower good workers!

Crowdwork Setting

- Validation/relabeling task:
 - Used for development and test data, plus small section of training data.
 - Mostly done with a *private* qualification.
 - \$0.10 for five pairs. Again, likely too low.
 - Small fraction of HITs (0.1%-0.5%) give a bonus (\$1?) for agreement with our judgments on each pair.

Crowdwork Setting

- Linked FAQ document
- 'Problems' field in HITs, mostly for misformatted inputs.
- ~1-10 email questions/day.
- Some inter-annotator communication through private qualification group.

What we got

Some sample results

Premise: *Two women are embracing while holding two packages.*

Hypothesis: *Two women are holding packages.*

Label: Entailment

Some sample results

Premise: *A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.*

Hypothesis: *A man is repainting a garage*

Label: Neutral

Nutrition Facts

Data set sizes:

Training pairs	550,152
Development pairs	10,000
Test pairs	10,000

Sentence length:

Premise mean token count	14.1
Hypothesis mean token count	8.3

Parser output:

Premise ‘S’-rooted parses	74.0%
Hypothesis ‘S’-rooted parses	88.9%
Distinct words (ignoring case)	37,026

General:

Validated pairs	56,951
Pairs w/ unanimous gold label	58.3%

Individual annotator label agreement:

Individual label = gold label	89.0%
Individual label = author’s label	85.8%

Gold label/author’s label agreement:

Gold label = author’s label	91.2%
Gold label \neq author’s label	6.8%
No gold label (no 3 labels match)	2.0%

Fleiss κ :

<i>contradiction</i>	0.77
<i>entailment</i>	0.72
<i>neutral</i>	0.60
Overall	0.70

MultiNLI

Sam Bowman

MultiNLI

- Same intended definitions for labels, but no longer specialized to photos.
- More genres—not just concrete visual scenes.
 - Includes transcribed speech, business documents, etc.
- More detailed guidelines
 - Need to cover, e.g., question-question relationships.

Crowdwork Setting

- Same overall design.
- hybrid.io: Short-lived MTurk competitor
- Private qualification for all HITs.
- Incrementally higher pay than SNLI—exact numbers lost.
Total cost ~\$60,000 for ~400k examples.

What we got

Typical Development Set Examples

Premise: *someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny*

Hypothesis: *No one noticed and it wasn't funny at all.*

Label: Contradiction

Genre: Switchboard (telephone speech)

Typical Development Set Examples

Premise: *In contrast, suppliers that have continued to innovate and expand their use of the four practices, as well as other activities described in previous chapters, keep outperforming the industry as a whole.*

Hypothesis: *The suppliers that continued to innovate in their use of the four practices consistently underperformed in the industry.*

Label: Contradiction

Genre: Oxford University Press (academic books)

Key Linguistic Phenomena

Tag	SNLI	MultiNLI
Pronouns (PTB)	34	68
Quantifiers	33	63
Modals (PTB)	<1	28
Negation (PTB)	5	31
‘Wh’ Words (PTB)	5	30
Belief Verbs	<1	19
Time Terms	19	36
Conversational Pivots	<1	14
Presupposition Triggers	8	22
Comparatives/Superlatives (PTB)	3	17
Conditionals	4	15
Tense Match (PTB)	62	69
Interjections (PTB)	<1	5
>20 Words	<1	5
Existentials (PTB)	5	8

Known Issues

Clara Vania

Annotation Artifacts

For SNLI:

P: ???

H: *Someone is **not** crossing the road.*

Label: entailment, contradiction, neutral?

Annotation Artifacts

For SNLI:

P: ???

H: *Someone is not crossing the road.*

Label: entailment, **contradiction**, neutral?

Annotation Artifacts

For SNLI:

P: ???

H: *Someone is not crossing the road.*

Label: entailment, **contradiction**, neutral?

P: ???

H: *Someone is outside.*

Label: entailment, contradiction, neutral?

Annotation Artifacts

For SNLI:

P: ???

H: *Someone is not crossing the road.*

Label: entailment, **contradiction**, neutral?

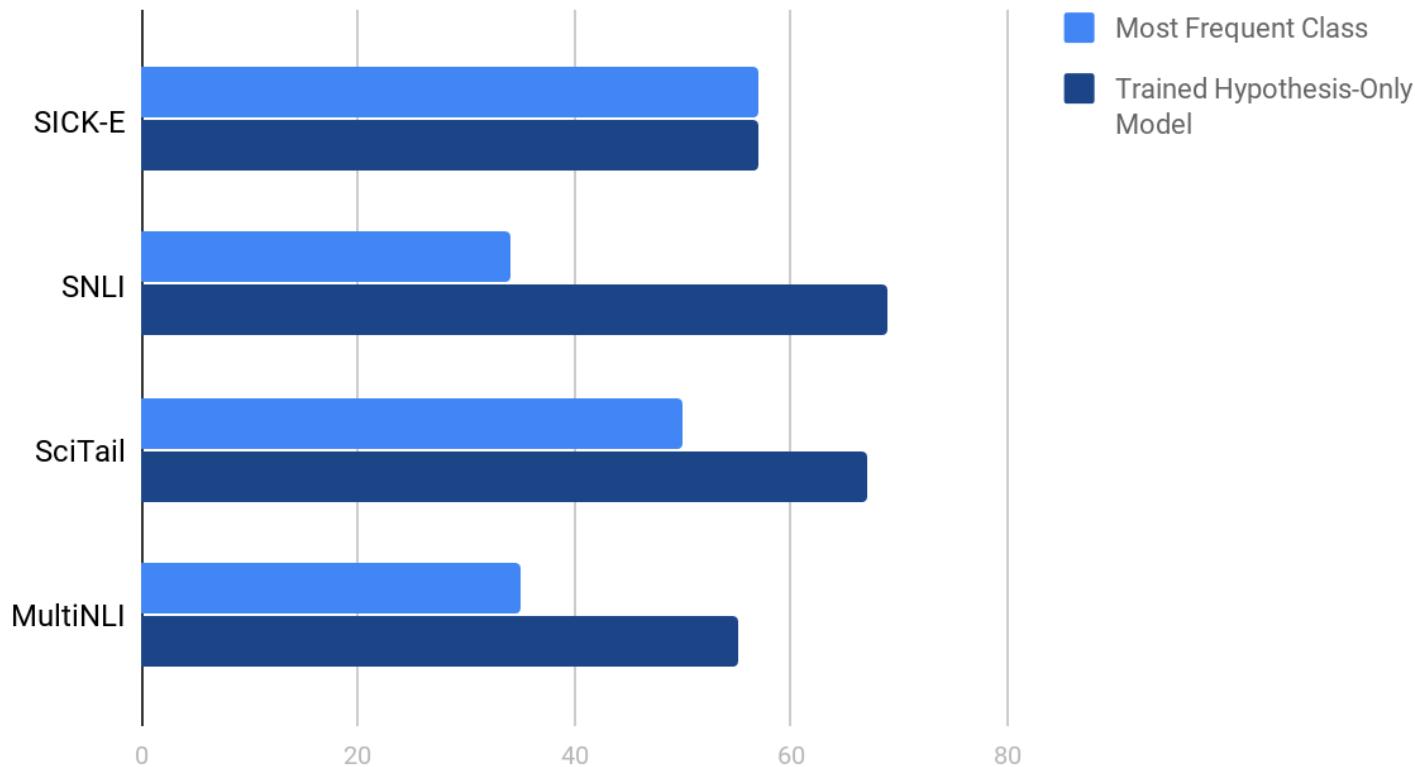
P: ???

H: *Someone is outside.*

Label: entailment, contradiction, neutral?

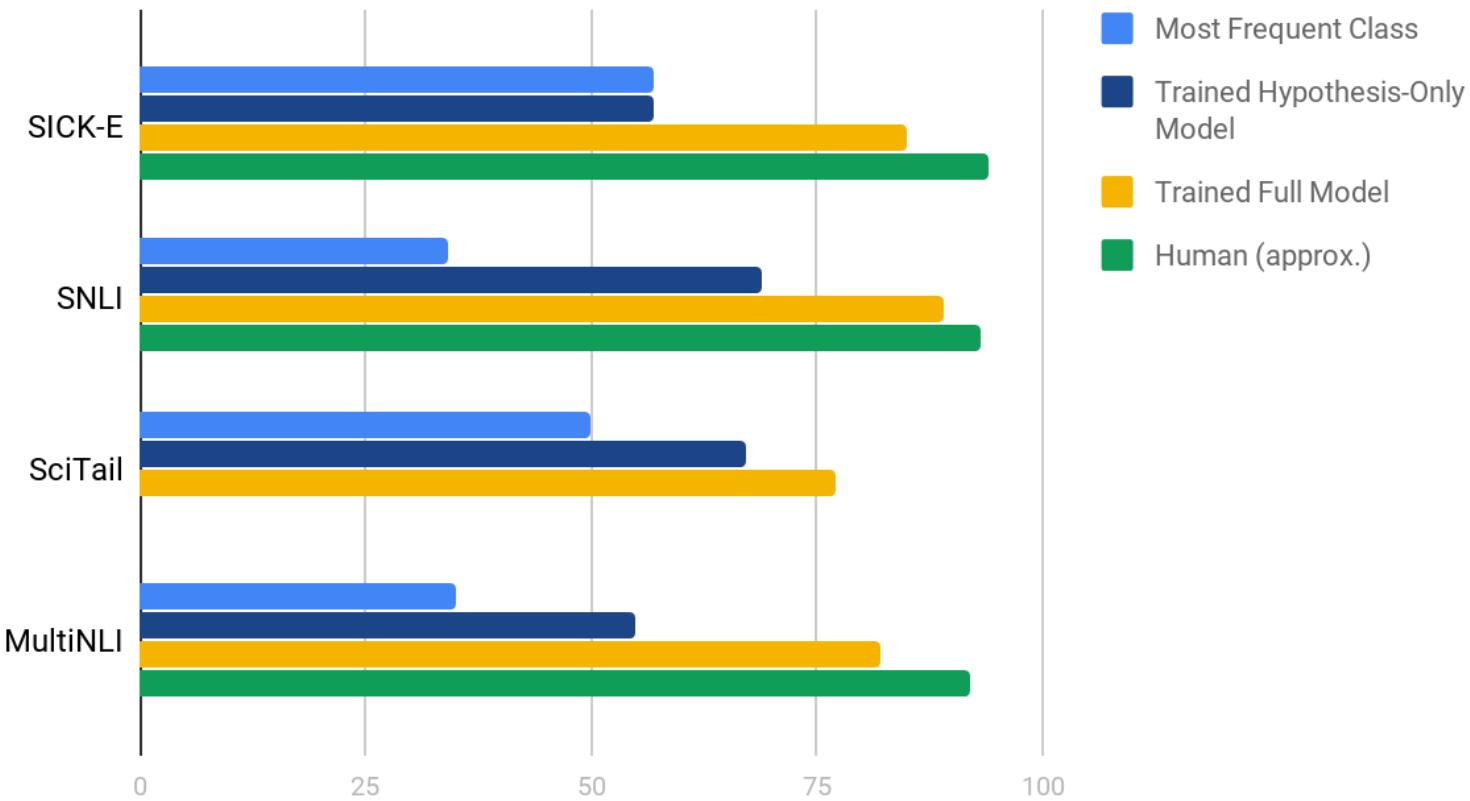
Annotation Artifacts

Models can do moderately well on NLI datasets without looking at the premise!



Single-genre SNLI especially vulnerable. SciTail not immune, despite using no crowdworker writing.

Annotation Artifacts



...but hypothesis-only models are still far below ceiling.

[Poliak et al. '18](#), [Tsuchiya '18](#), [Gururangan et al. '18](#)

Social Bias

SNLI data demonstrates social stereotypes that we won't want models to use in many settings, both from the distribution of the original Flickr photos and from the crowdworkers.

woman	hairdresser [‡] fairground grieving receptionist widow
women	actresses [†] husbands [‡] womens [‡] gossip [‡] wemon [‡]
girl	schoolgirl piata cindy pigtails [‡] gril
girls	fifteen [‡] slumber sking [‡] jumprope [†] ballerinas [‡]
mother	kissed [‡] parent [‡] mom [‡] feeds daughters

Top hypothesis terms by PMI with the given premise term.

Crowdsourcing Experiments & Alternative Protocols

Adversarial Data Collection

Clara Vania

Collect a Large Benchmark That Can Last Longer

Adversarial human-and-model-in-the loop

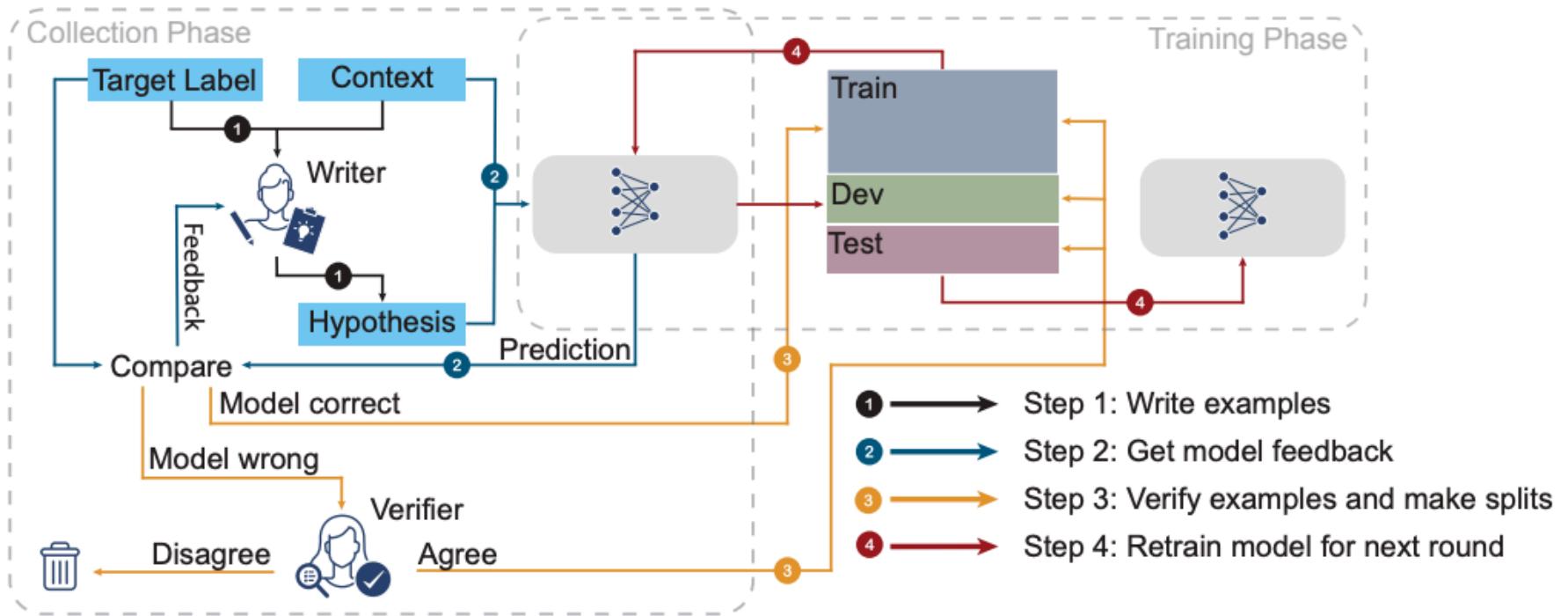


Figure 1: Adversarial NLI data collection via human-and-model-in-the-loop enabled training (HAMLET). The four steps make up one round of data collection. In step 3, model-correct examples are included in the training set; development and test sets are constructed solely from model-wrong verified-correct examples.

Model Performance on is Low Compared to SNLI/MNLI

Rounds Become Increasingly More Difficult

Train Data	A1	A2	A3	S	M-m/mm
ALL	73.8	48.9	44.4	92.6	91.0/90.6
S+M	47.6	25.4	22.1	92.6	90.8/90.6
ANLI-Only	71.3	43.3	43.0	83.5	86.3/86.5
ALL ^H	49.7	46.3	42.8	71.4	60.2/59.8
S+M ^H	33.1	29.4	32.2	71.8	62.0/62.0
ANLI-Only ^H	51.0	42.6	41.5	47.0	51.9/54.5

Table 6: Performance of RoBERTa with different data combinations. ALL=S,M,F,ANLI. Hypothesis-only models are marked H where they are trained and tested with only hypothesis texts.

Model trained on ANLI-Only is quite good at SNLI & MNLI
Training model on all training data obtains the best performance

Hypothesis-only Models Also Perform Poorly

Fewer Annotation Artifacts?

Train Data	A1	A2	A3	S	M-m/mm
ALL	73.8	48.9	44.4	92.6	91.0/90.6
S+M	47.6	25.4	22.1	92.6	90.8/90.6
ANLI-Only	71.3	43.3	43.0	83.5	86.3/86.5
ALL ^H	49.7	46.3	42.8	71.4	60.2/59.8
S+M ^H	33.1	29.4	32.2	71.8	62.0/62.0
ANLI-Only ^H	51.0	42.6	41.5	47.0	51.9/54.5

Table 6: Performance of RoBERTa with different data combinations. ALL=S,M,F,ANLI. Hypothesis-only models are marked H where they are trained and tested with only hypothesis texts.

In rounds 2 and 3, model performs similarly as hypothesis-only.

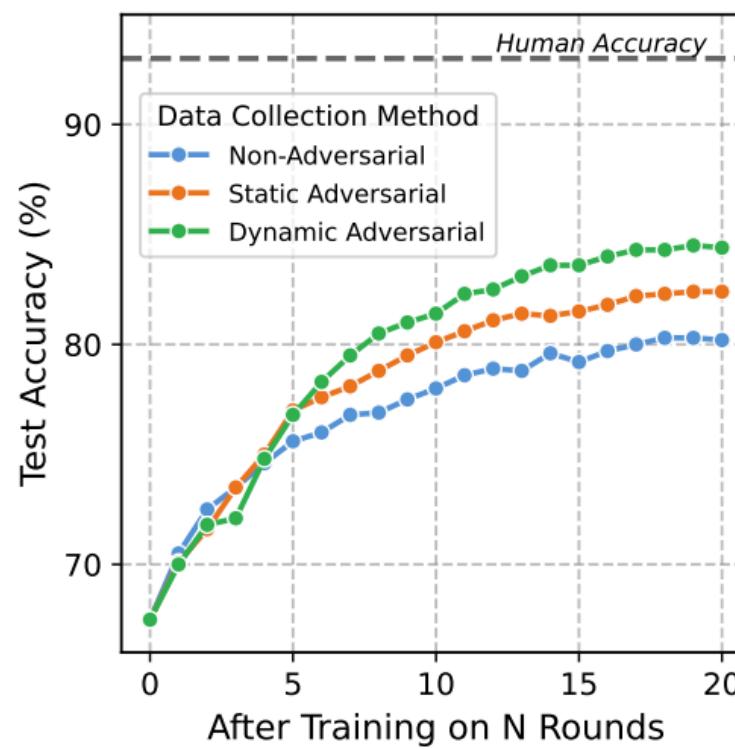
More Robust Models

Does Training on Adversarial Data Help?

- Compare fine-tuning on standard vs. single-round adversarial data on similar and out-of-domain data distributions.
- Adversarial training data does not offer clear benefits robustness under distribution shift.
- Fine-tuning on adversarial data leads to better performance on previously collected adversarial data.
- But worse performance on out-of-domain datasets, compared to fine-tuning model on standard data.

More Robust Models

Training Over Many Rounds Maximizes the Benefits of Adversarial Data Collection



Collected data from later rounds also seem to be more diverse with fewer annotation artifacts.

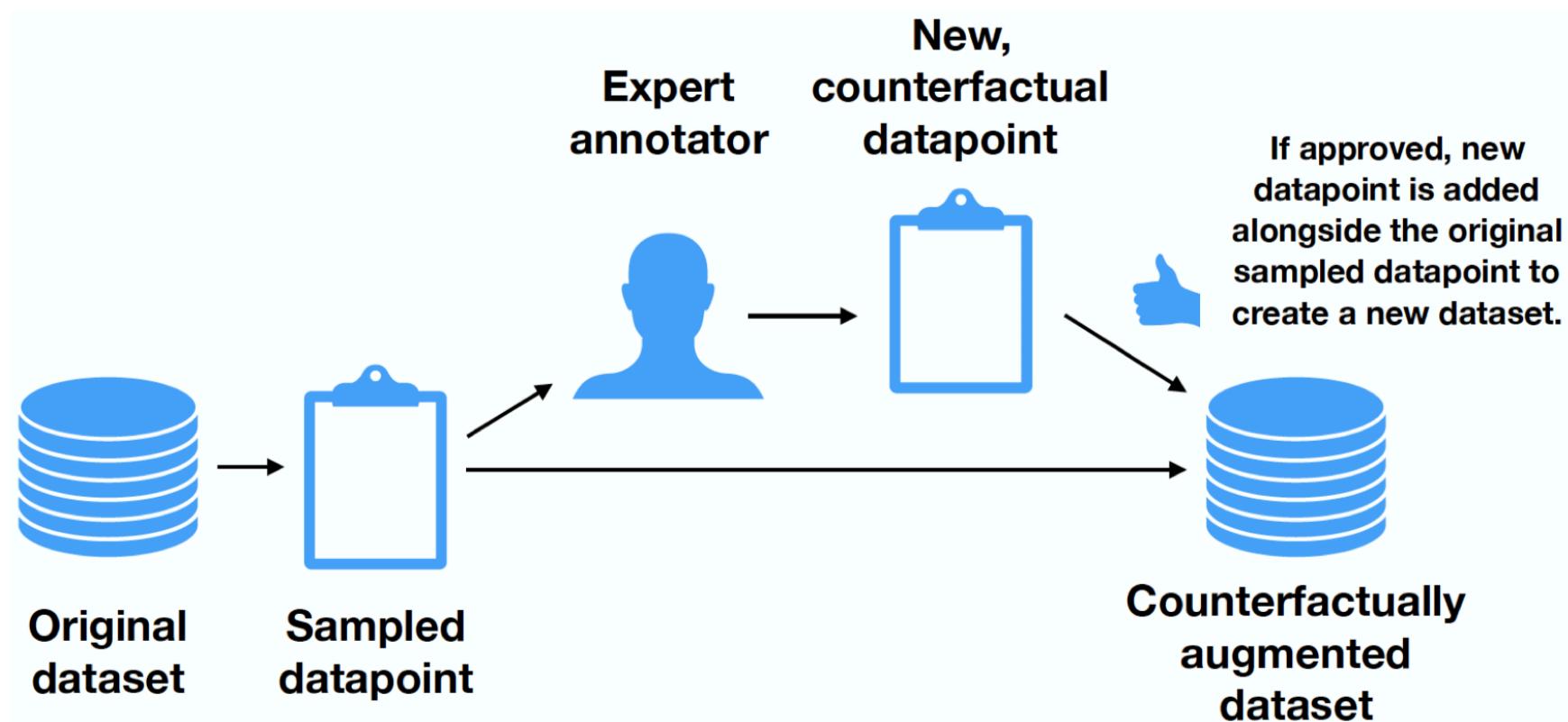
Wallace et al., '21

Counterfactual Data Augmentation

Sam Bowman

We can Target Artifacts by Editing Examples

Minimally Edit Existing Examples to Change the Label



We can Target Artifacts by Editing Examples

Minimally Edit Existing Examples to Change the Label

OP: An elderly **woman** in a crowd pushing a wheelchair. (Entailment)

NP: An elderly **person** in a crowd pushing a wheelchair. (Neutral)

H: There is an elderly woman in a crowd.

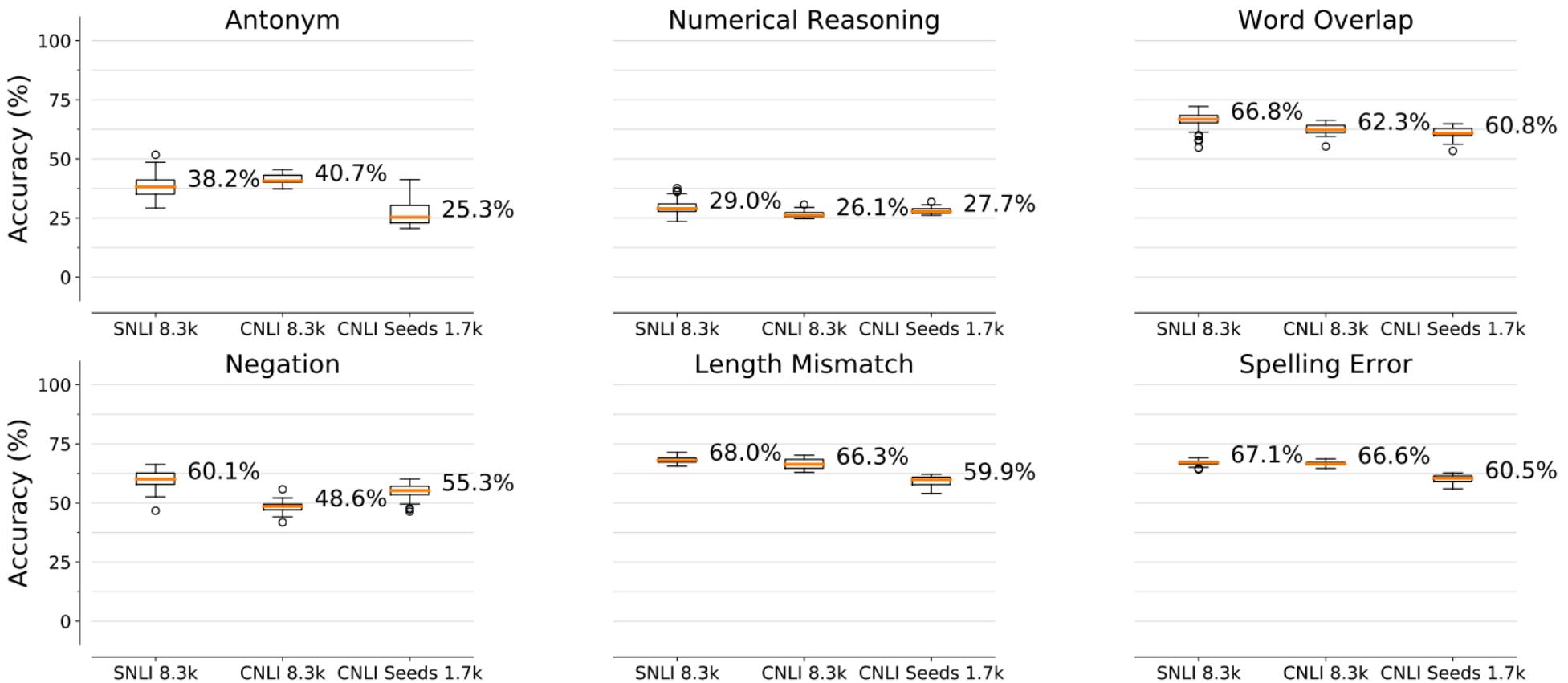
Significant Drops to SNLI Hypothesis-Only Accuracy

...suggesting fewer artifacts.

Train/Test	Original	RP	RH	RP & RH
Majority class	34.7	34.6	34.6	34.6
RP & RH (6.6k)	32.4	35.1	33.4	34.2
Original w/ RP & RH (8.3k)	44.0	25.8	43.2	34.5
Original (8.3k)	60.2	20.5	46.6	33.6
Original (500k)	69.0	15.4	53.2	34.3

...but no consistent improvements to robustness

(Tested on Naik et al.'s probing sets and others.)



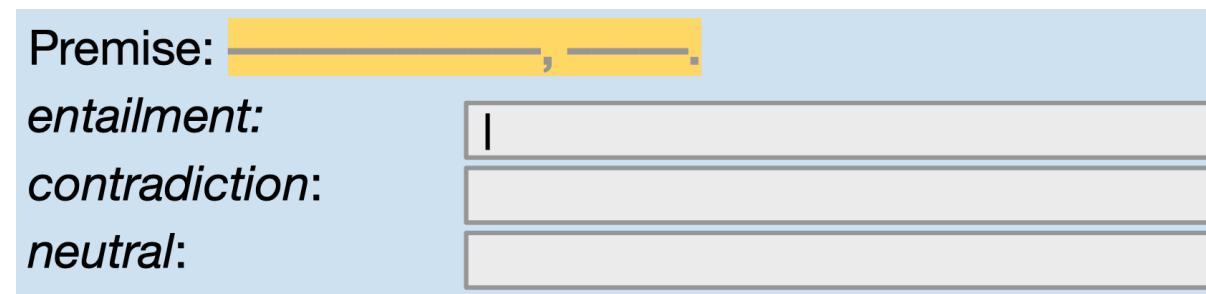
Four New Task Designs

Sam Bowman

We've basically tried only one task design so far.

It's not great.

- The three biggest training datasets (SNLI, MNLI, ANLI) were all crowdsourced, using essentially the same setup:



- No research went into this design. My advisors and I just thought it was a reasonable thing to try...
- Can alternate protocols help with artifacts or robustness?

Longer Premises

More text means 'simple' heuristics look less simple?

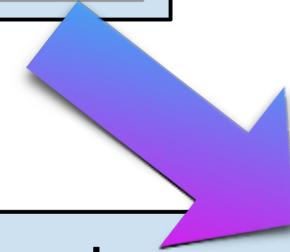
Base

Premise: ,

entailment:

contradiction:

neutral:



Paragraph

Premise: , . (—).
 , , . , .
 . . , ,

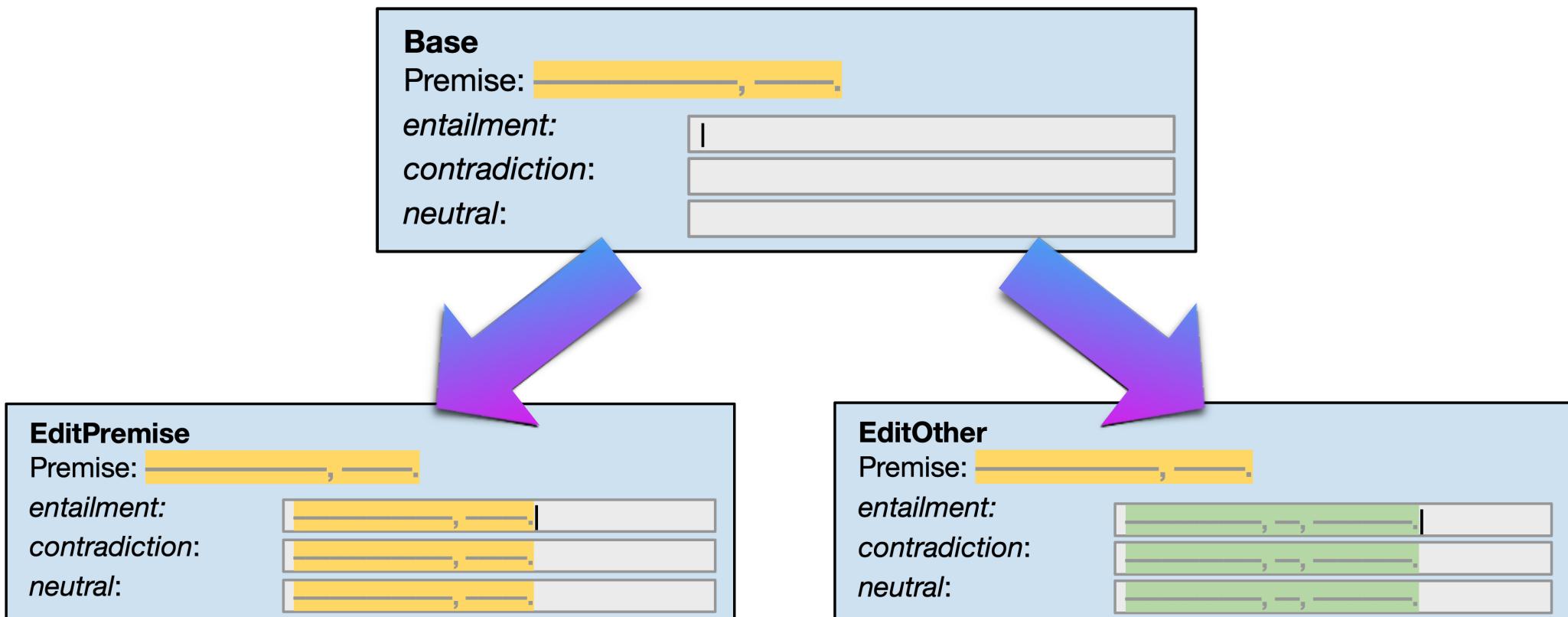
entailment:

contradiction:

neutral:

Pre-Filled Starter Text

Asking for *minimal edits* will eliminate the source of spurious artifacts?



Contrastive Writing

Making sure each entailment is *not* entailed by some other sentence makes many artifacts unlikely? And maybe encourages creativity?

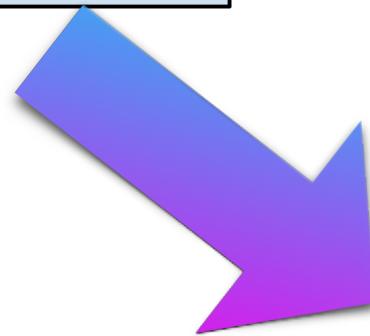
Base

Premise: , .

entailment:

contradiction:

neutral:



Contrast

Main Premise: , .

Contrasting Premise: , , .

entailment:

contradiction:

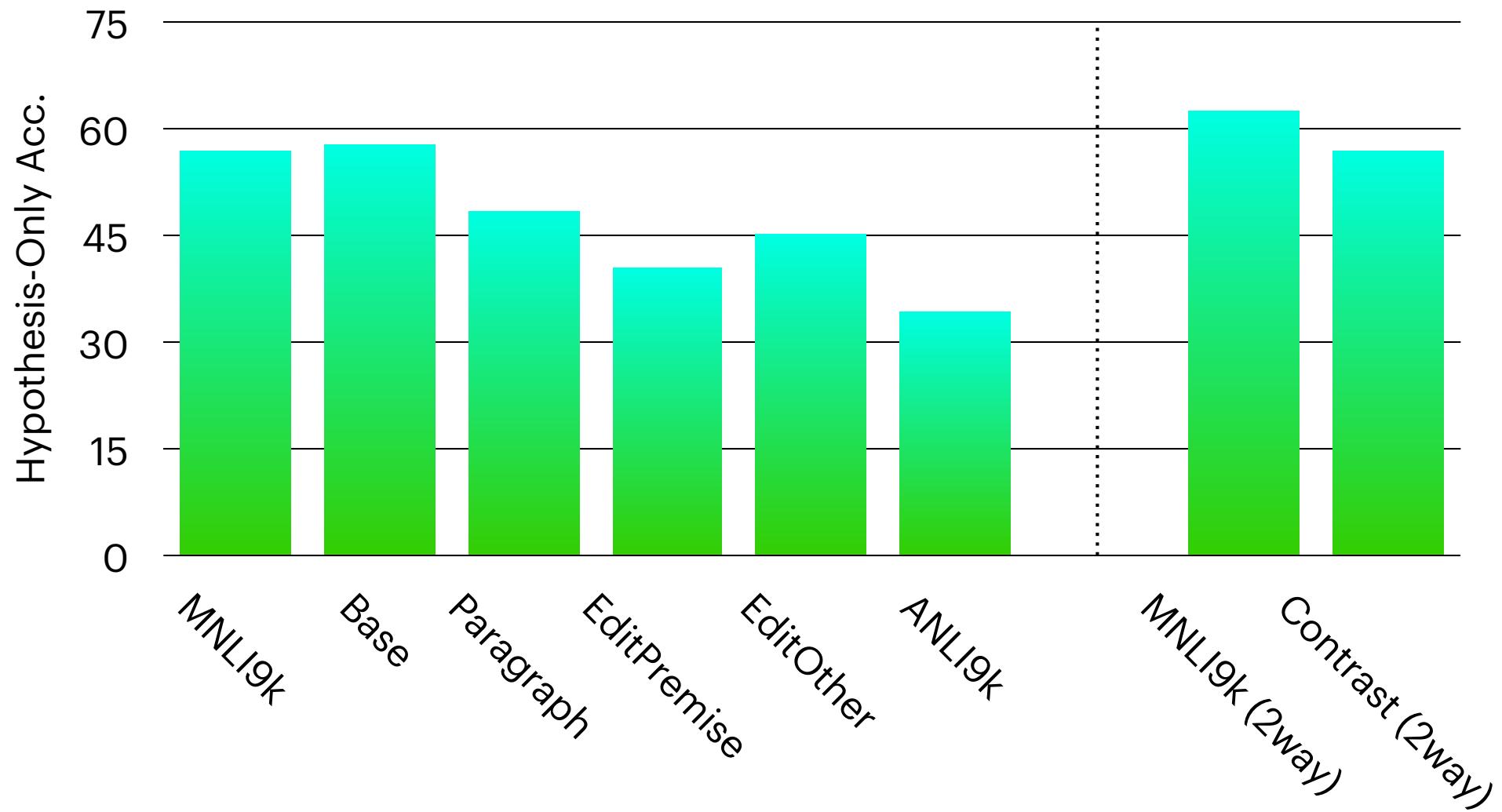
The Experiment

We collected five mid-sized datasets.

- All four new protocols produce subjectively reasonable data—similar to baseline.
- All four new protocols took about the same amount of time per premise as the baseline.
 - Switching from writing to editing text *didn't save time*.

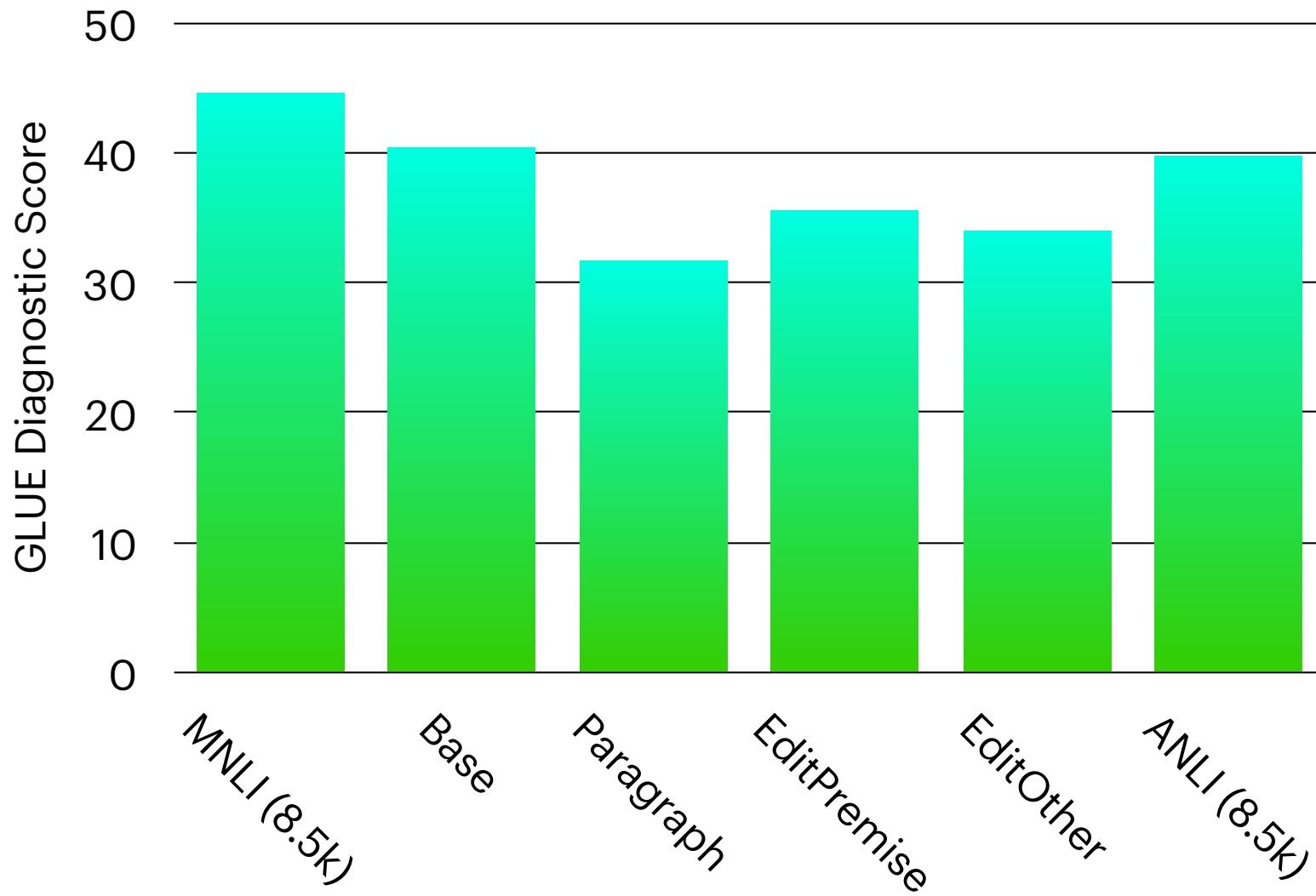
The Results

Artifacts don't seem to be as bad with any of the new protocols!



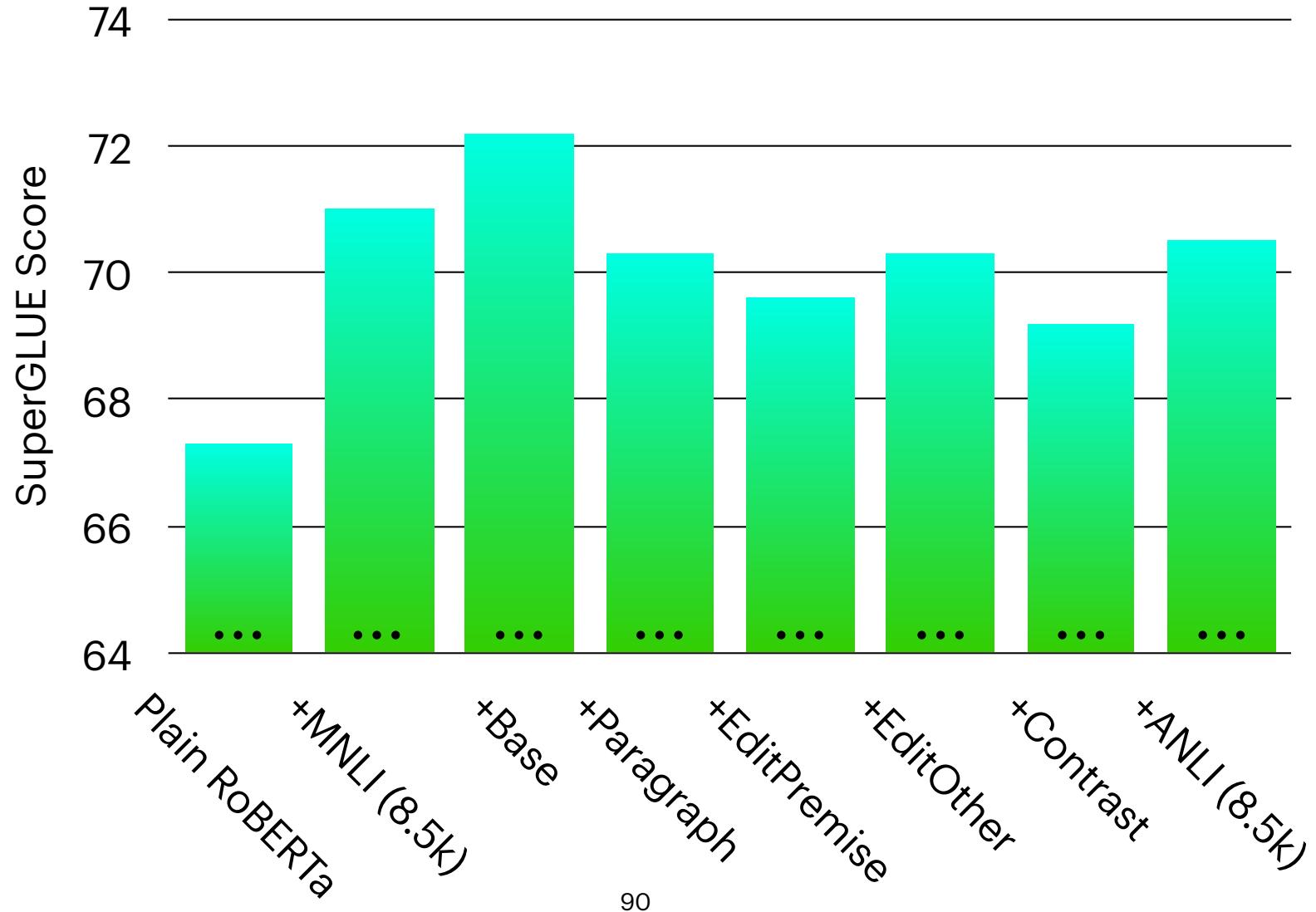
The Results

...but out-of-domain generalization is also a lot worse.



The Results

...and transfer learning performance is also a lot worse.



Takeaways!

- The basic crowdworker-writing protocol for MNLI is hard to beat.
 - Why? We're not sure. Creativity seems to matter.
 - Aside: Even ANLI is no better than MNLI as *training data*, controlling for dataset size.

Many-Way Validation & Annotator Disagreement

Sam Bowman

Disagreements can be Genuine

...and persist with large numbers of annotators

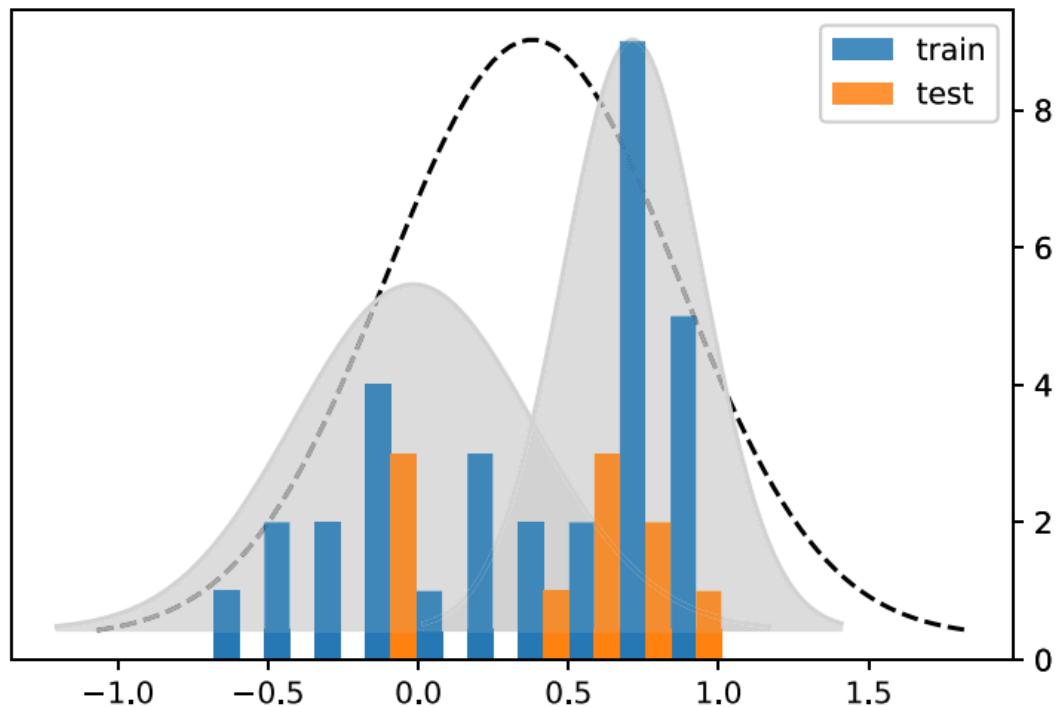
- NLI label choice is sensitive to individual variation in annotators' language use and understanding of the world.
 - Disagreements between annotators don't always mean that some annotator is wrong.
- Experiment: Reannotate a sample of NLI data from five sources on a continuous scale 50x each...
 - ...look for examples with multimodal label distributions.
 - About 20% of examples show significant multimodality!

Disagreements can be Genuine

...and persist with large numbers of annotators

p: Paula swatted the fly.

h: The swatting happened in a
forceful manner.



Disagreements can be Genuine

...and persist with large numbers of annotators

- Paragraph-length premises yield *more* multimodality.
- Models trained on singly-labeled data *do not* capture multimodal behavior in their output distributions.

Disagreements can be Genuine

...and persist with large numbers of annotators

- ChaosNLI: Largest resource with many-way annotations
 - 100x annotations for 4,600 NLI examples.
 - Most model errors are on lower-agreement examples.
 - ALBERT reaches 90–95% accuracy on high-agreement cases!
 - Most remaining headroom involves ambiguity/subjectivity.

OCNLI

Clara Vania

New Strategies for Eliciting Diverse Hypothesis

Multi-Hypothesis Elicitation

- Writer is asked to write three sentences per label (total nine hypotheses per premise)
 - Difficulty: easy (1st), medium (2nd), hard (3rd)
 - More challenging and higher inter-annotator agreement than the single-hypothesis writing

New Strategies for Eliciting Diverse Hypothesis

Control for Hypothesis-Only Bias

- Encourage annotators to write more diverse hypothesis
 - Tell them which types of data are expected, e.g., contradiction without negator, diverse way of inferences
 - Give incentives if written hypothesis matches the given criteria
- Put constraints on hypothesis generation
 - Only one contradiction can contain a negator
 - No hypothesis should overlap with the premise > 70%

Experiment Results

	SINGLE	MULTI	MULTIENC	MULTICON
BERT: fine-tune on XNLI				
dev_full	77.3	73.6	68.6	65.8
easy	na.	74.0	70.1	68.4
medium	na.	74.3	69.6	65.9
hard	na.	72.5	66.2	63.1
RoBERTa: fine-tune on XNLI				
dev_full	78.9	77.3	71.3	70.8
easy	na.	77.2	72.8	73.5
medium	na.	78.6	71.7	70.2
hard	na.	76.2	69.4	68.7

Multi-hypothesis elicitation yields more challenging data.

Experiment Results

	SINGLE	MULTI	MULTIENC	MULTICON
BERT: fine-tune on XNLI				
dev_full	77.3	73.6	68.6	65.8
easy	na.	74.0	70.1	68.4
medium	na.	74.3	69.6	65.9
hard	na.	72.5	66.2	63.1
RoBERTa: fine-tune on XNLI				
dev_full	78.9	77.3	71.3	70.8
easy	na.	77.2	72.8	73.5
medium	na.	78.6	71.7	70.2
hard	na.	76.2	69.4	68.7

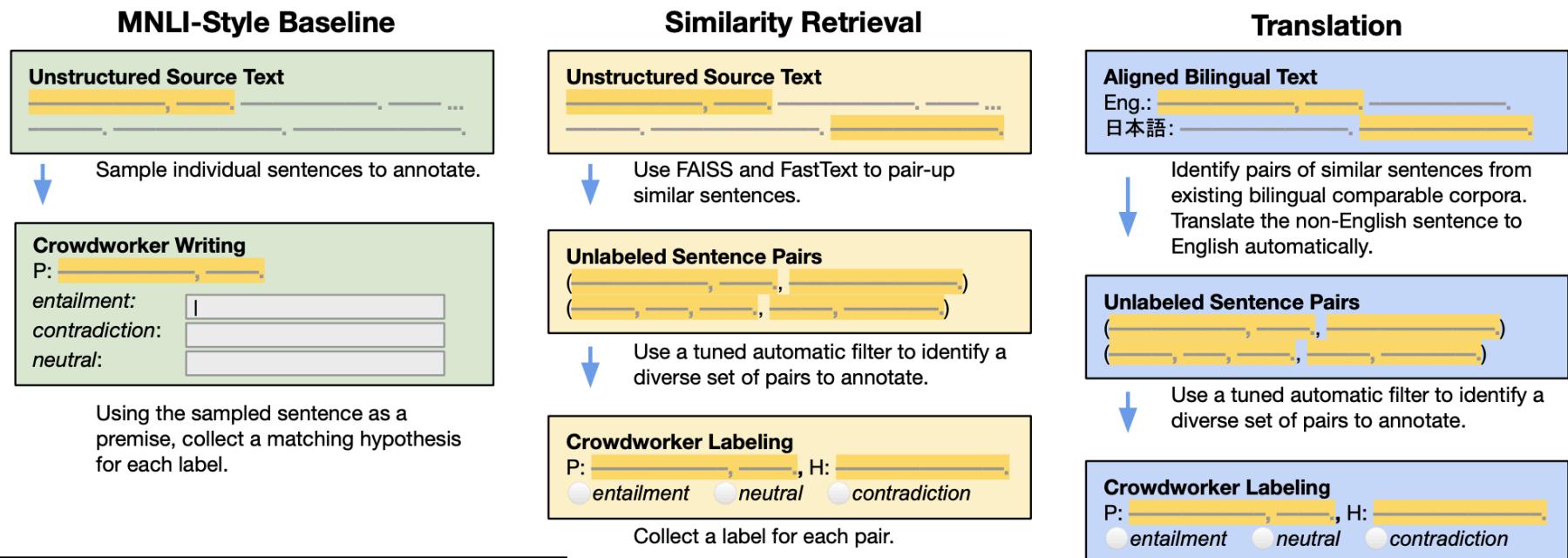
Give constraints and encouragement also lead to more challenging data.

Semi-Automatic Data Collection

Clara Vania

Replace Human Writing with Existing Natural Sentences

Baseline: MNLI-style human writing protocol



Pros:

- We can potentially collect more data
- Faster to annotate
- Might solve annotation artifacts issue

Cons:

- Not guaranteed to have a balanced distribution
- Similarity can be noisy

New proposed protocols

For another NLI dataset collected using semi-automatic protocol, see SciTail ([Khot et al., AAAI '18](#))

[Vania et al., AACL 2020](#)

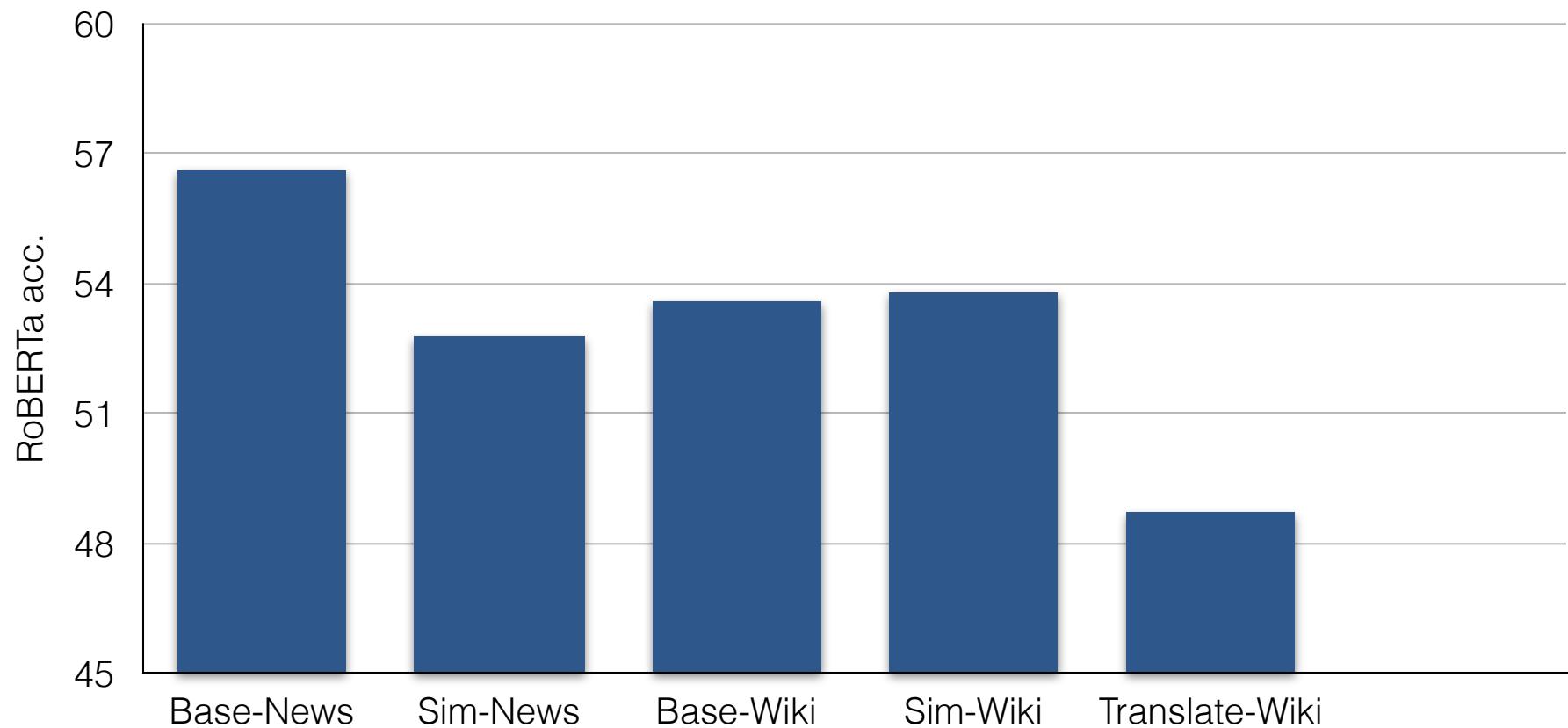
Experimental Setup

- Five new datasets:
 - Base-News, Base-Wiki (3k each)
 - Sim-News, Sim-Wiki, Translate-Wiki (6k each)
- Evaluation:
 - Generalization on NLI data
 - Transfer learning using intermediate-task training (Phang et al., 2018)



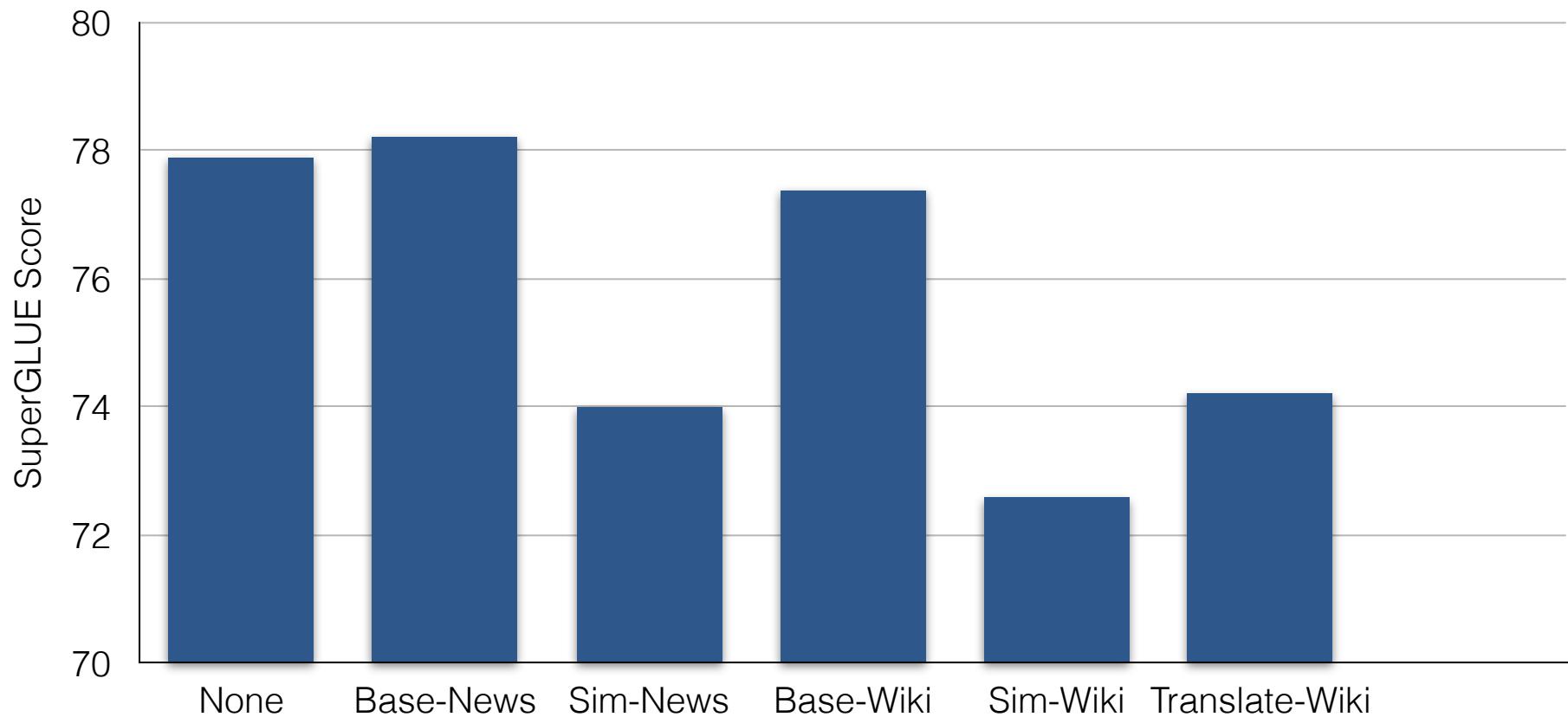
Evaluation on NLI

Human writing baseline is still the best one for NLI generalization!



Transfer Learning Results

The baseline protocol also yields model that transfer better.



Linguist-in-the-Loop Crowdsourcing

Nikita Nangia

Putting a Linguist in the Data-Collection Loop

Identifying and addressing gaps during data collection

Post-hoc analysis of NLU datasets has identified biases and unwanted artifacts in the data that models easily learn. E.g.: hypothesis-only bias

P: ???

H: Someone is not crossing the road.

Label: entailment, **contradiction**, neutral

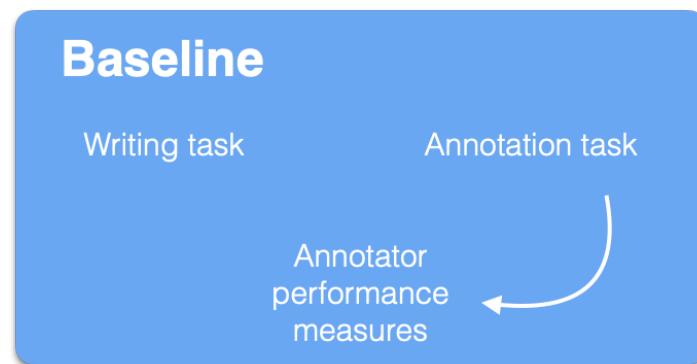
Can these biases and artifacts be identified **during** data collection and then addressed for the remaining data-collection process

Protocols

Test three protocols in parallel

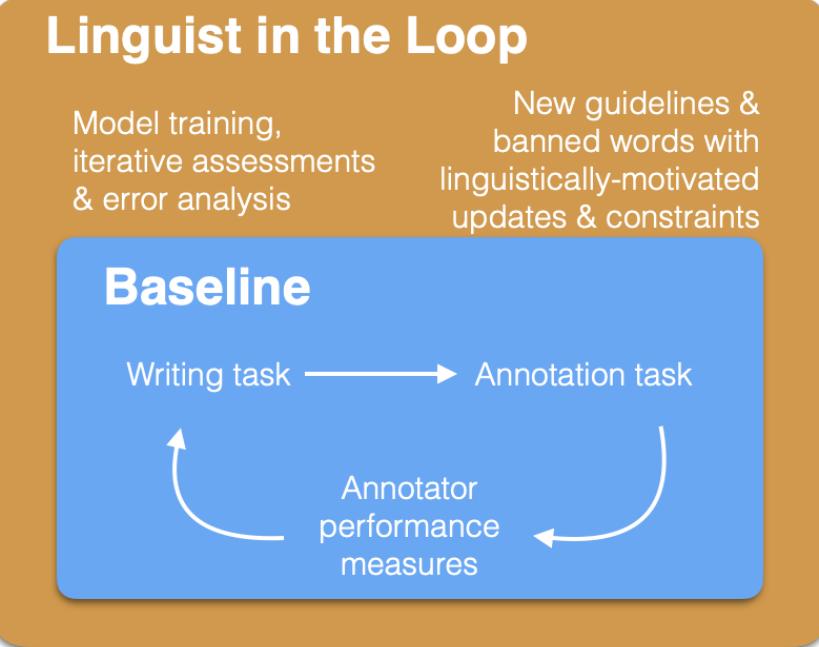
Protocols

Test three protocols in parallel



Protocols

Test three protocols in parallel



Protocols

Test three protocols in parallel

Linguist in the Loop with Chat

Expert and crowd workers stay active in a chatroom for task discussion and guidance

Linguist in the Loop

Model training,
iterative assessments
& error analysis

New guidelines &
banned words with
linguistically-motivated
updates & constraints

Baseline

Writing task → Annotation task

Annotator
performance
measures



Optional Constraints

Constraint	Premise	Hypothesis	Label	Attempt rate	
				LitL	Chat
Hypernym or hyponym	Does anyone know what happened to chaos ?	Whatever happened to the lack of order is certainly a mystery.	E	22.8	23.7
Banned word in diff. label	Inflation is supposed to be a deadly poison, not a useful medicine.	All people believe inflation is supposed to be a useful medicine	C	43.7	27.7
Temporal reasoning	John Kasich dropped his presidential bid.	They said that earlier , John Kasich had dropped his presidential bid.	E	34.1	10.0
Synonym or antonym	2) This particular instance of it stinks .	This instance is perceived to be a good thing .	C	39.5	24.5
All overlap	News argues that most of America's 93 million volunteers aren't doing much good.	News argues that volunteers aren't doing much good.	E	21.8	30.4
Register change	First, the horsemen brought out a teaser horse.	Teaser horses are commonly thought to be both entertaining and tragic.	N	25.3	15.0
No overlap	and she doesn't floss while driving.	The woman has an automated car.	N	29.2	22.3
Relative clause	Sun Ra's spaceships did not come, as it were, out of nowhere.	The spaceships that belong to Sun Ra came out of nowhere	C	35.0	24.3
Reverse argument order	After an inquiry regarding Bob Dole 's ...	It is illegal for Bob Dole to receive inquiries .	N	36.7	29.4
Grammar change	The Bush campaign has a sweet monopoly on that.	The Obama campaign had a sweet monopoly on that.	C	22.6	13.4
Sub-part	He was crying like his mother had just walloped him .	He cried a lot, as though he were walloped on his behind .	E	23.2	19.1
Background knowledge	In both Britain and America , the term covers nearly everybody.	The term generally applied to countries in two opposite sides of the world .	E	32.9	15.9

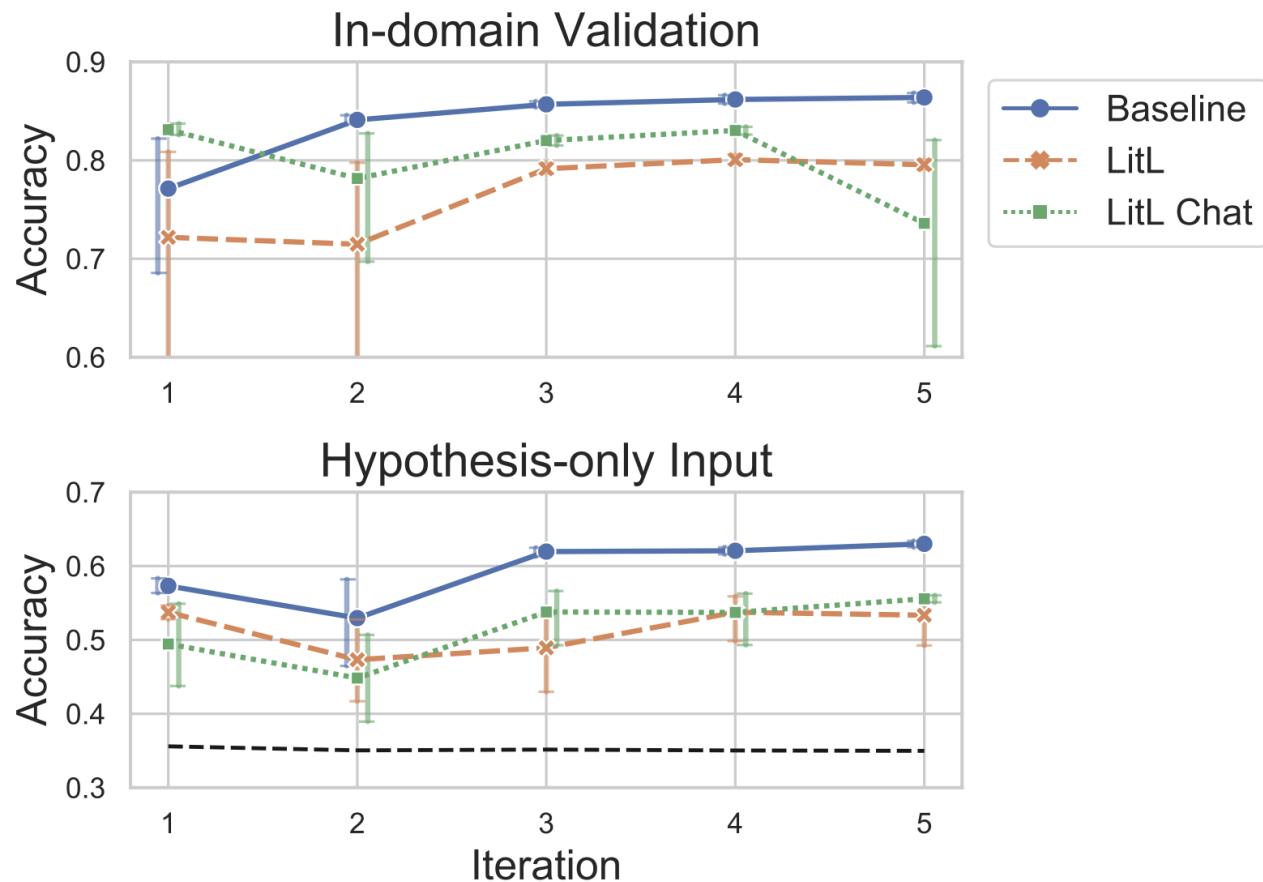
Table 1: Sentence pairs displaying each challenge option. Where applicable, relevant contrasts are bolded. Examples are randomly drawn from data that passed validation on the constraint with the restriction that both sentences be fewer than 80 characters (~ 32% of the data). The last column shows the percentage of the challenges attempted.

Results

The interventions help in some ways

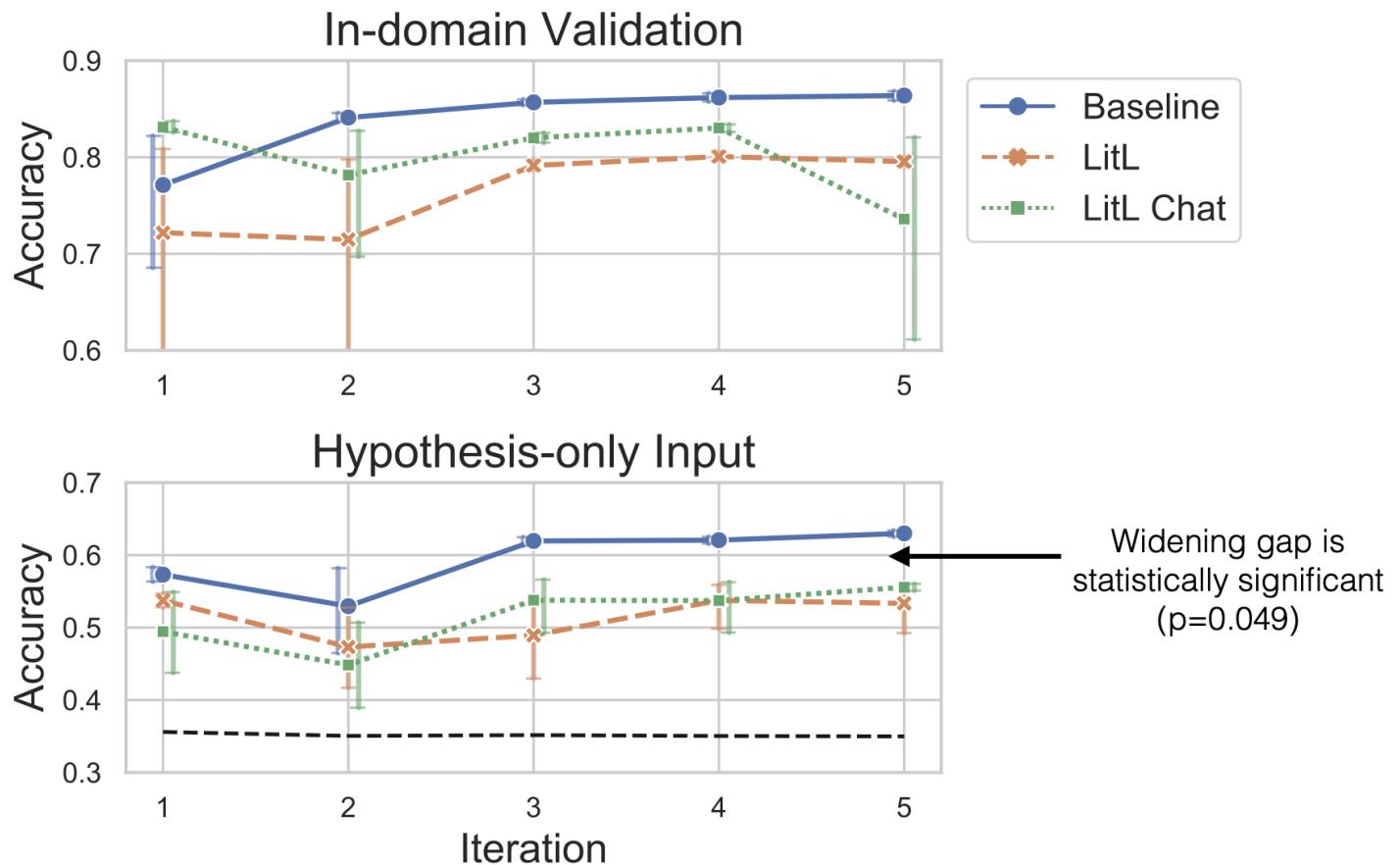
Results

The interventions help in some ways



Results

The interventions help in some ways

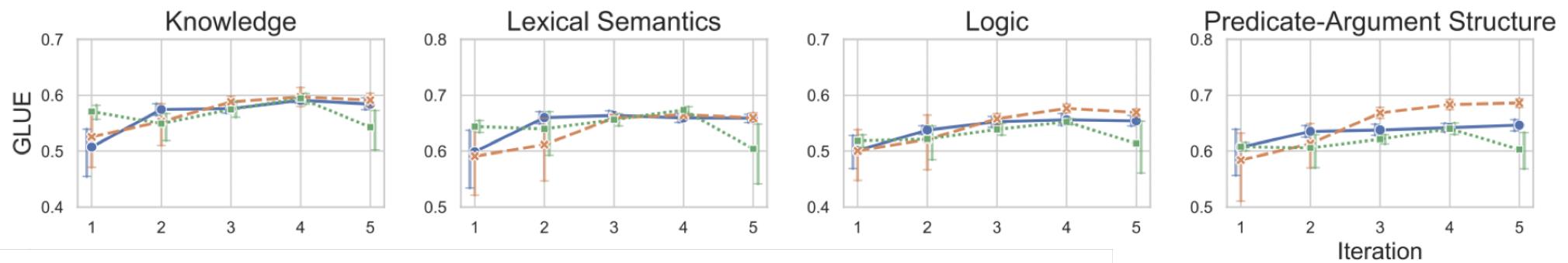


Results

...but not in others

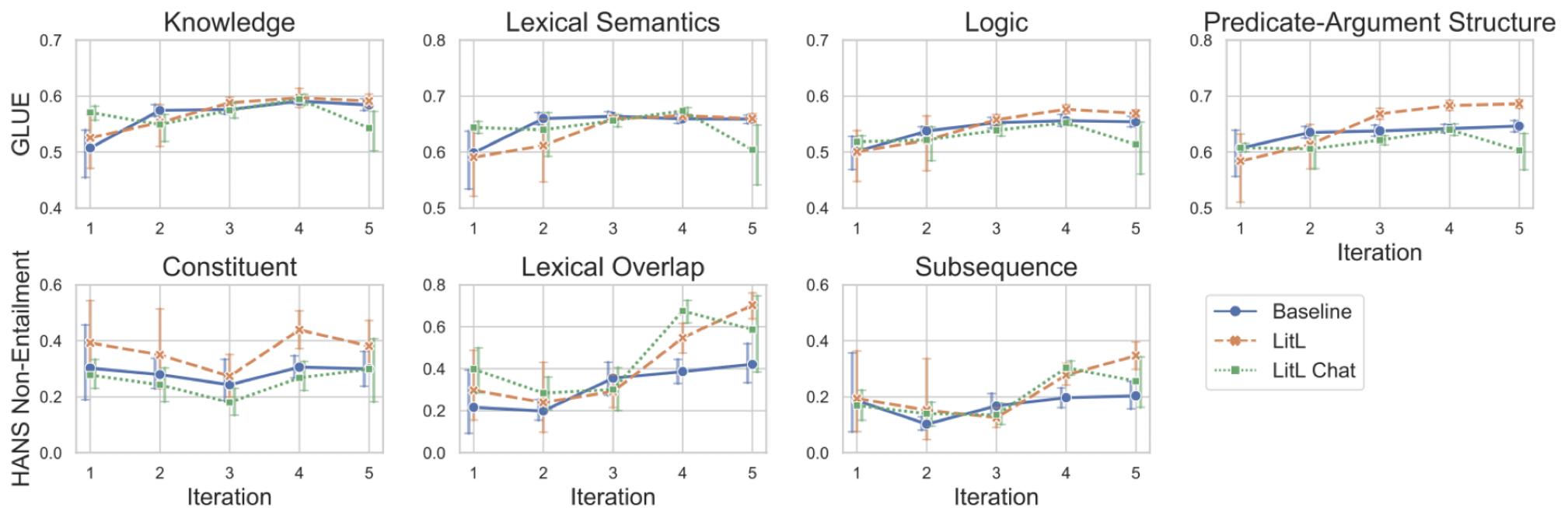
Results

...but not in others



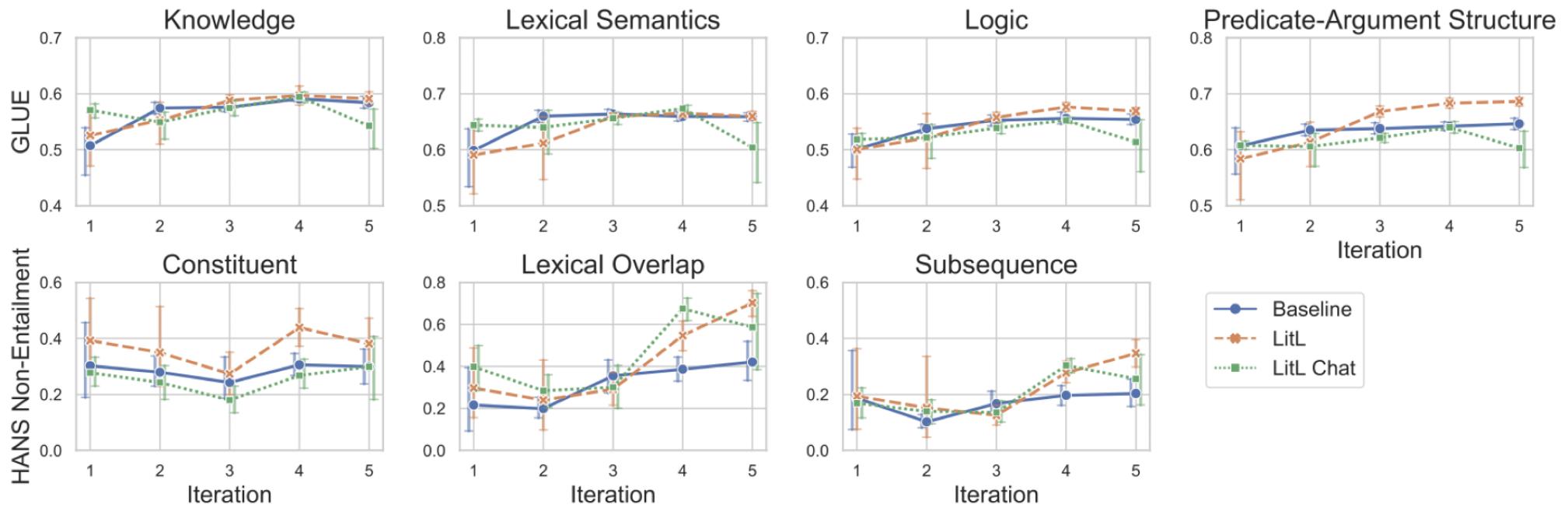
Results

...but not in others



Results

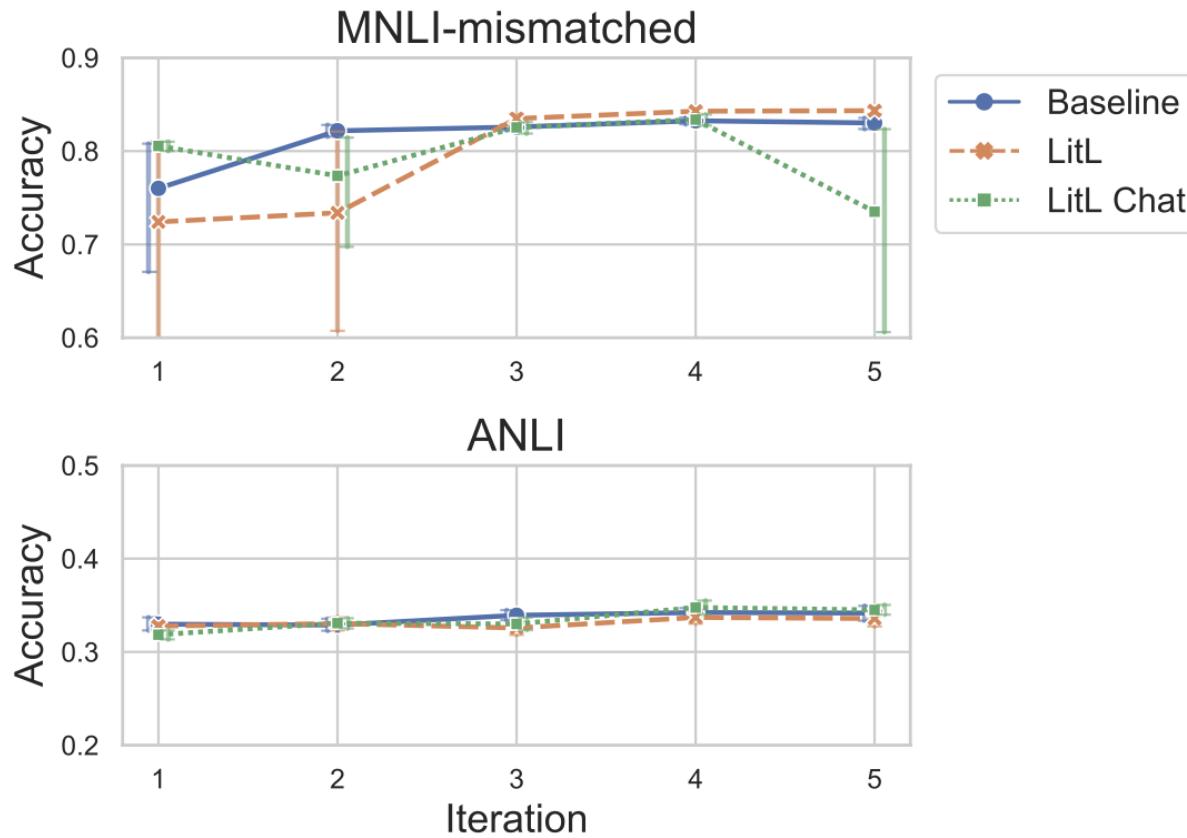
...but not in others



LitL and LitL-Chat have statistically significant higher accuracy

Results

...but not in others



Takeaways!

- Adding an expert/linguist in the loop is beneficial for collecting more challenging data with fewer dataset artifacts!
- However, this doesn't help with out-of-domain accuracy
- LitL-chat offered no advantages over LitL, real time interactions with crowd workers was not helpful in this study

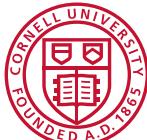
EMNLP 2021 Tutorial

Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection

Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar,
Sam Bowman, and Yoav Artzi

Case Study II: NLVR

Presented by Alane Suhr



ML²



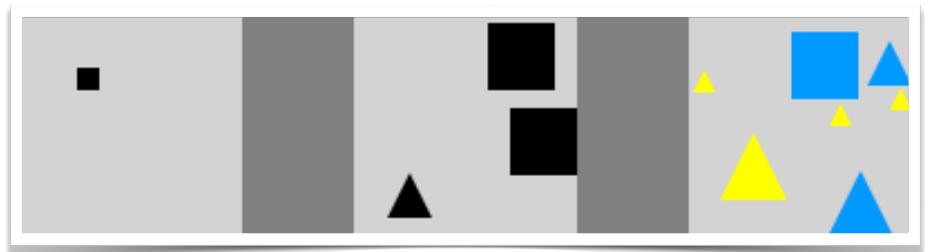
Penn
UNIVERSITY OF PENNSYLVANIA

Natural Language for Visual Reasoning

- **Our goal:** large corpus of natural language paired with images, focusing on a diverse set of language phenomena
- **Task:** given an image and sentence, determine whether the sentence is true or false about the image

Natural Language for Visual Reasoning

*there are exactly three squares
not touching any edge
(NLVR, Suhr et al. 2017)*



TRUE

All dogs are corgis with upright ears, and one image contains at least twice as many real corgis as the other image.

(NLVR2, Suhr et al. 2019)



TRUE

Outline

- Design goals of NLVR and NLVR2
- Data collection process
- Data quality measures
- Managing crowdsourcing
- Resulting corpora

Corpus Goals

- Images that include diversity of spatial relations, attributes, and grouping of objects
- Language that requires reasoning about spatial relations, sets, counts, negation, etc.
- In parallel, done with synthetic images and language (e.g., CLEVR)
- How do we elicit natural language without allowing for reasoning shortcuts?

Corpus Design: Task

- Should be easy to evaluate
 - Use a binary classification task
- Should measure model robustness in language understanding
 - Measure consistency of model prediction for each sentence across multiple paired images

Corpus Design: Pitfalls

- Avoiding spurious correlations between inputs (image + text) and labels
 - Pair each sentence with both positive and negative labels
 - Small label set and pairing sentences with multiple labels means no possible bias between sentence and label only
- Finding a balance between simplicity and complexity in language
 - Require sentences to apply to more than one, but not all, images
- **Added benefit:** lower cost per sentence

Task Design

- Sourcing image contexts
 - NLVR: generating synthetic images
 - NLVR2: image web search
- Sentence-writing task
- Validation

Image Sourcing

- **Main principle:** want complex and interesting images
 - Different objects with diverse properties
 - Interesting spatial relations
- Solution:
 - First, synthetically generate images (NLVR)
 - To extend to real images, use web search and queries that result in complex images (NLVR2)

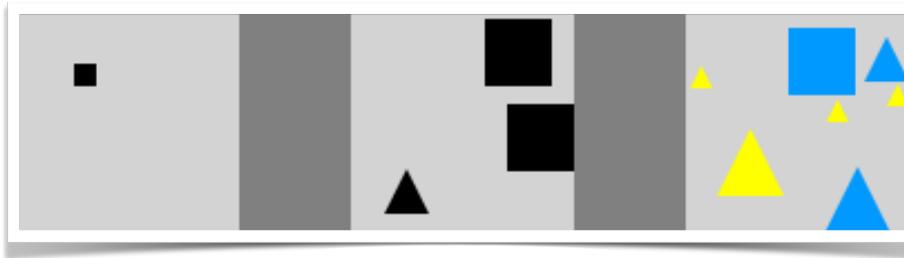
Sentence-Writing

- **Main principle:** want sentences paired with
 - Multiple image contexts
 - Multiple labels
- Solution: contrastive sentence-writing

Validation

- Main principle: make sure the labels implicitly derived from sentence-writing are correct
- Solution:
 - Task asking workers to label image+text pairs as true or false
 - Show image+text pairs independent of their original image contexts
 - Filter out low-agreement pairs (but see Leonardelli et al. 2021 for discussion on this)

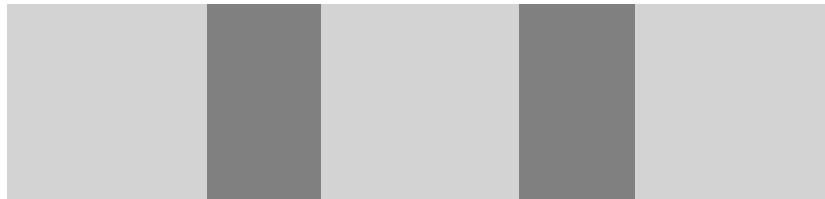
NLVR



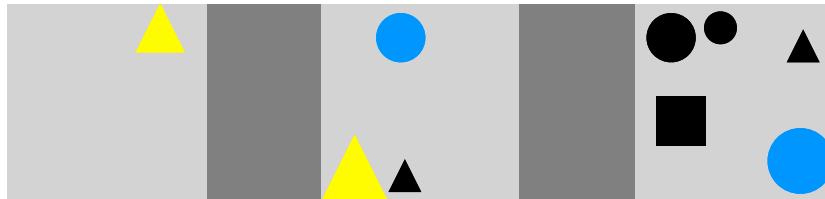
TRUE

*there are exactly three squares
not touching any edge*

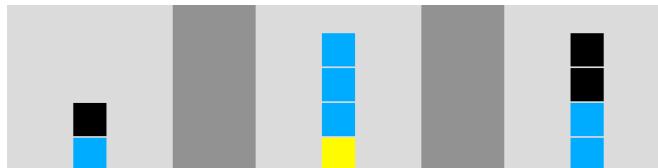
NLVR Image Generation



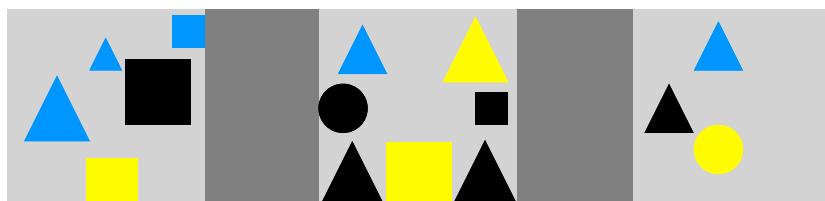
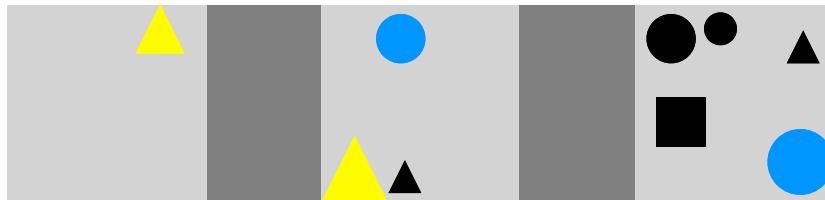
NLVR Image Generation



- Randomly generate a single image
- Colorblind-friendly
- Small number of properties
- 3 boxes to group objects
- Scatter and tower configurations

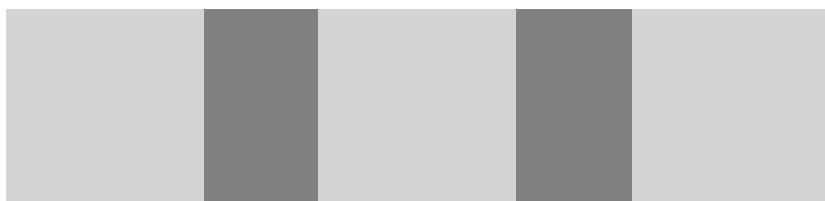
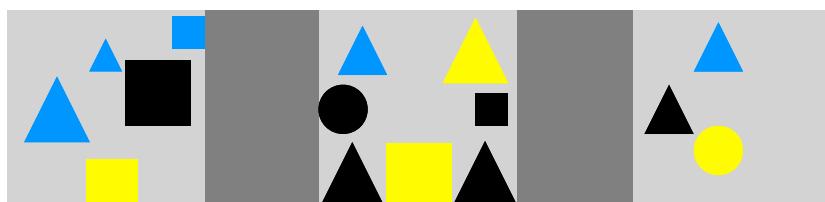
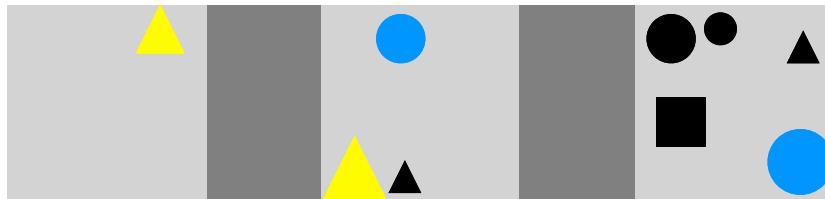


NLVR Image Generation



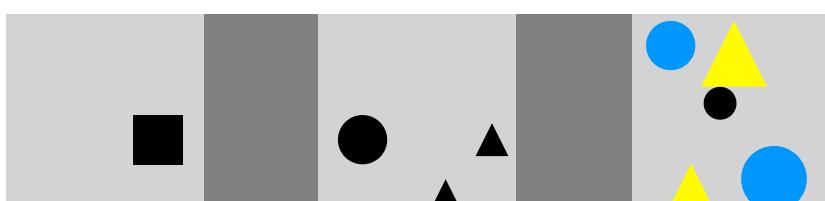
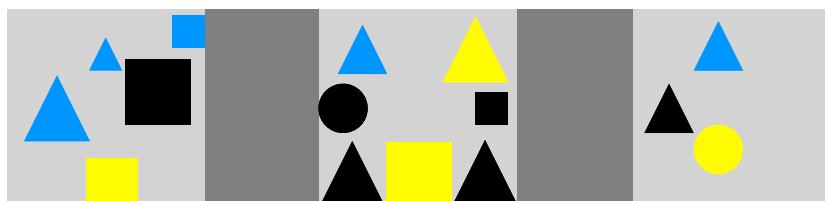
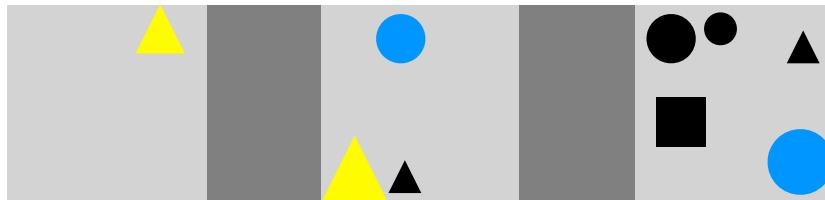
- Randomly generate a single image
- Randomly generate another image

NLVR Image Generation



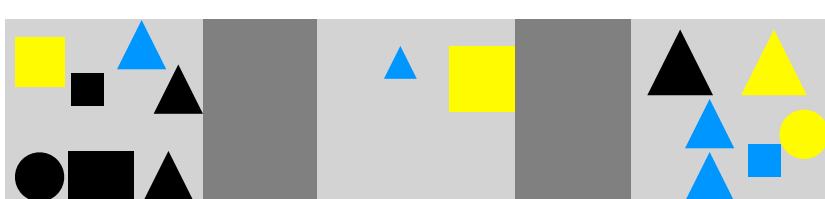
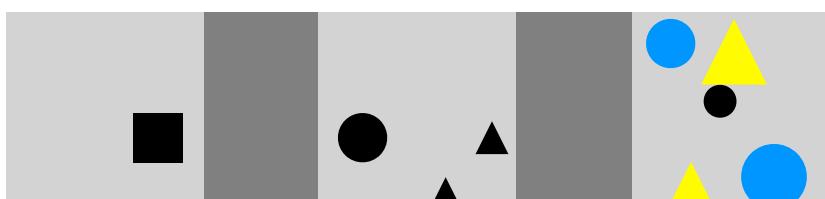
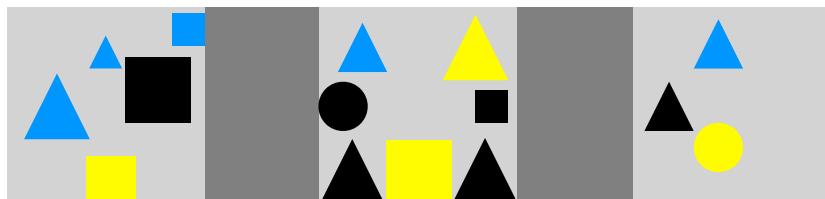
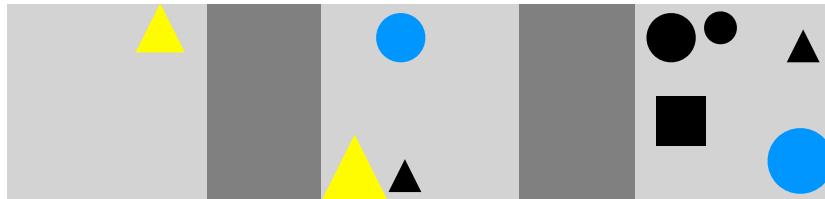
- Randomly generate a single image
- Randomly generate another image

NLVR Image Generation



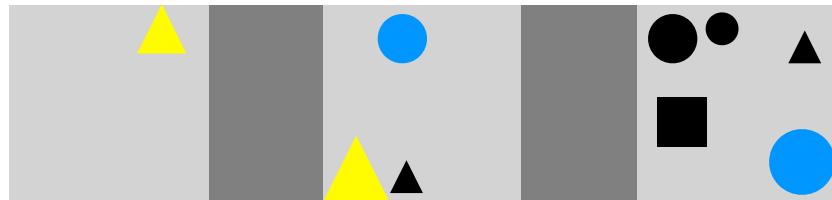
- Randomly generate a single image
- Randomly generate another image
- Generate a third image, using objects from top image

NLVR Image Generation

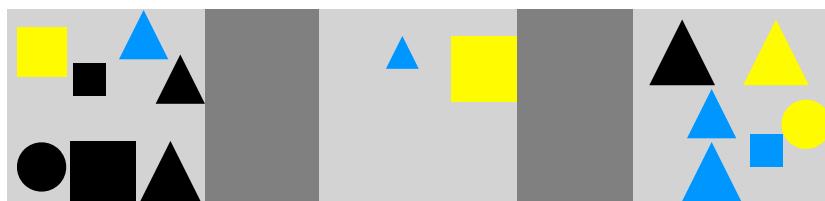
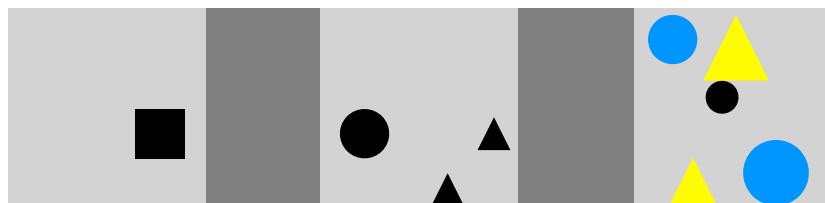
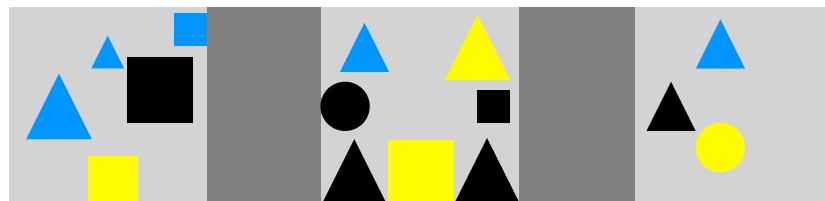


- Randomly generate a single image
- Randomly generate another image
- Generate a third image, using objects from top image
- Generate a fourth image similarly

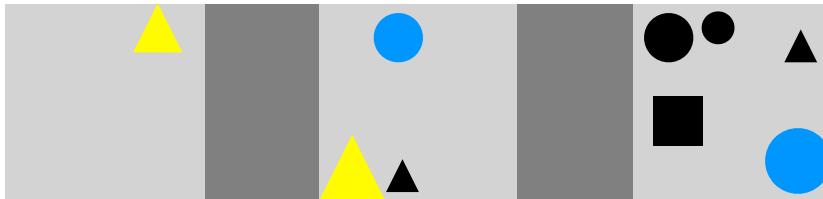
NLVR Sentence Writing



There is a box with 3 items of all 3 different colors.

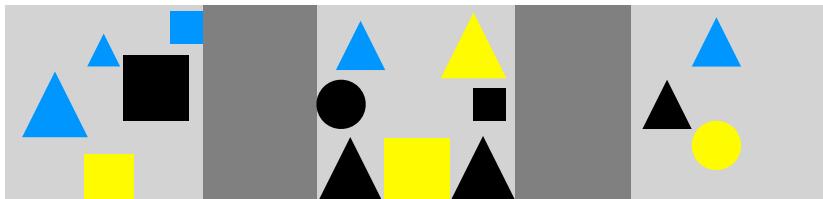


NLVR Sentence Writing



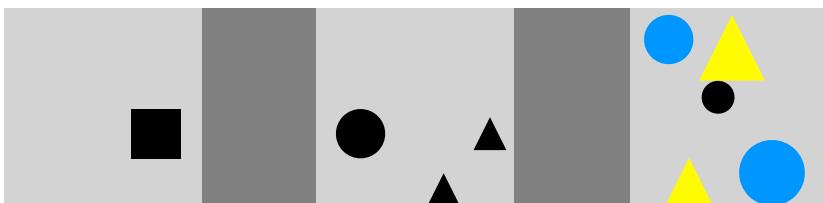
There is a box with 3 items of all 3 different colors.

TRUE



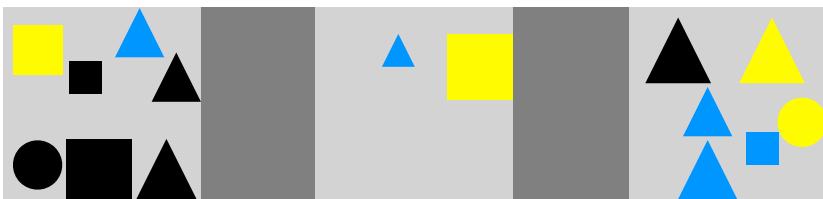
There is a box with 3 items of all 3 different colors.

TRUE



There is a box with 3 items of all 3 different colors.

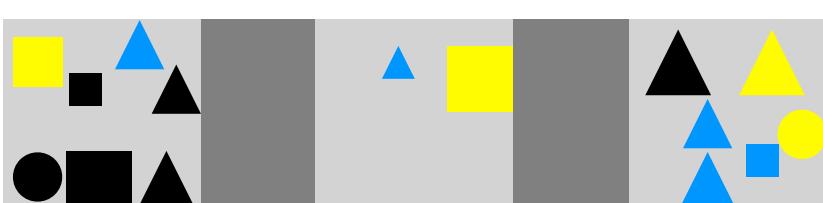
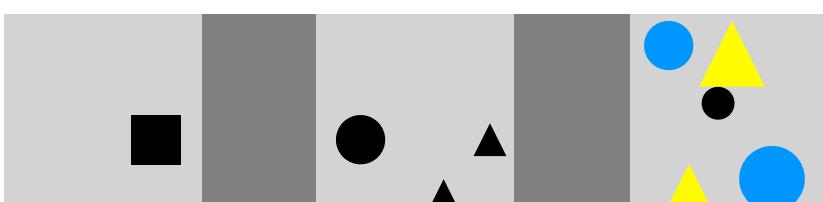
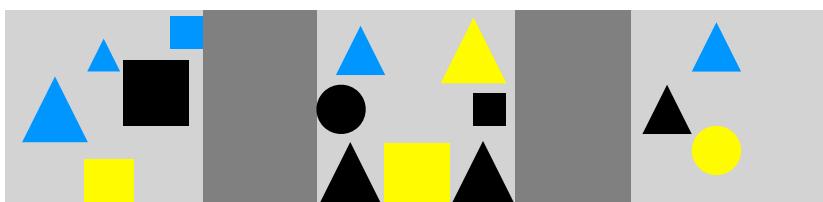
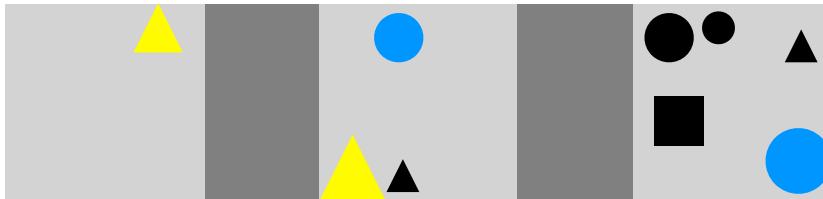
FALSE



There is a box with 3 items of all 3 different colors.

FALSE

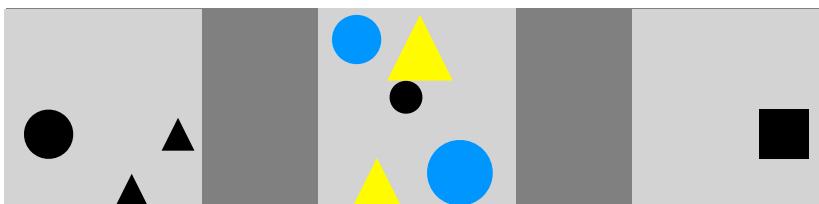
NLVR Sentence Writing



- Sentence must be general enough to be true for two images
- But not so general that it describes all images
- Shuffling objects prevents trivial descriptions, e.g., “there is a blue square”

NLVR Validation

There is a box with 3 items of all 3 different colors.



TRUE



FALSE

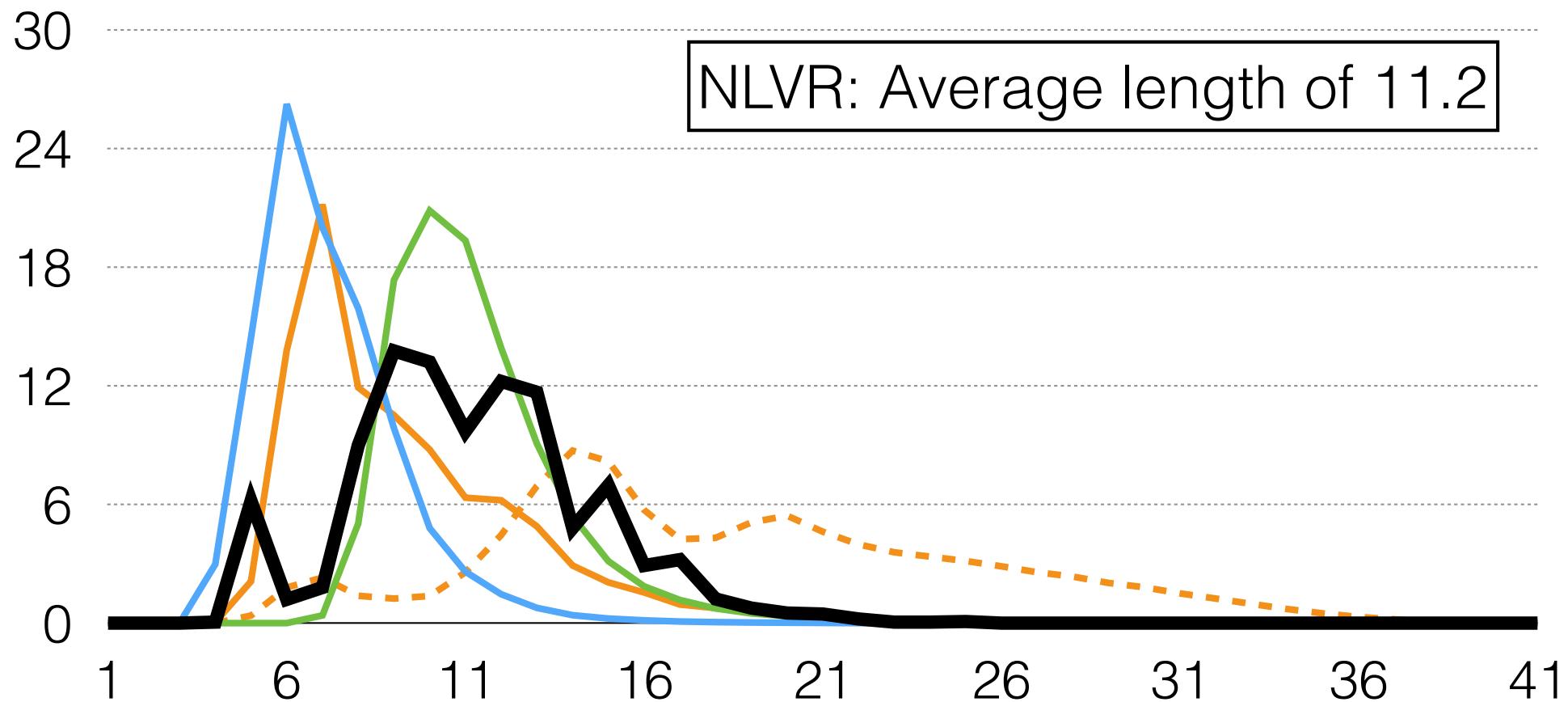
- Show another worker a sentence and image independently
- Also shuffle the boxes in the image

NLVR Stats

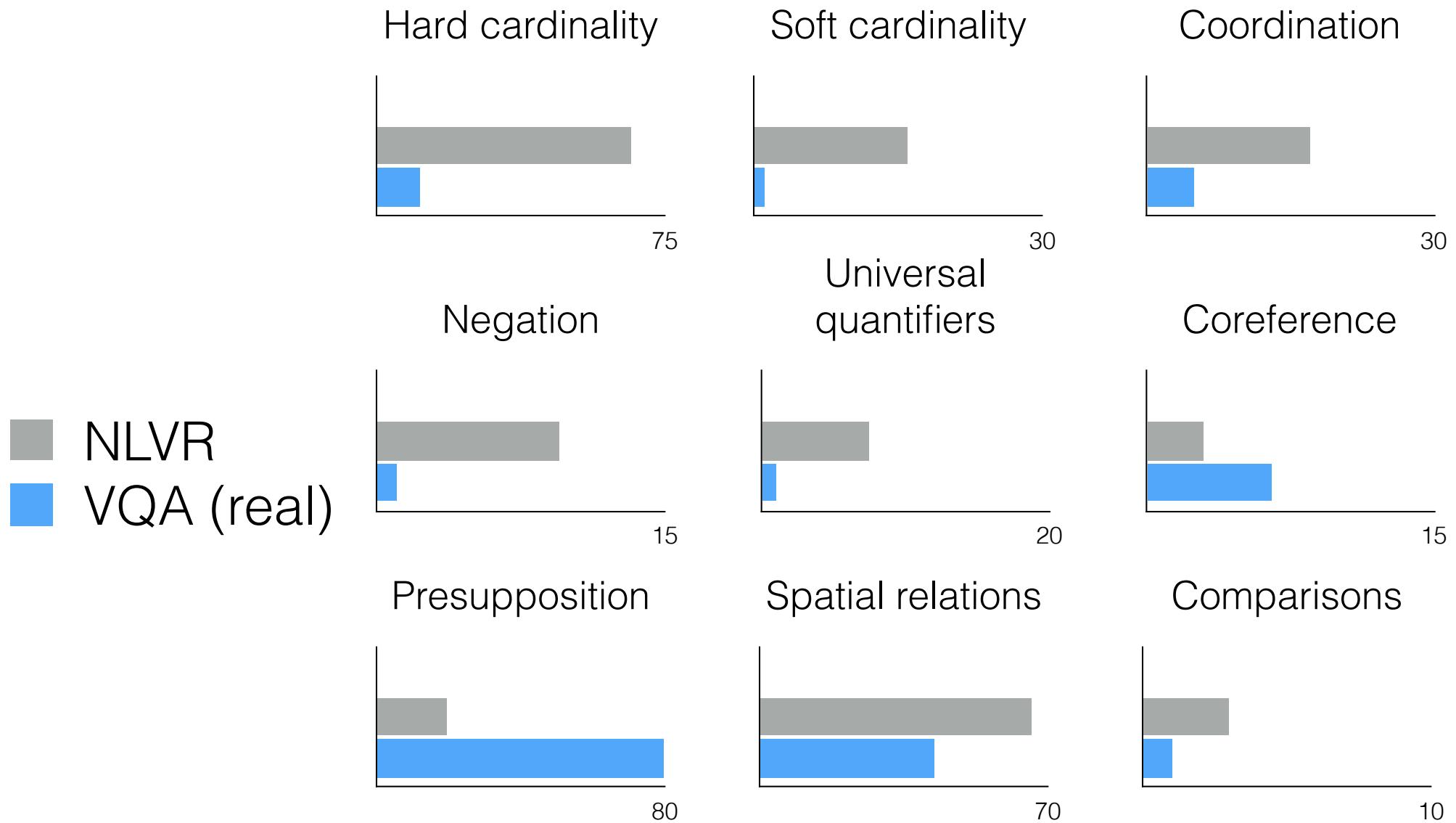
	# Examples	# Unique Sentences	Agreement (a)	Vocab Size
NLVR (Suhr et al 2017)	92,244	3,692	0.831	262

Sentence Lengths

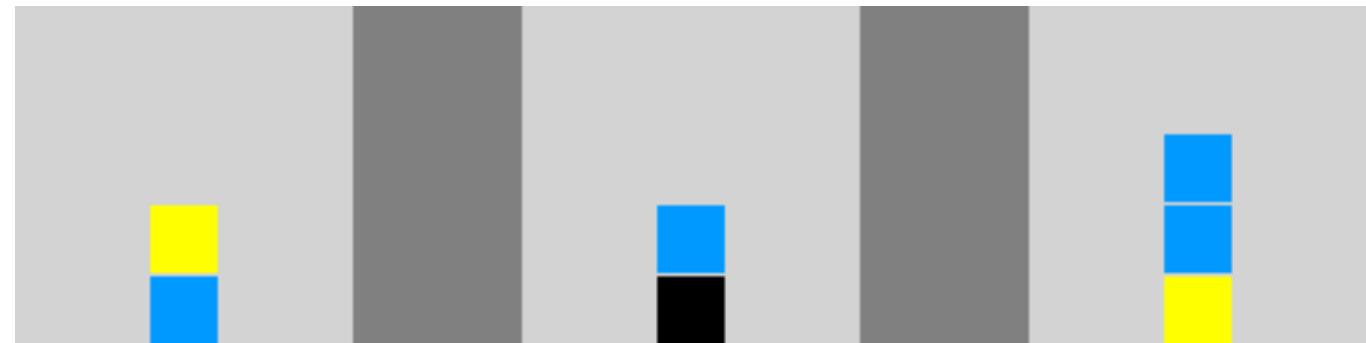
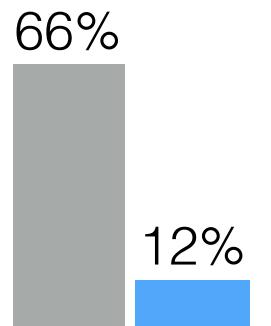
- NLVR
- MSCOCO
- - - CLEVR
- VQA real images
- CLEVR-Humans



Linguistic Analysis



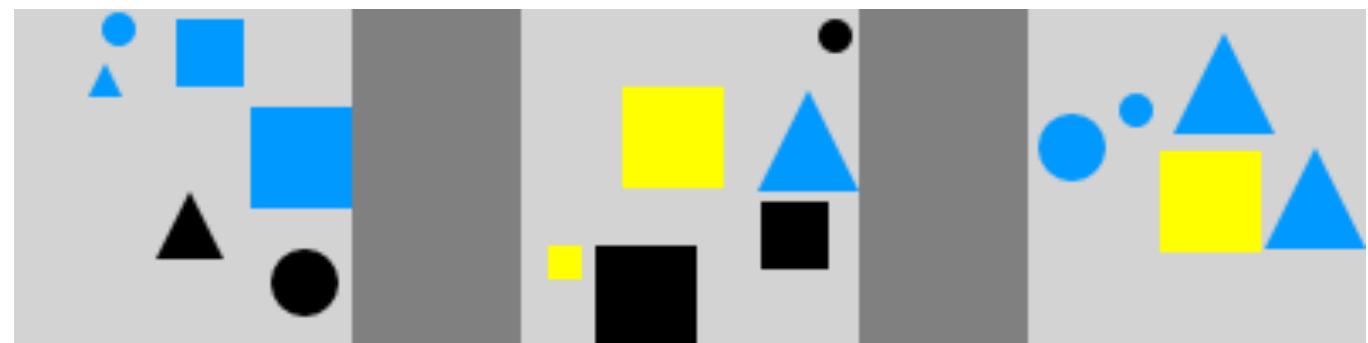
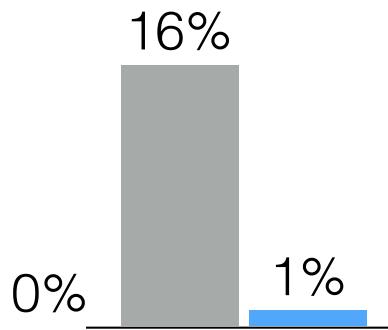
Hard Cardinality



*There is a tower with exactly three blocks,
and it has a yellow block and two blue blocks.*

TRUE

Soft Cardinality

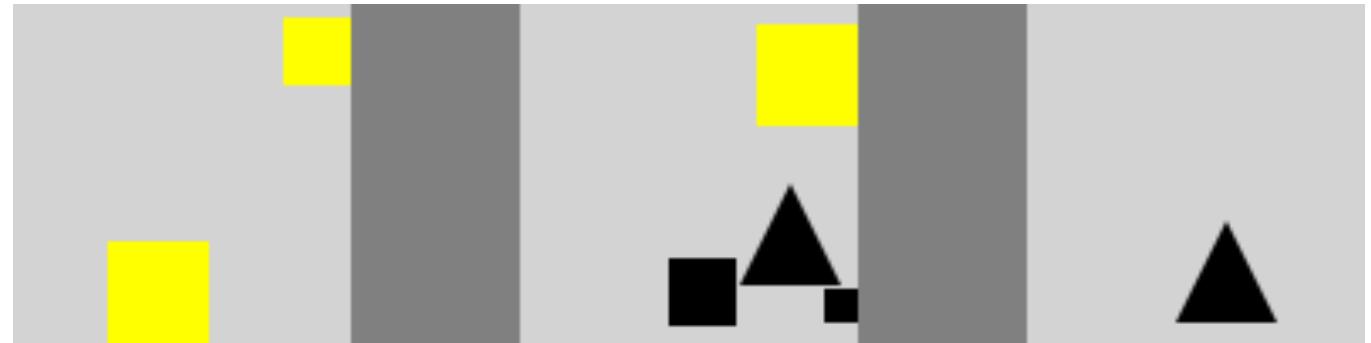
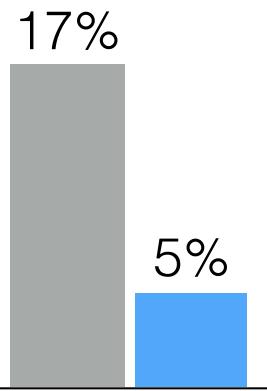


*there are at least two yellow squares not
touching any edge*

TRUE

- NLVR
- VQA (real)

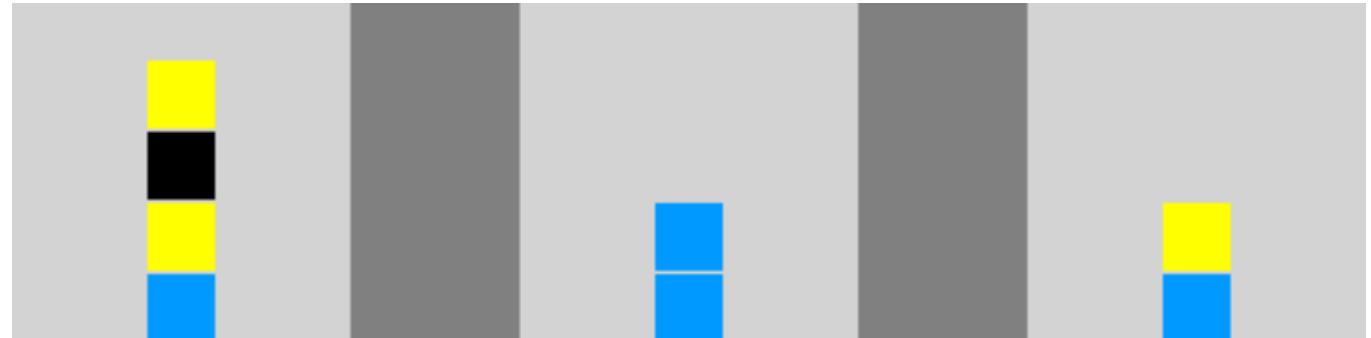
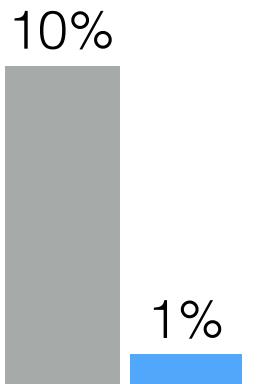
Coordination



*There is a box with a yellow item
and three black items.*

TRUE

Negation



*There is a box with a black item between 2 items
of the same color and no item on top of that.*

TRUE

- NLVR
- VQA (real)

NLVR2

All dogs are corgis with upright ears, and one image contains at least twice as many real corgis as the other image.



TRUE

NLVR2 Image Sourcing

- Can't generate images to control content, but still want complex (and similar) images
- Solution: image search engine + similarity tools
 - Design queries that elicit complex images
 - Use Similar Image tools to construct sets of image contexts
 - Filtering step for ensuring quality

Image Collection

1. Pick 124 synsets from ImageNet

Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump

- Allows use of ImageNet models and tools
- Allows for weak annotation of image content



Image Collection

1. Pick 124 synsets from ImageNet

Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump

2. Generate and execute search queries

Combined synset names with numerical phrases, hypernyms, and similar words



Image Collection

3. Remove low-quality images

Don't contain synset, drawings, inappropriate content



4. Construct sets of eight images

Each set must contain at least three *interesting* images
(e.g., multiple objects)

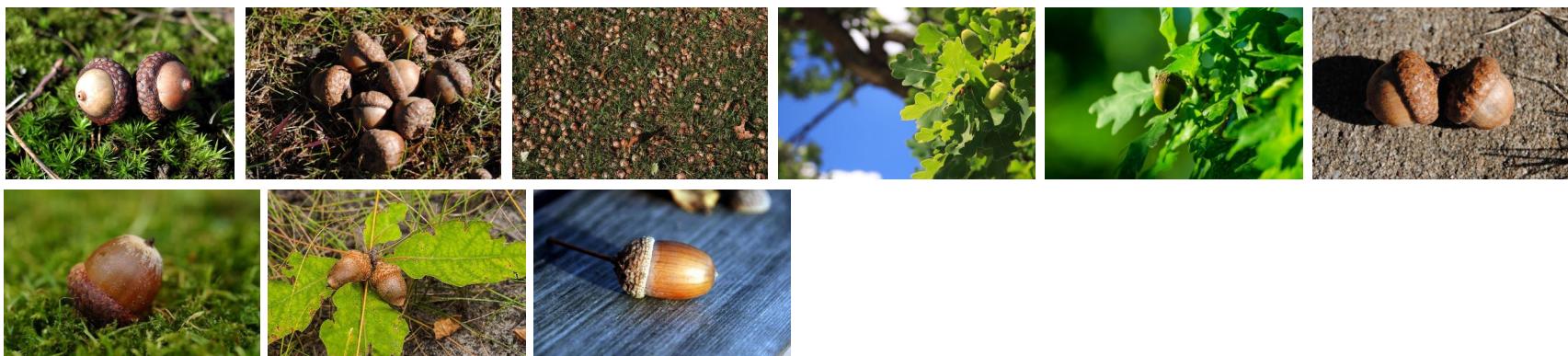


Image Collection

3. Remove low-quality images

Don't contain synset, drawings, inappropriate content



4. Construct sets of eight images

Each set must contain at least three *interesting* images
(e.g., multiple objects)

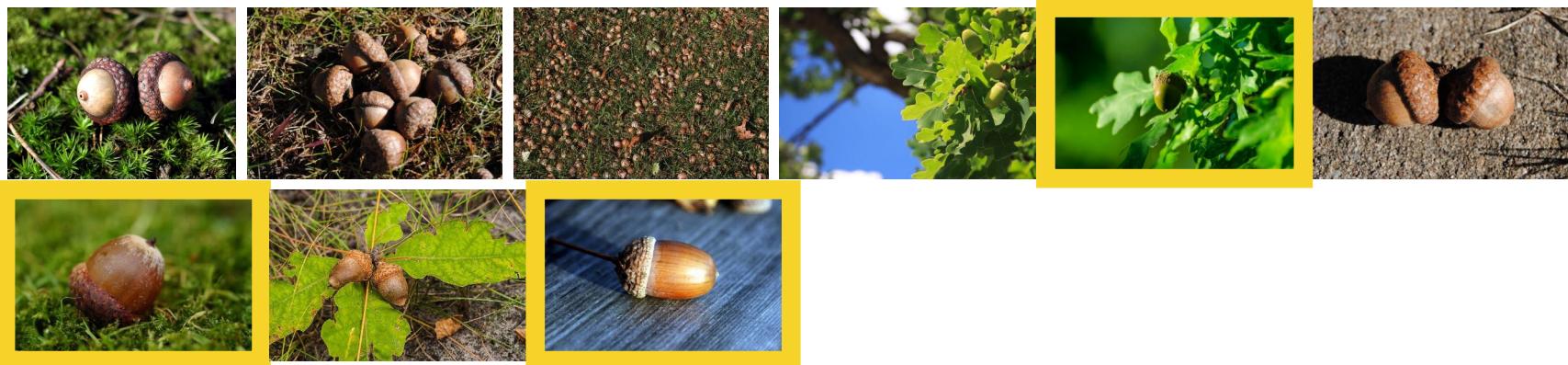


Image Collection

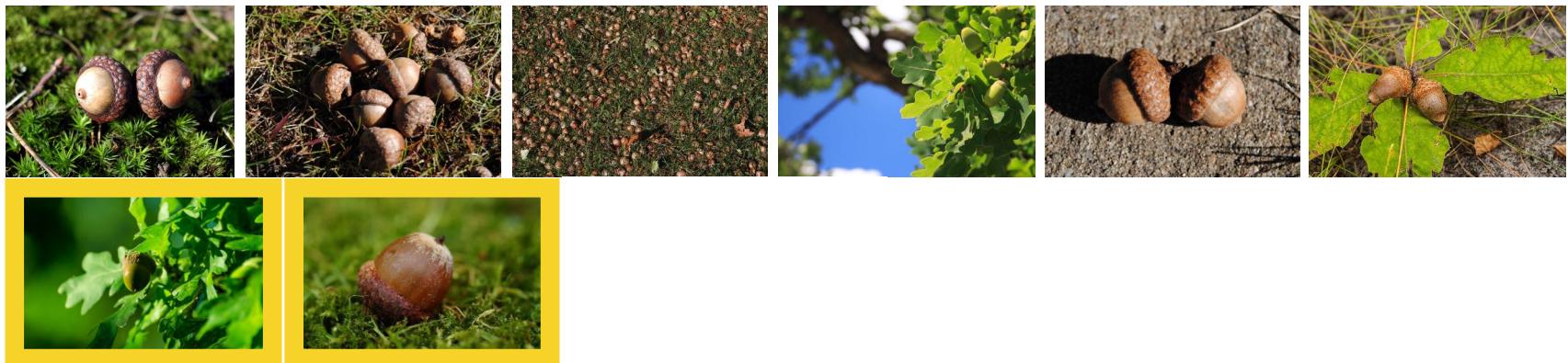
3. Remove low-quality images

Don't contain synset, drawings, inappropriate content



4. Construct sets of eight images

Each set must contain at least three *interesting* images
(e.g., multiple objects)



NLVR2 Sentence Writing

5. Display a set of randomly paired images

6. Ask workers to select two pairs

7. Workers write a sentence **true about the selected pairs, but **false** about the others**



One image shows exactly two brown acorns in back-to-back caps on green foliage.

Validation Tasks



One image shows exactly two brown acorns in back-to-back caps on green foliage.



TRUE



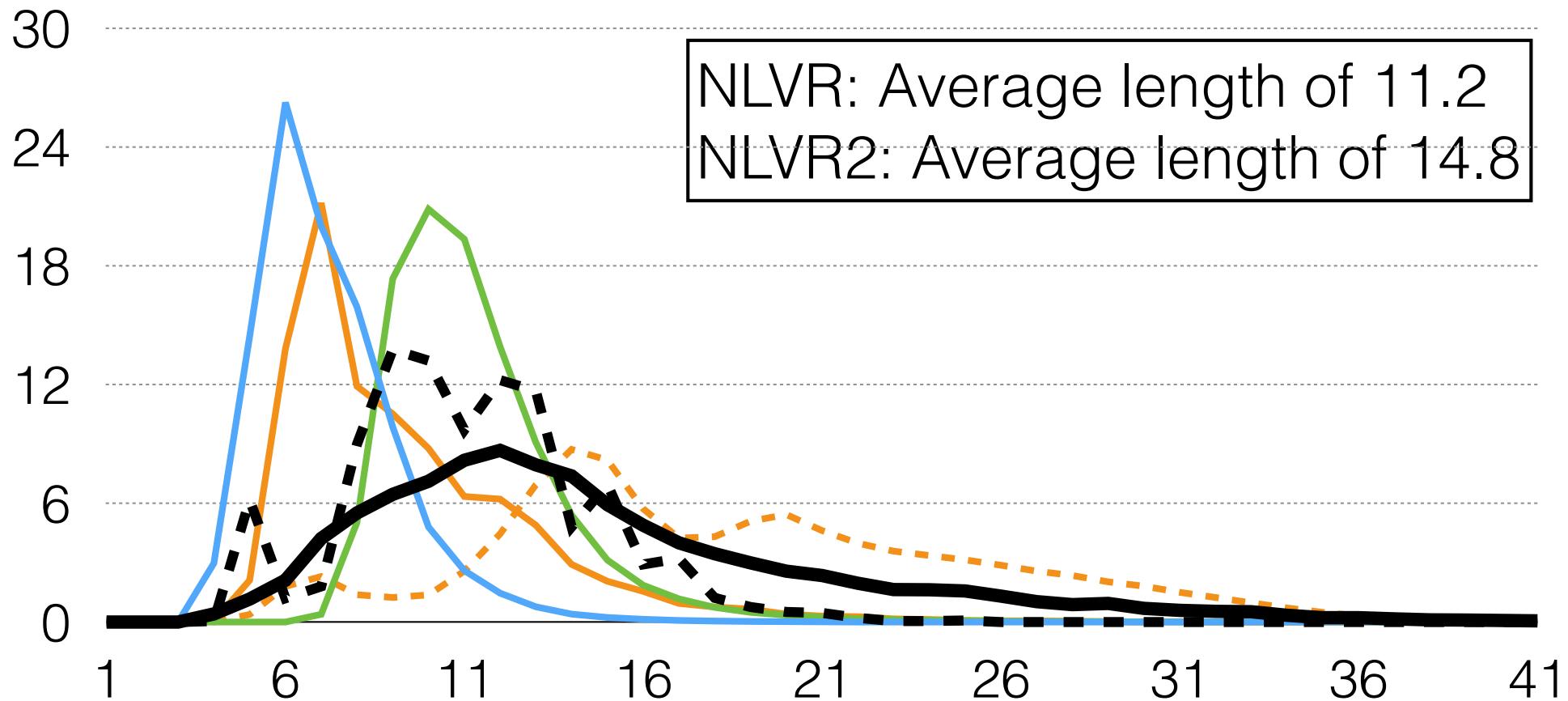
FALSE

NLVR Stats

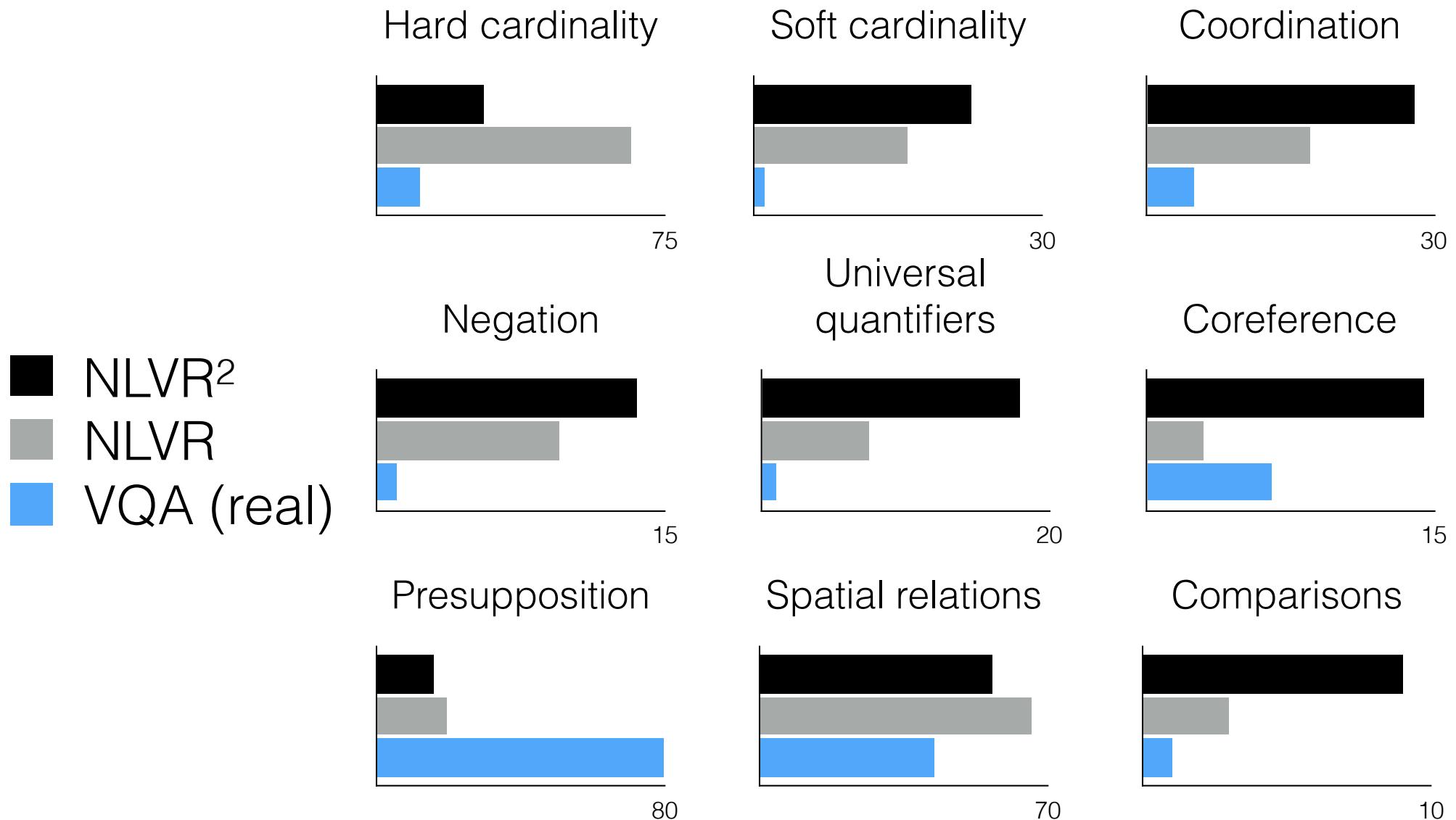
	# Examples	# Unique Sentences	Agreement (a)	Vocab Size
NLVR (Suhr et al 2017)	92,244	3,692	0.831	262
NLVR2 (Suhr et al 2019)	107,296	29,680	0.912	7,500

Sentence Lengths

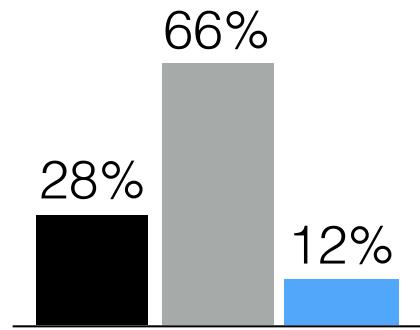
- NLVR2
- VQA real images
- CLEVR-Humans
- NLVR
- MSCOCO
- CLEVR



Linguistic Analysis

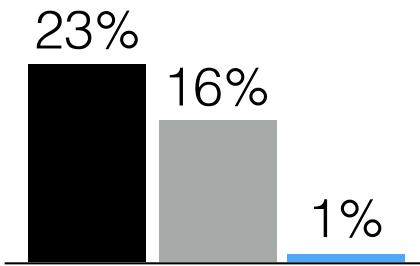


Hard Cardinality



There are two, and only two, people.

Soft Cardinality



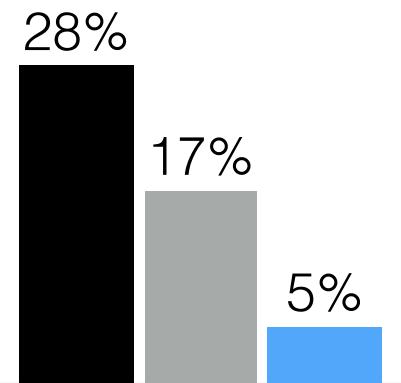
There are no more than eight bottles in total.

■ NLVR²

■ NLVR

■ VQA (real)

Coordination



■ **NLVR²**

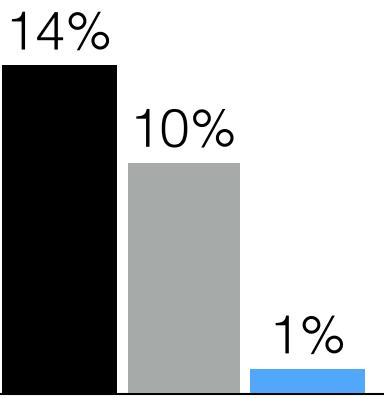
■ NLVR

■ VQA (real)



*Each image contains just one bird,
and the wires of a cage are behind
the bird in one image.*

Negation

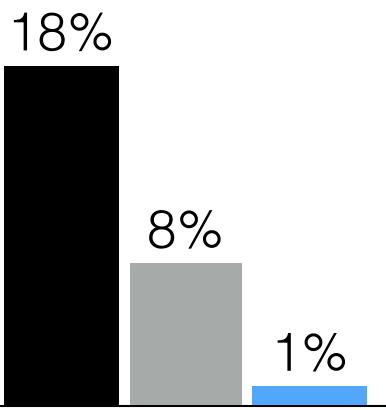


- **NLVR²**
- NLVR
- VQA (real)



A mitten is being worn in one image and the mittens are not being worn in the other image.

Universal Quantifiers



■ **NLVR²**

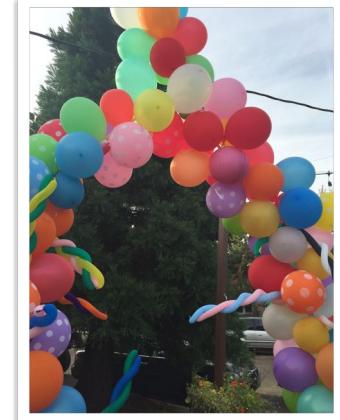
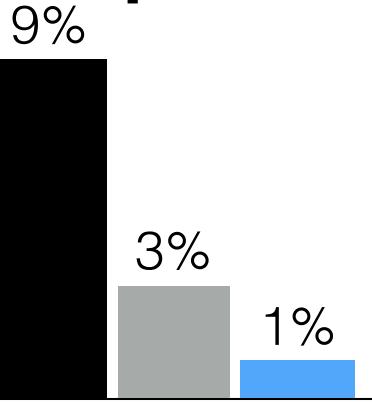
■ NLVR

■ VQA (real)



*Both images shows a silver pail
being used as a flower vase.*

Comparisons



*the left image has 4 balloons
of all different colors*

Spurious Correlations in NLVR2

- While we avoided spurious correlations between text and label, we found that there may be spurious correlations between images and labels
- Problem: allowing workers to select pairs to be true and false during sentence-writing
- Quantifying this effect
 - Looked at image pairs that were annotated twice with different sentences, where we expect labels for those pairs to be uniformly distributed
 - Found that there were more pairs that had the same label for both sentences than expected!
 - Assigning most common label for image pair leads to high accuracy, without even looking at the sentence
 - However, evaluating on a balanced subset of the data shows that existing models mostly did not take advantage of this bias
- URL of analysis: <https://lil.nlp.cornell.edu/nlvr/NLVR2BiasAnalysis.html>

Managing Crowdworkers

- Platforms
- Qualifications
- Pay and incentives

Crowdsourcing Platforms

- NLVR
 - Upwork
 - 10 total workers
 - Cost: \$5,526
- NLVR2
 - MTurk
 - 167 total workers — harder to scale
 - Cost: \$19,133
- Regular communication with workers via email and forums
- English

Base Qualifications

- English proficiency

NLVR2 Qualifications

- For image curation and sentence-writing
 - Read a short tutorial about the guidelines and task
 - Short quiz on guidelines for 19 sentences with 2 sets of pre-selected images
 - One sentence-writing task with a pre-selected image set
- For validation
 - Eight validation tasks on pre-selected images and sentences

NLVR2 Pay and Incentives

- Workers receive a bonus for each task they complete (image pruning, sentence-writing, or validation)
- Refined sentence-writing expertise through novice/expert pools:
 - Novice pool has fewer available HITs and a lower bonus
 - Regularly sample twenty sentences written by each worker
 - Evaluate each for following guidelines
 - If at least 75% follow guidelines, they receive a bonus for each sentence written, and are part of an **expert** pool
 - If between 50-75% follow guidelines, they receive a slightly lower bonus and are moved to the **novice** pool

Procedure Summary

1. Planning
2. Collecting source data
3. Designing beta qualifications and tasks
4. Pilots and refinement cycle
5. Deploy main data collection phase cycle
 1. Continually check quality and communicate with workers
 2. Run validation
6. Filter and prune data and split into training/testing sets

Takeaways

- Think about clever ways to measure model consistency and avoid spurious correlations
- Validation is very important
- Novice / expert pools
- Language analysis

EMNLP 2021 Tutorial

Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection

Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar,
Sam Bowman, and Yoav Artzi

Case Study III: CerealBar

Presented by Yoav Artzi and Alane Suhr



ML²



Outline

- Game design, incentive structure, and tasks
- The crowdsourcing data collection process
- Resulting corpus

CerealBar

A situated collaborative game with sequential natural language instruction

- **Interaction:** participants respond to each others' language and behavior across multiple turns
- **Collaboration:** participants are incentivized to work together and must coordinate using language

CerealBar

A situated collaborative game with
sequential natural language instruction



Game Environment

- Passable terrain
- Impassable terrain
- Landmarks
- Cards



Collaboration

Leader

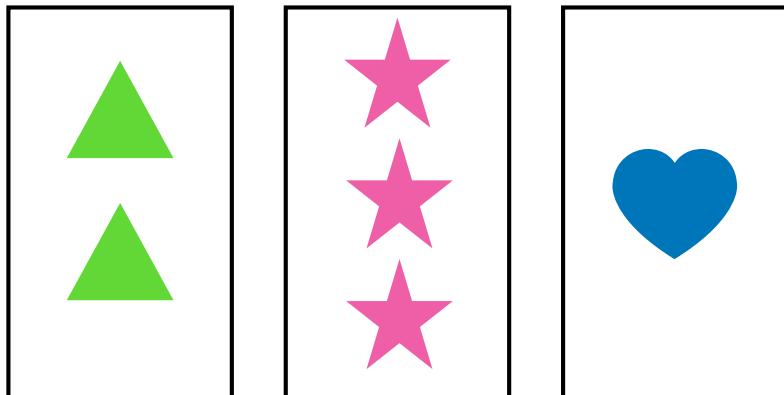


Follower

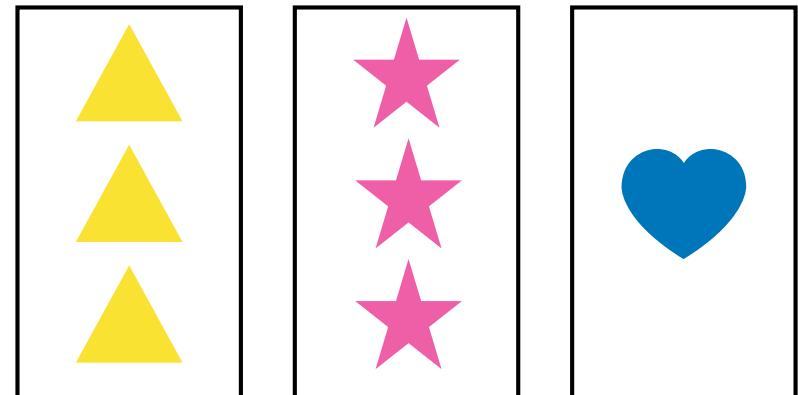


Collaboration

- Collect valid sets of three cards
- Valid: unique color, shape, and count
- Each set completed is one point
- Goal: maximize game score



✓ **Valid Set**



✗ **Invalid Set**

181 (two cards with three objects)

Collaboration

Leader

Follower



Collaboration

- Reward successful collaborations: when games go well, let them keep playing
- On each set completion: players get a point and additional turns
- So now they can play longer, complete even more sets, and get even more points
- During crowdsourcing: we get more data from effective interactions, and effective workers make more money
- This further reinforce the game incentives

Language

- Because players construct sets together, they must coordinate their actions
- Coordination is only possible via unidirectional (leader → follower) natural language instruction
- Why not bidirectional? Dramatically complicates the problem of learning agent models
- This simplification allows to study the learning problem with limited (or no) “crutches”

Instruction

- **Leader's role:** give instructions to the follower
 - Write as many instructions as they want per turn, as long as the follower has one to follow
- **Follower's role:** follow the instructions
 - Follow as many instructions as they want per turn, or take multiple turns for an instruction

Instruction-Action Alignment

- If leaders give multiple instructions, how would we know which actions correspond to which instruction
- Solution: learn it? More complexity, to already complex setup
- We use a FIFO queue to store instructions, and only show the next one to the follower
- This also prevents reasoning about future tasks when executing the current instruction

Instruction-Action Alignment

- Follower only sees the first incomplete instruction in the queue
- Follower presses “DONE” when they’ve finished a command
 - If there are more instructions in the queue, they will see the next one
 - If there are no more instructions, their turn will end
- This gives us accurate alignments
 - Follower can’t mark “DONE” early, or they risk ending their turn early
 - They can’t mark “DONE” late, because they don’t know what’s coming next, or if there’s another instruction
- Follower sees previously-completed instructions, to have access to interaction history

Incentivizing Instruction

- Players have different abilities and knowledge, and must use language to bridge those differences
- **Observability:** leader sees the whole board, follower sees a first-person view
 - Leader is responsible for planning what cards both players should get
 - Follower is disincentivized to wander off or select unmentioned cards
 - Leader's instructions need to be grounded in the follower's partial observability first-person view (e.g., contain spatial relations)
- **Action:** follower has 10 steps per turn, while leader has only 5
 - Encourages leader to delegate longer, more complex paths to the follower (i.e., more interesting language)
 - If leader ignores follower, they can't do well in the game

Leader



Grab the three red stripes behind you

Follower



Multi-turn Interaction

- Fundamental to CerealBar: interaction across multiple turns
- This allows:
 - Adaptation to the other player's behavior
 - Correction of mistakes
 - Formation of common ground

Leader view



Player Role

Leader

END TURN

Time Left: 9987
Moves: 5 Moves Left in Turn
Turns Left: 4

Score 0 --- [DONE] turn left and get the green heart and three red triangles
--- [DONE] turn left and go past the windmill for two black crosses
--- [CURRENT] then get the two blue stars
--- [NOT SENT] get the three orange squares

CAMERA

HELP

Send Command...

SEND

Follower's View:

Time, moves, and turns remaining

Command window, with instruction statuses

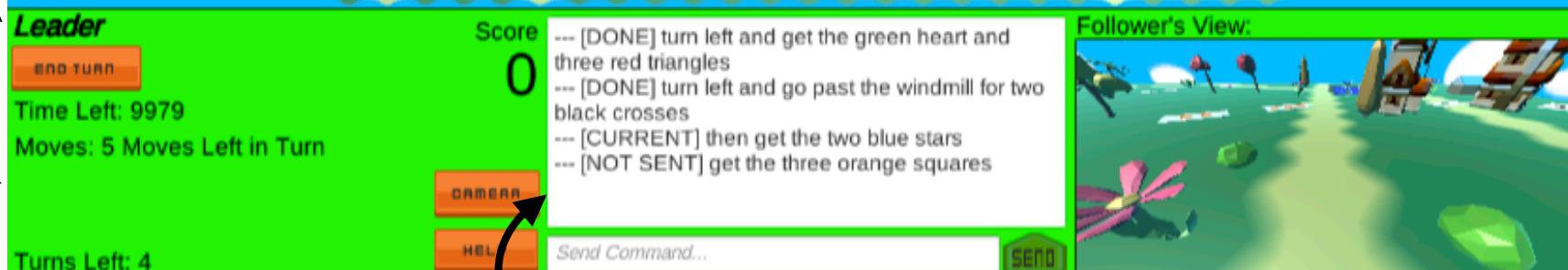
Cards in an invalid set have a red outline

Leader sees follower's first-person view

Leader view

Top-down view lets leader see occluded cards

Player Role



Time, moves, and turns remaining

Command window, with instruction statuses

Leader sees follower's first-person view

Follower view

The screenshot shows a 3D hexagonal grid environment. The player is positioned on a light green hexagon, facing towards a blue triangular card in the distance. The path is indicated by a series of hexagons leading to the card. In the background, there are trees, a street lamp, and a large pink sphere. The sky is blue with a white sun.

Player Role
(Leader or Follower)

Follower

Score 0

Time Left: 9975

Moves Left: 988 Moves Left in

Turns Left: 1000

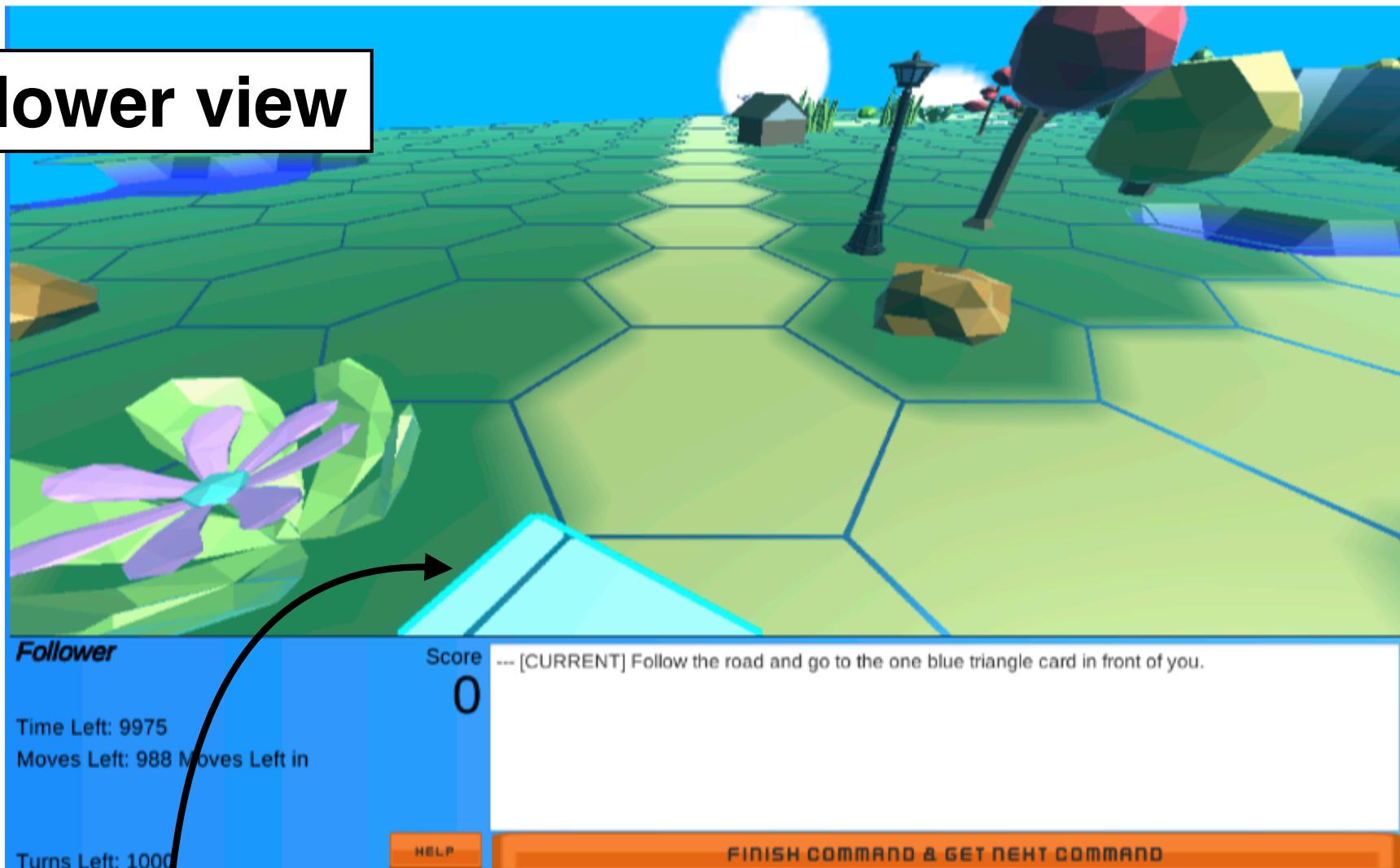
HELP FINISH COMMAND & GET NEXT COMMAND

--- [CURRENT] Follow the road and go to the one blue triangle card in front of you.

Time, moves, and turns remaining

Command window,
with instruction statuses

Follower view



Follower can't see
when current set is
invalid

- Hex grid helps follower navigate
- Not visible to leader (to avoid exact instructions)

Some Technicalities

- CerealBar is implemented in the Unity cross-platform game engine
- WebGL compilation is important for in-browser support, which is critical for crowdsourcing
- Backend multiplayer server is in Python with support for various agent controllers (human, Pytorch, etc)
- Non-supervised learning is done using a simple replication in Python, so is lightweight and fast

Tasks Studied in CerealBar

Task I: map leader instructions to follower actions

$f(\text{instruction}, \text{history}, \quad) = \text{actions}$



[Suhr et al. 2019]

Task II: generate leader instructions

$f(\quad , \text{history}) = \text{instruction}$



[Kojima et al. 2021]

Crowdsourcing

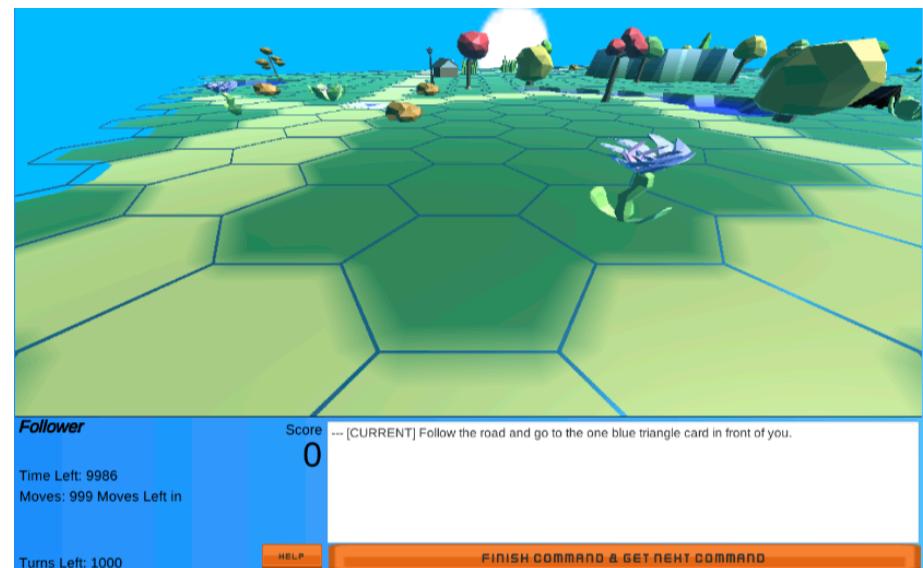
- Tutorials and qualifications
- Managing games
- Pay and incentives
- Novice and expert pools
- Communication with workers
- Filtering games

Tutorials and Qualifications

- To qualify for the CerealBar tasks, workers need to:
 - Complete a tutorial
 - Pass a short qualification quiz
 - Live in a country where English is widely spoken
- Total of 264 qualified workers

Tutorials

- Leader tutorial
 - Write an instruction
 - Pick up 3 sets
- Follower tutorial
 - Follow 3 instructions
 - Teaches users about interface and control



Qualification Quiz

- Set-making
- Player responsibilities
- Additional bonus if passing the qual + tutorials

Managing Games and Matching Players

- Server matches players and assigns roles randomly
- Players only have access to one game at a time
- Waiting room that times out if nobody else is online
- We let players know ahead of time when a batch is coming, so they plan to be online

Pay and Incentives

- Small base pay per game, regardless of their score
- Additional bonus for each point
- Bonus per point increases as they score more points → incentivized to keep playing for longer
- Both roles receive the same bonus
- Median cost per game is \$5.80

Novice and Expert Pools

- Workers start out in the novice pool, where pay is slightly lower and they are paired with other novices
- Expert pool has more HITs and higher pay per point
- After at least two games as novice, in both roles, with at least one point each → become an expert
- Separating by expertise also ensures expert workers have the best experience

Communication with Workers

- Very responsive to worker questions, comments, suggestions, etc.:
 - Email
 - MTurk forums
 - Recently: Discord

Filtering Games

- Filtered out games using heuristics to identify when guidelines were not followed
- E.g., games where the follower moved a lot before marking instructions as done
- Kept 0-score games unless violated guidelines

Data Statistics

- Total of 1,202 games and 23,979 instructions
- Median score of 9
- 24 instructions per interaction on average, median length of 13 tokens
- 8 follower actions per instruction on average
- Vocabulary size of 3,641



turn left twice and head straight , toward the dog house and look for 2 green circles to pick up



Procedure Summary

1. Game design
2. Designing qualifications and tasks
3. Pilots and refinement cycle
4. Main data collection cycle
5. Filtering data and splitting into training/test sets

Takeaways

- Directly incentivize success
- Design game so that collaboration is critical for success
- Carefully consider how design may affect language use
- Communicate with workers often

Thanks to the Team!

- Claudia Yan
- Jack Schluger
- Stanley Yu
- Marwa Mouallem
- Hadi Khader
- Iris Zhang
- Yoav Artzi

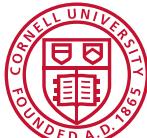
EMNLP 2021 Tutorial

Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection

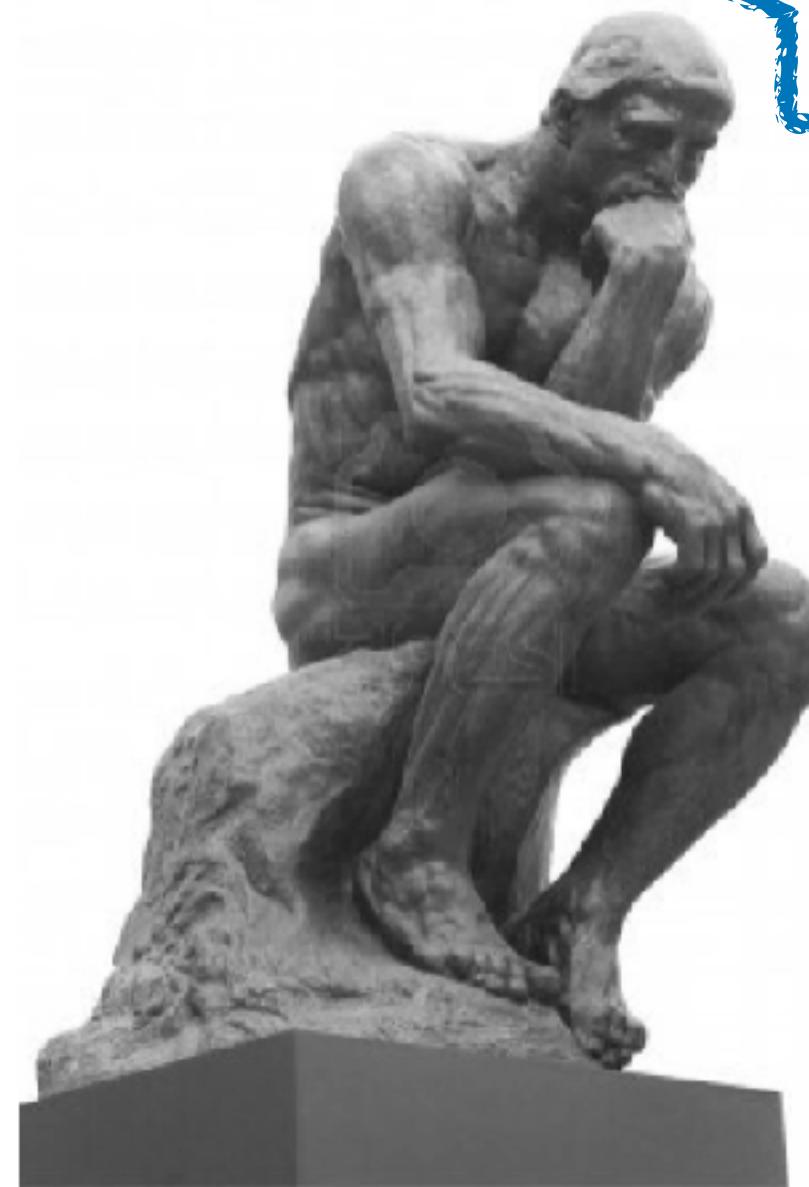
Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar,
Sam Bowman, and Yoav Artzi

Case Study IV: QuAC

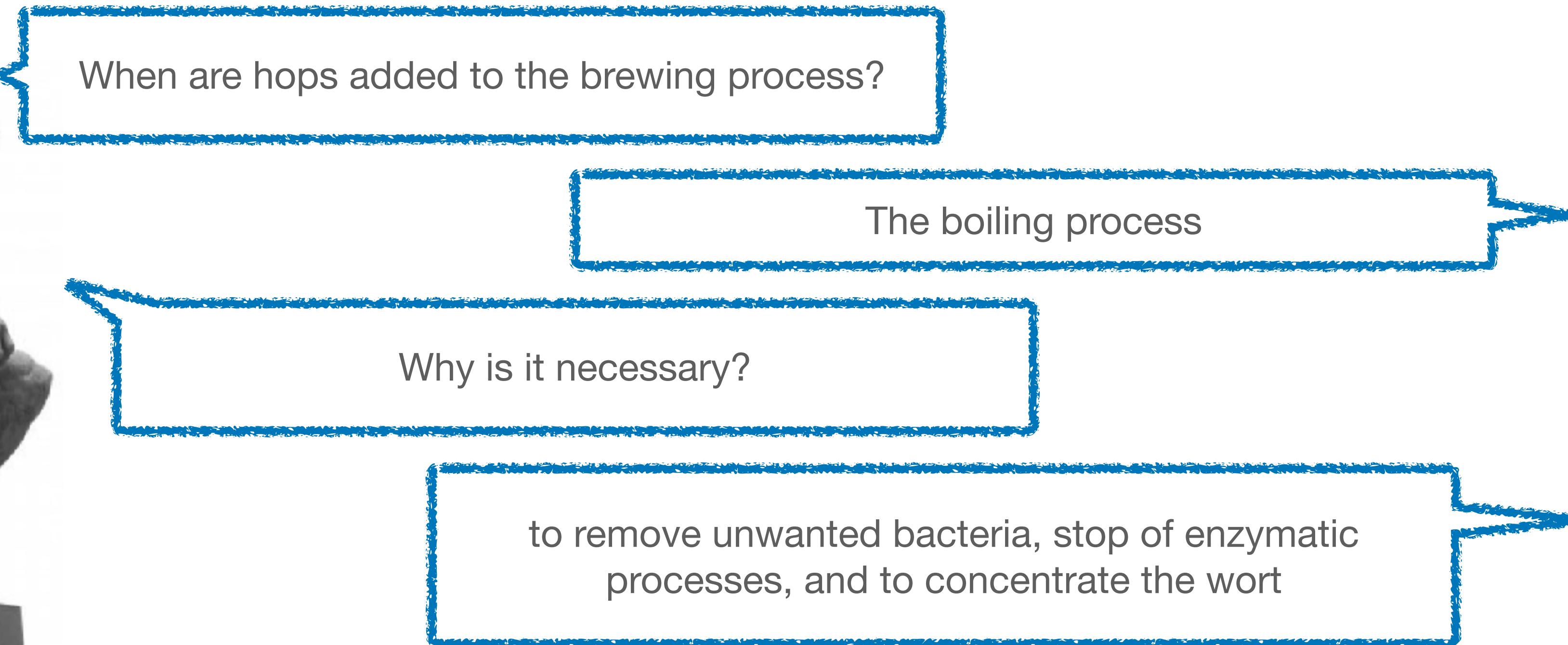
Presented by Mark Yatskar



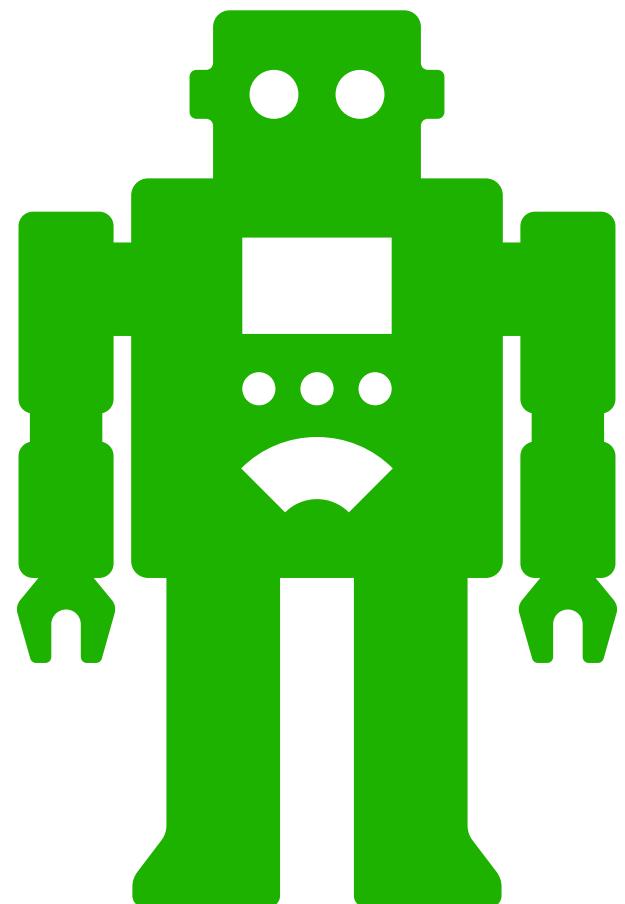
Multiple Turn Question Answering



Questioner



Beer Brewing



Model

Single Turn:
SQuAD dataset [Rajpurkar et al 2016], TriviaQA, NewsQA, RACE, etc

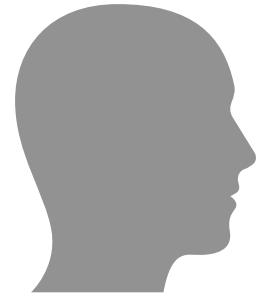
Multiple Turn:
QuAC, CoQA [Reddy et al. 2019], ect.



QuAC Formulation

Given:

entity name and the first paragraph of Wikipedia page



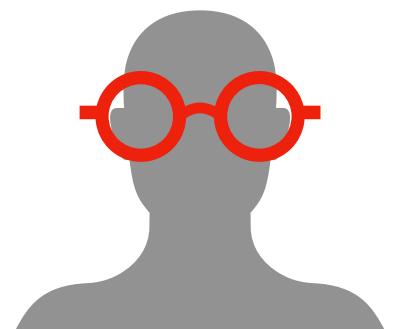
Student

Topic: title of section about entity**Do:**

Ask questions to learn as much as possible about this topic!

Given:

entity name and first paragraph of Wikipedia page, text of topic section



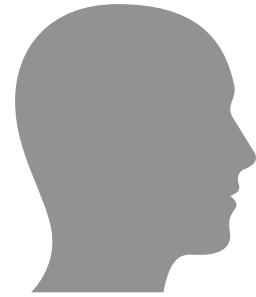
Teacher

Do:

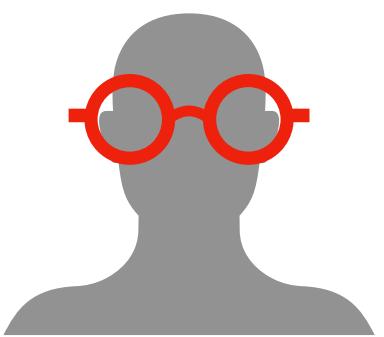
Answer the questions by choosing a span or return 'cannot answer'

Example Dialog on Daffy Duck's Origin

Q: What is the origin of Daffy Duck?



Student

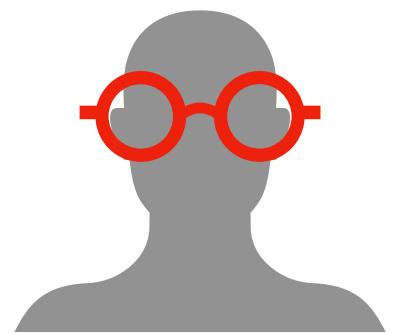
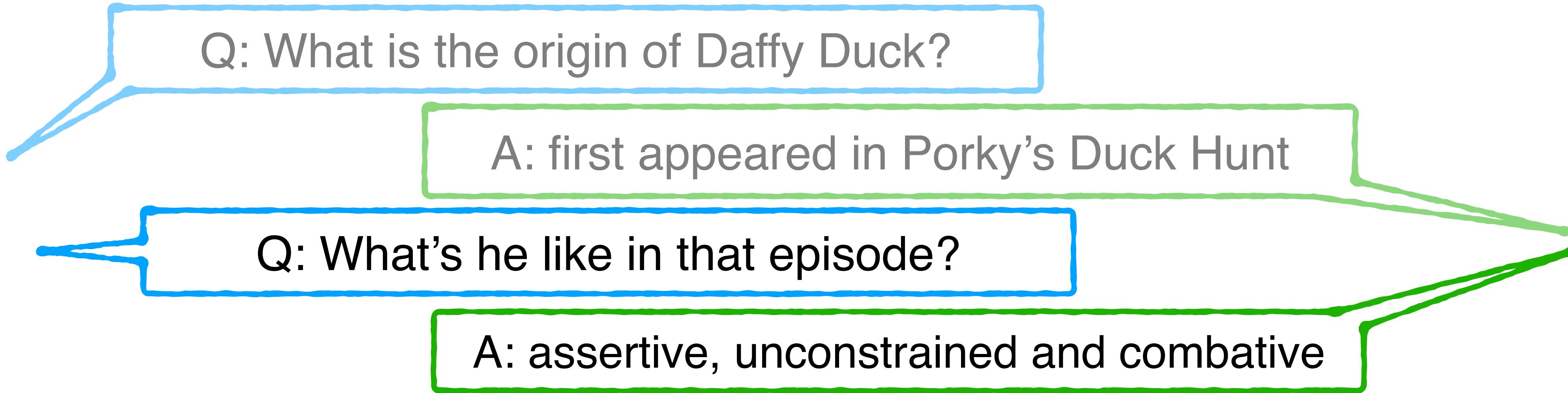


Teacher

Example Dialog on Daffy Duck's Origin

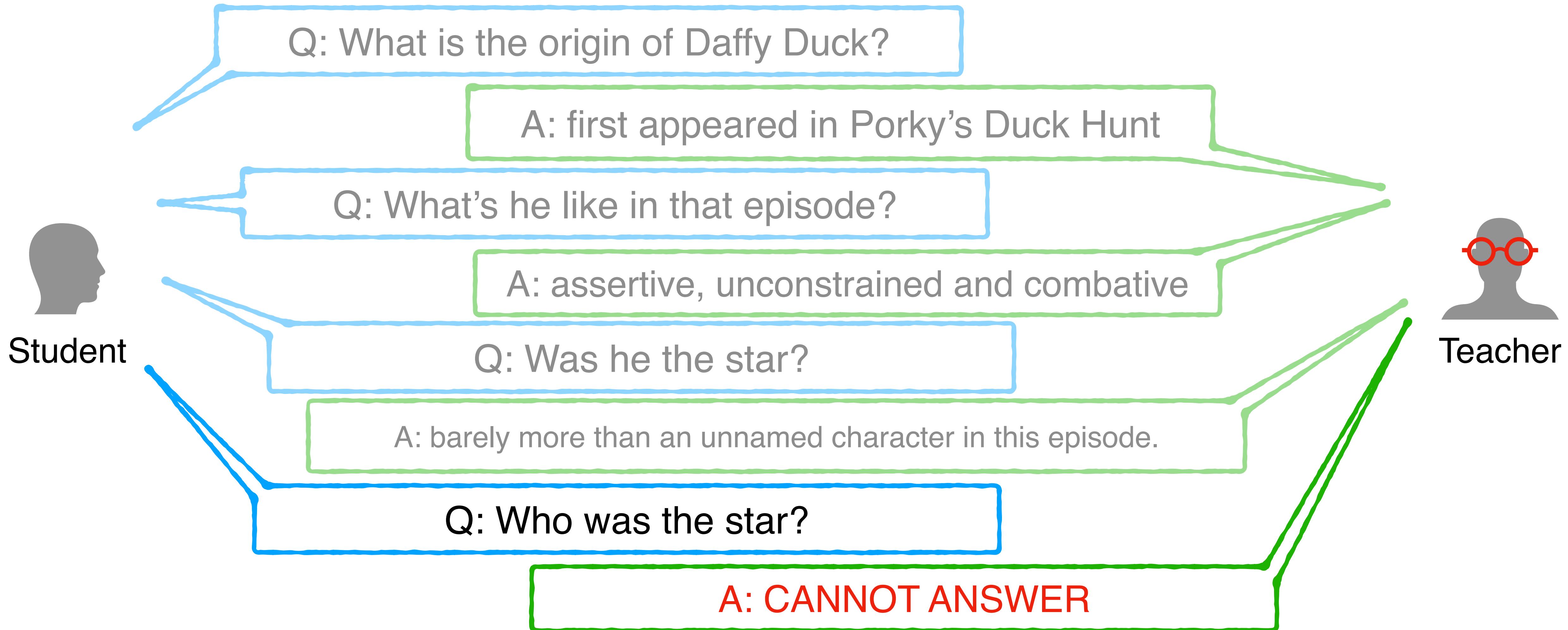


Student

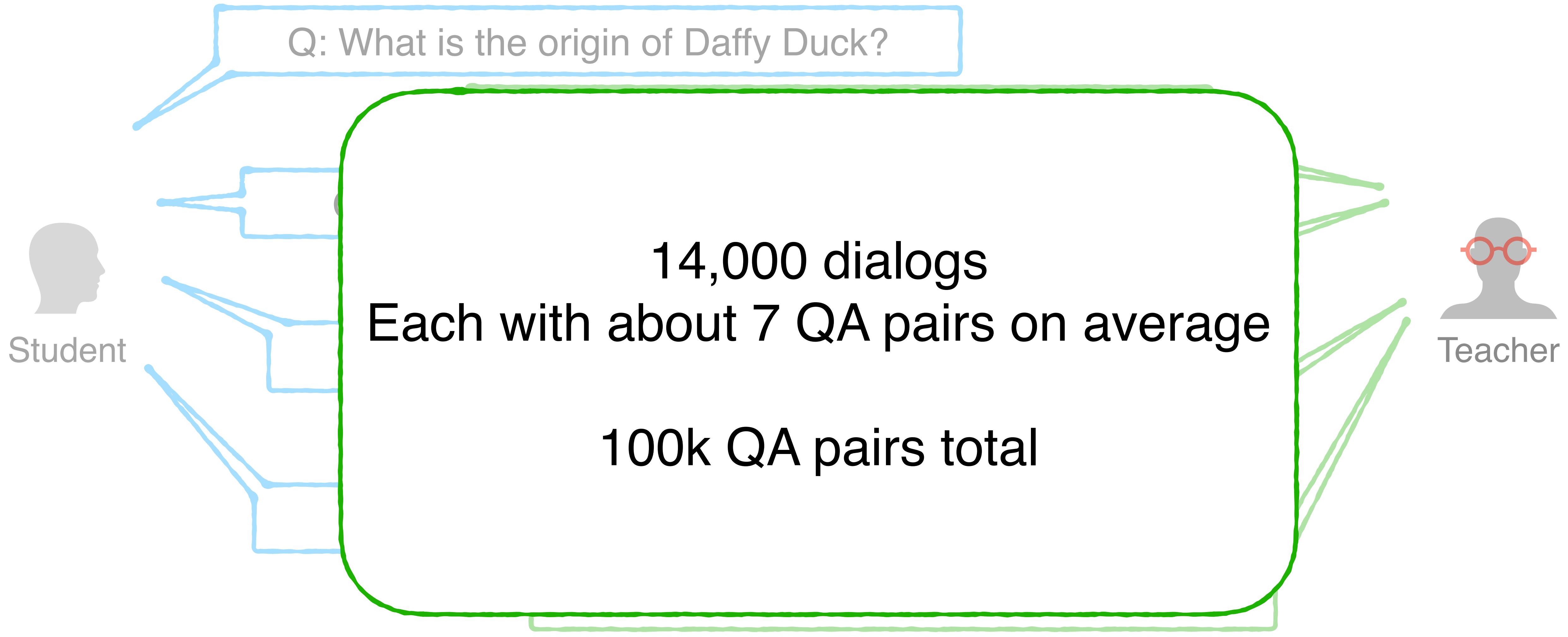


Teacher

Example Dialog on Daffy Duck's Origin

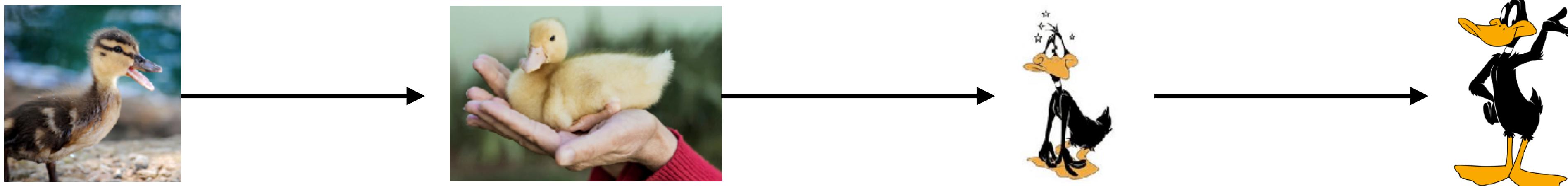


Example Dialog on Daffy Duck's Origin

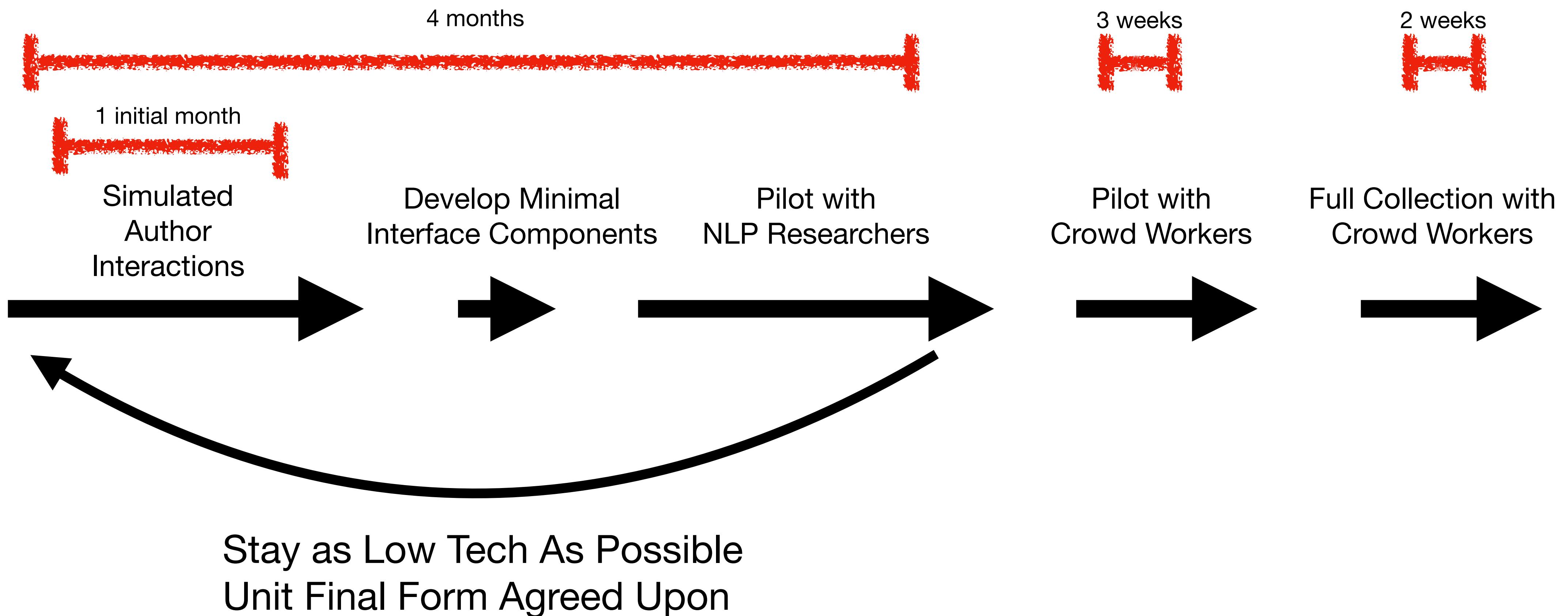


Case Study Focus

- Progression of QuAC from early data collection attempts to final form
- What didn't work
 - How did we decide something was failing?
 - How did we adapt?



Crowdsourcing Process



Initial Goals

- An **unconstrained conversation** between a teacher and student driven by a curiosity on any topic on Wikipedia
- After the conversation, the student has **demonstrably learned** important information about the topic

Concerns

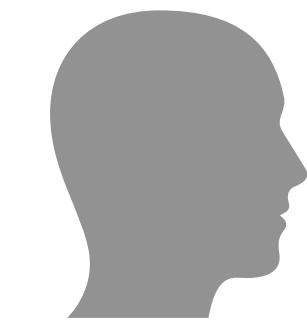
- Can we define a task that isn't too burdensome?
 - How long will it take annotators to have a dialog?
- Can we capture enough teacher-student behavior to make data challenging and realistic?
 - Will the dialogs be long enough?
- How do we avoid biases in the data?
 - Will the data be immediately solved by community?

Avoiding Biases

- The student should **not see supporting information** that teacher is using to answer questions[1]
- Teacher must avoid telling the student content unrelated to their questions



QuAC, Version 0



Student

Given:

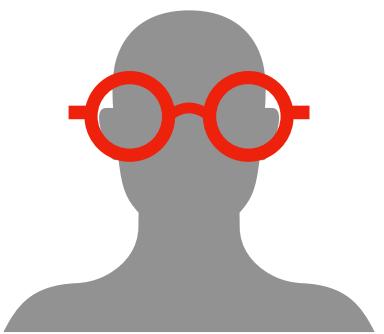
The name of a wikipedia article

Task 1:

Learn as much as possible by asking questions

Task 2:

Answer a quiz of the teacher's design



Teacher

Given:

A whole wikipedia article

Task 1:

Read article

Task 2:

Design a quiz.

Task 3:

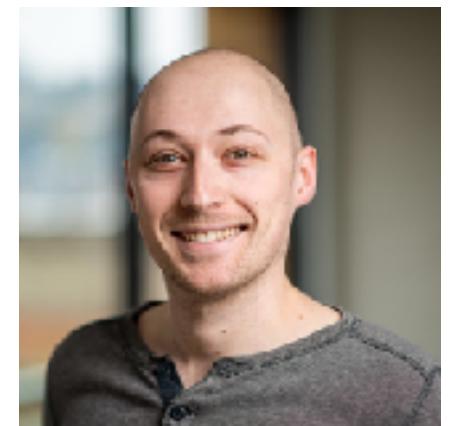
Answer, in free form, students questions.

Office Conversations

- A text file and a Slack chat window



Student



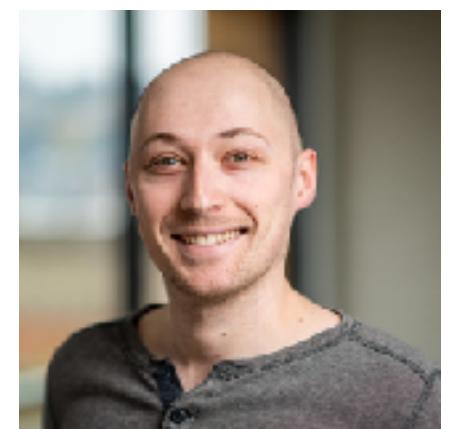
Teacher

Office Conversations



Student

Sure, spiders are cool.
What's special about orb spiders?



Teacher

They are a common spider that builds circular webs

Ok, what do they eat?

...

Office Conversations



Student

I'm tired. Give me your quiz.

This is horrible. And I've wasted 15 minutes

No.

Great! you got 1 out of 10 questions right!

You didn't ask about anything important!
Why don't you like animals?
I've been doing this for an hour!
Did you at least like my quiz?



Teacher

Let's go get snacks and try this with someone else. You just weren't trying.

Modified Goals

- An **unconstrained conversation** between a teacher and student driven by a curiosity on **any simple** topics
- After the conversation, the student has **demonstrably learned** important information about the topic

Justification: Less than 50 office conversations



QuAC, Version 1

Given:

The name of a wikipedia article

Task 1:

Learn as much as possible by asking questions



Student

Given:

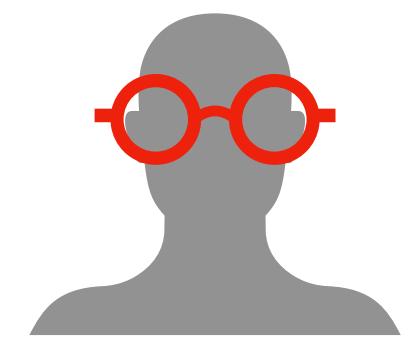
A whole wikipedia article

Task 1:

Read article

Task 3:

Answer, in free form, students questions.



Teacher

Pilots with Non-Author Researchers

- Assemble two ‘workers’ and QuAC team in one room
 - Watch: How did workers spend their time?
- 15 minutes post-mortem interview
 - Questions:
 - What did you like?
 - What did you find hard?
 - What didn’t work?

Qualitative Example

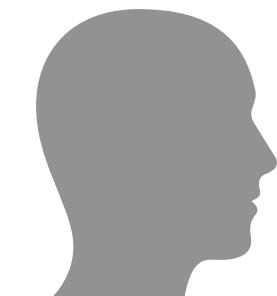
Topic: Daffy Duck

Reads Wikipedia Page

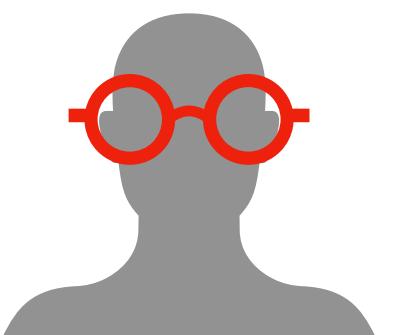
5 minutes

Q: When was Daffy Duck created?

A: 1937



Student



Teacher

2 minutes

Q: Did his appearance ever change?

2 minutes

A: Cannot Answer

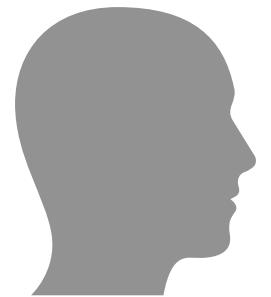
2 minutes

2 minutes

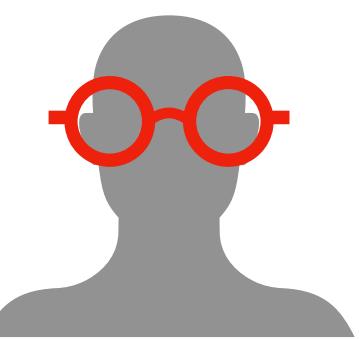
A: Cannot Answer

Feedback

- Waiting a lot for partner
- Each exchange is hard to execute



Student



Teacher

Really hard to ask good questions

Unable to communicate information to help
the student ask useful questions

Restating information in article, and typing
it in the text box. Waste of time.

Unconstrained Conversation

- Student

Really hard to ask good questions

- Give background on topic (i.e Daffy Duck first paragraph)
- Topic is narrowed to a single wikipedia section (i.e. Daffy Duck History)

- Teacher

Just restating information in the article by typing it in the text box.

- Select spans from the article

~~Curious~~ Guided Student

Unable to communicate information to help the student ask useful questions

- Problem: Too much information introduces bias
- Solution: Limit the bandwidth of guidance
- Teacher
 - Each response has a categorical label how important current line of questioning is, and if it should continue

Definitely Follow Up

Maybe Follow Up

Don't Follow Up

Initial

- An unconstrained conversation between a teacher and student driven curiosity on any topic
- After the conversation, the student has successfully learned important information about the topic

Final

- An **extractive** conversation between a teacher and student **guided by teacher** on **simple** topics

Justification: Less than 50 office conversations

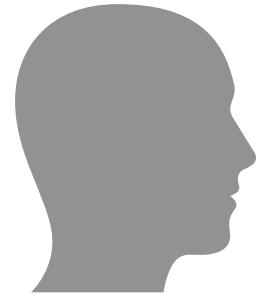
Less than 25 pilot studies



QuAC

Given:

Entity, first paragraph of Wikipedia page, and name of teacher's section



Student

Task: Read paragraph**Task:**

Ask questions to learn as much as possible about teacher's section!

Given:

Entity and a Wikipedia section

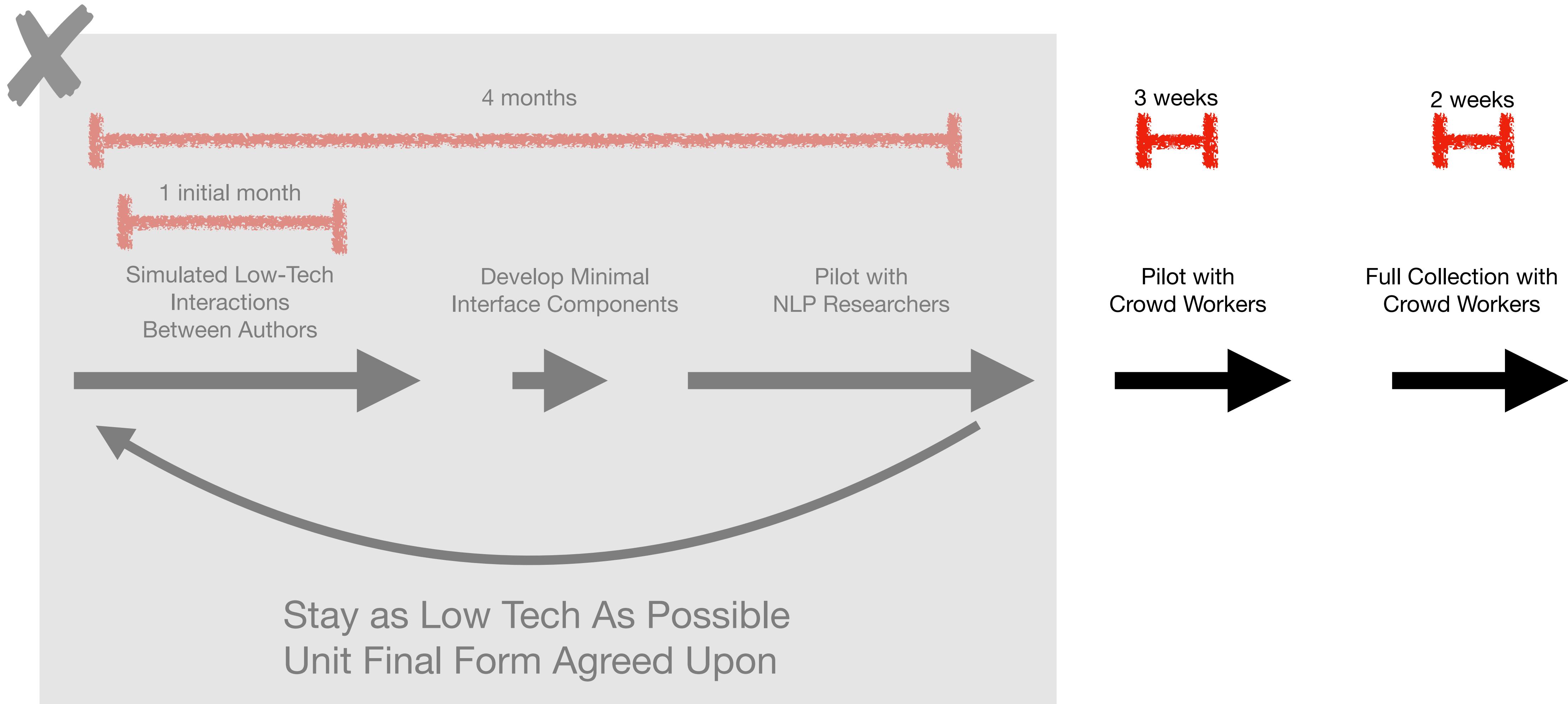


Teacher

Task: Read section**Task:**

Answer the questions by choosing a span or return 'cannot answer'. Inform student if they should follow up.

Crowdsourcing Process



Moving to Amazon MTurk

- QuAC modified existing interface for 2017 paper on collaborative dialogs[1]:
 - 1.Serve external annotation application using Flask web server
 - 2.Crowd workers accept a task and are routed to application
 - 1.Wait for a partner
 - 2.Randomly be assigned to Teacher or Student
 - 3.Complete a QuAC dialog and receive a hash code
 - 3.Enter hash code on MTurk for payment

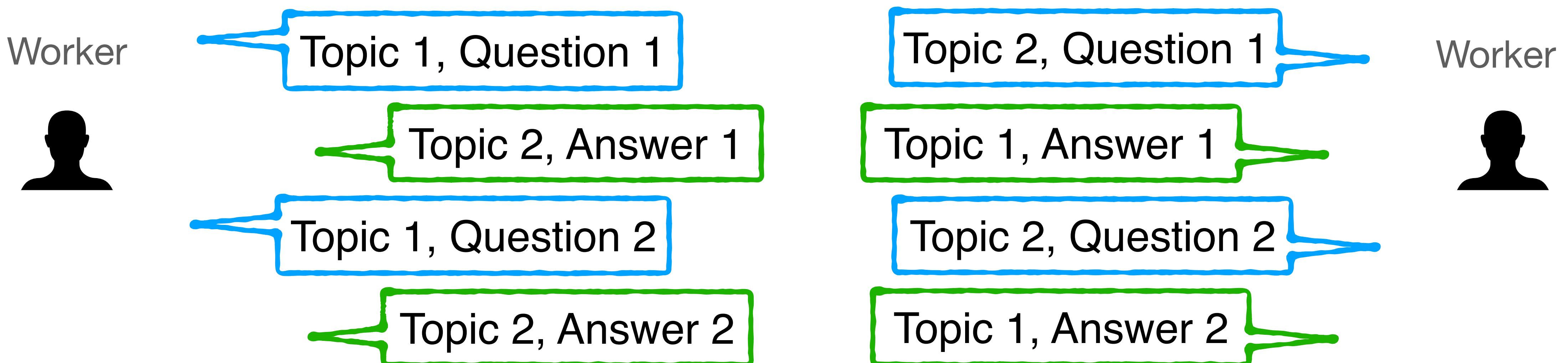
Results: A Lot of Very Frustrated Turkers



- Problem: Asymmetry meant that teachers waited for students, then students for teachers
- Conversations were abandoned at high rates, out of boredom, with few exchanges (reported on forums)
- Solution: give workers more to do, better incentives

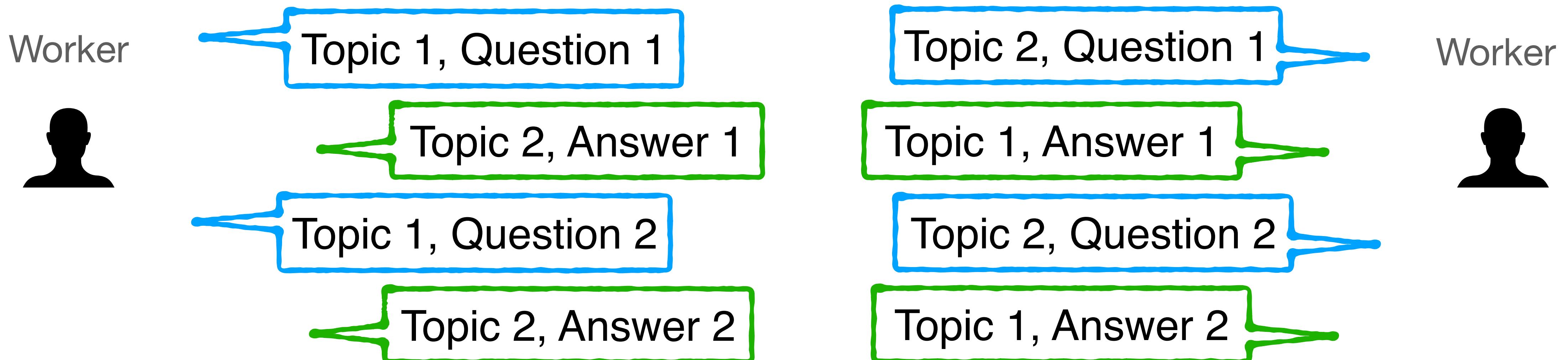
Parallel Collection

- Have a worker play both teacher and student together
- Same article, different paragraphs as student and teacher
- Interface alternates between roles



Advantages

- Work is predictable
 - When, waiting for your teacher to answer, you can answer



Compensation

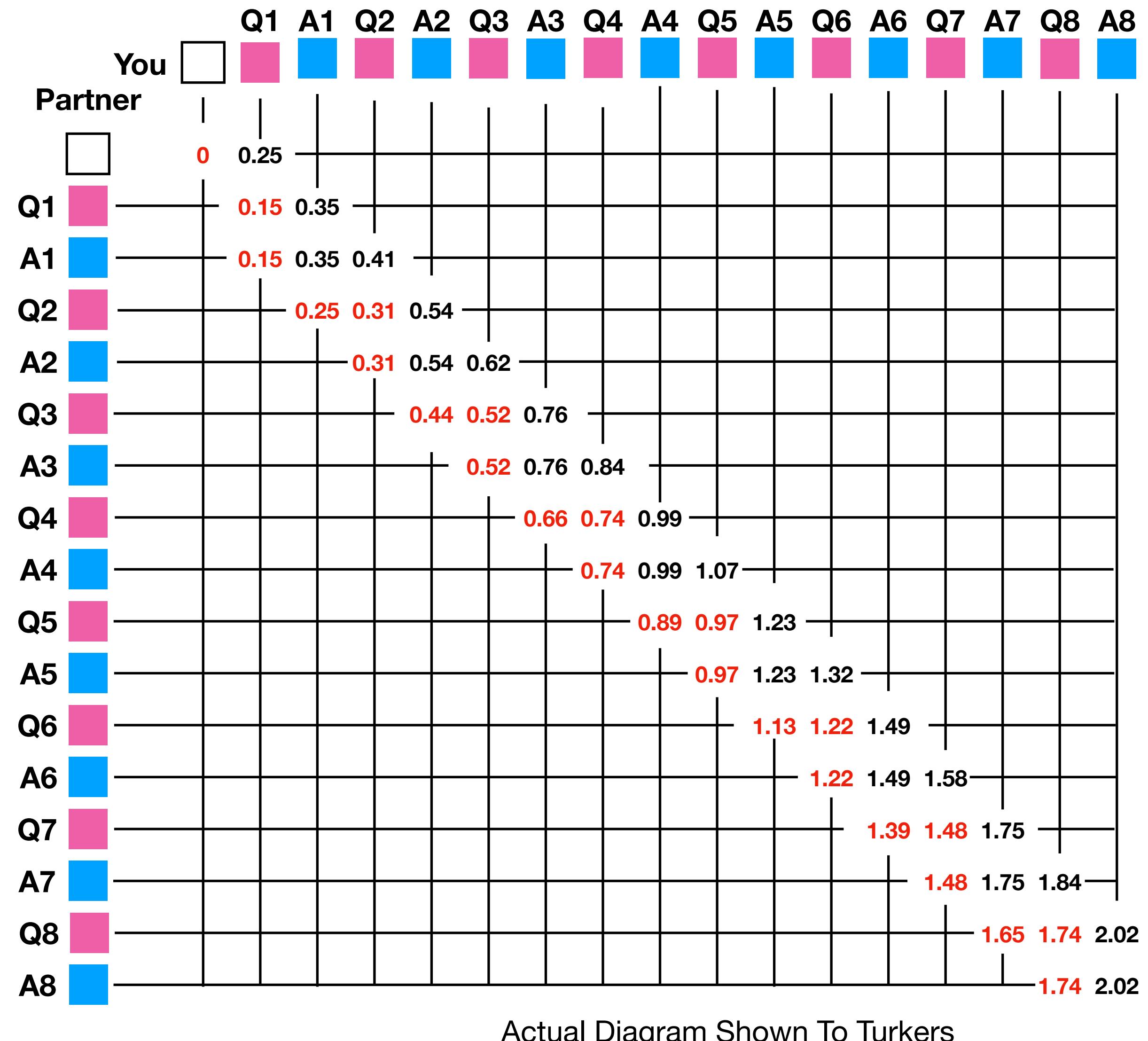
- Initial interaction worth .25\$
- Reward grows polynomially with length
 - First answered question = .10\$
 - Last answered question = .18\$
 - Encouraged long conversation
- Punitive
 - Unanswered question? = -.10\$
 - Discourage wasted partner effort

$$\text{Reward \$} = .20 * \min(Q, 1) + .05 * (Q^{1.2}) + .10 * (A^{1.2}) - .10 * U$$

Q = # of questions you asked

A = # of partner's questions you replied to

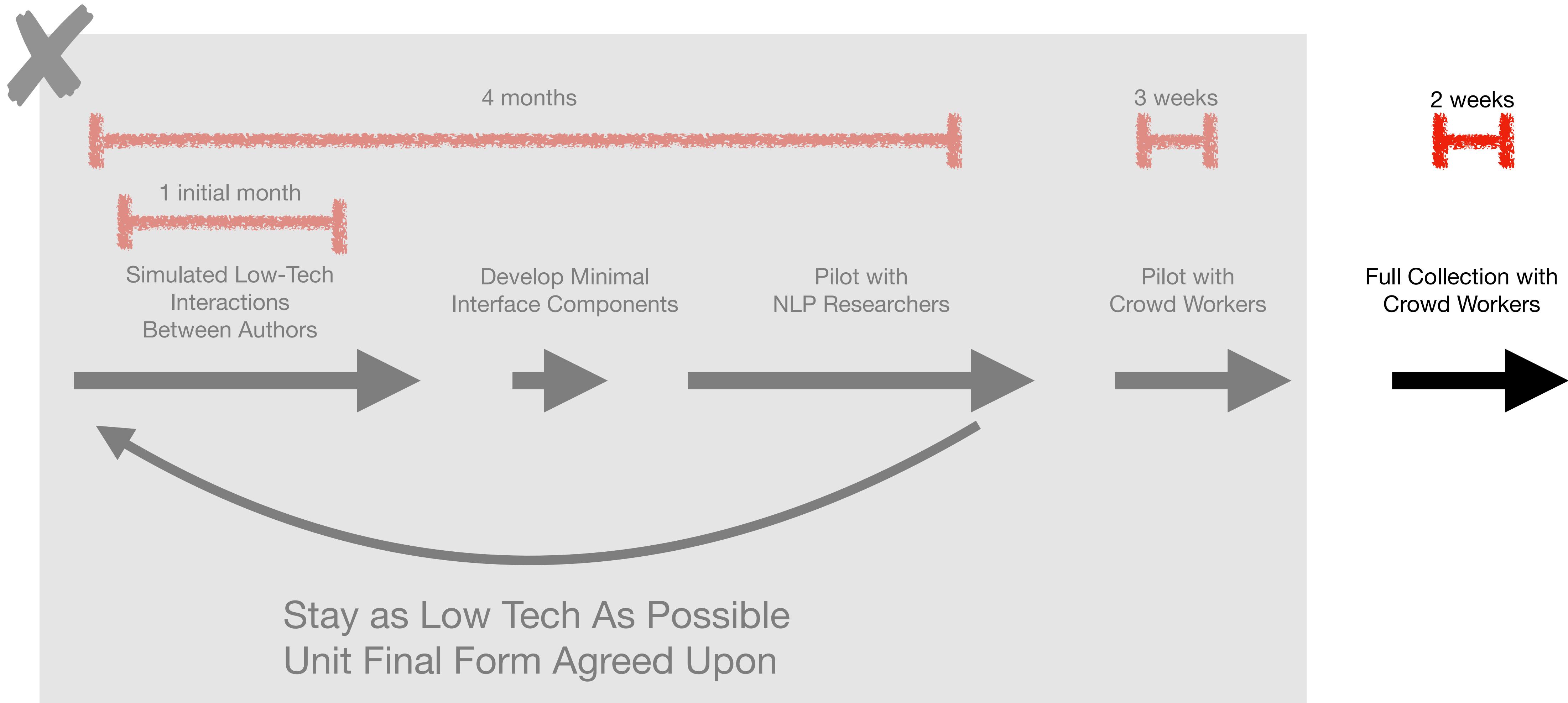
U = # of partner's questions left without reply



Pilot: How Big?

- ~100 Dialogs, less than 1,000 \$ USD
 - Enough to be certain that there are no bugs in interface
 - Enough to compute statistics and figures to be assured of quality

Crowdsourcing Process





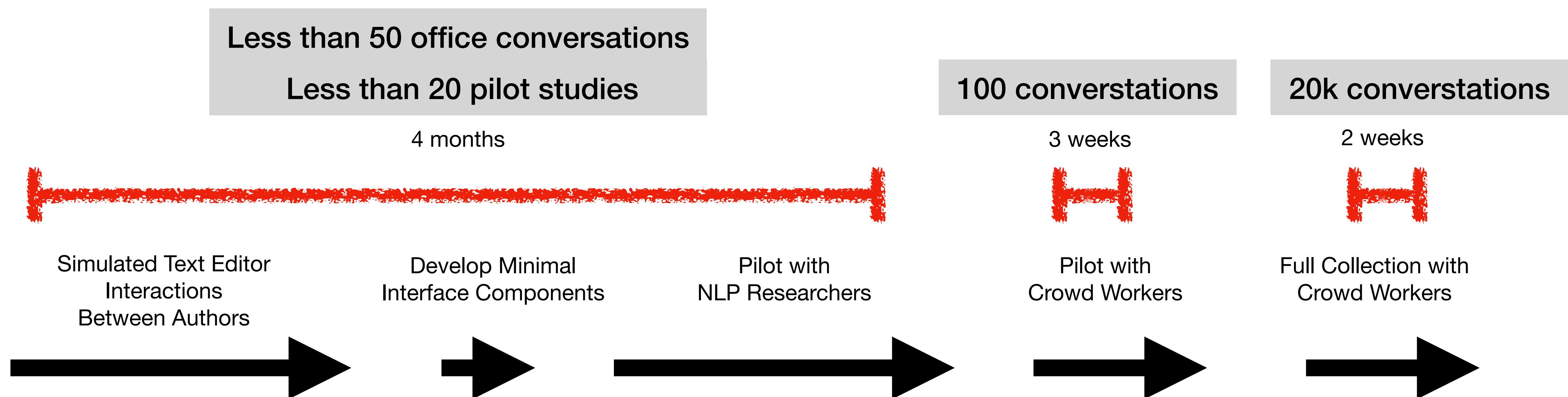
Full Crowdsourcing

- Pool of workers rapidly increased
 - Added a qualification where new workers have 1 conversation
 - Authors manually inspect this conversation, and if its long enough and not repetitive, we new workers are qualified



Conclusions

- Most designs decisions done with very little data, on high quality feedback in first 4 months
- Once moved to Amazon MTurk, a few tricks, but largely stable after pilots



EMNLP 2021 Tutorial

Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection

Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar,
Sam Bowman, and Yoav Artzi

Case Study V: SocialIQA

Presented by Maarten Sap



ML²



Penn
UNIVERSITY OF PENNSYLVANIA

Social Intelligence



Alex spilt food all over the floor and it made a huge mess.

Physical: Where will the food land?

Social: What will Alex want to do next?

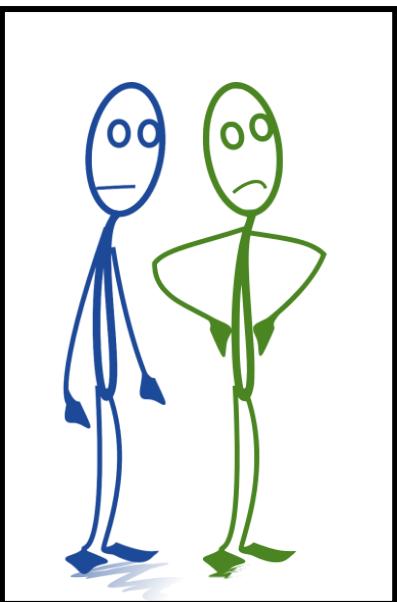
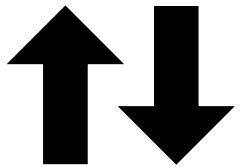
run around in the mess

less likely

mop up the mess

more likely

Social Intelligence



Tracy accidentally pressed against Austin in the small elevator and it was awkward

Why did Tracy do this?

flirt with Austin

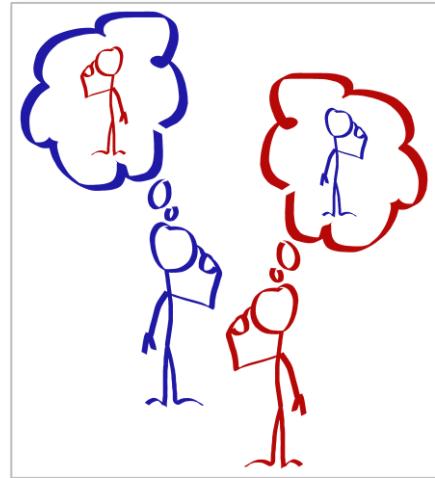
less likely

squeeze into the elevator

more likely

Models need to reason about social situations to properly interact with us

- Humans have **Theory of Mind**, allowing us to
 - make inferences about **people's mental states, next actions**
 - navigate social situations seamlessly [Moore '13]
- **AI systems lack social and emotional intelligence**
 - Pretraining on large text corpora ≠ commonsense
 - **reporting bias limits** the scope of knowledge learned [Mitchell '11; Gordon & Van Durme '13; Lucy & Gauthier, '17]
 - only find **complex correlational patterns** in training data [Davis and Marcus '15; Lake et al. '17; Marcus 2018, Talmor et al. '19]



SOCIAL IQA: the first large-scale benchmark to **quantify NLP models' ability to reason** about social situations

Related commonsense benchmarks

Winograd Schema Challenge
Levesque '11

COPA
Roemmele et al. '11

Physical IQa
Bisk et al. '20

Story Commonsense
Rashkin et al. '18

CommonsenseQA
Talmor et al. '19

SOCIALIQA
Sap, Rashkin, et al '19

Related commonsense benchmarks

small
scale

Winograd Schema Challenge
Levesque '11

COPA
Roemmele et al. '11

Expert curated, but small scale (less than 500 examples) poses challenge for training NLP systems

large
scale

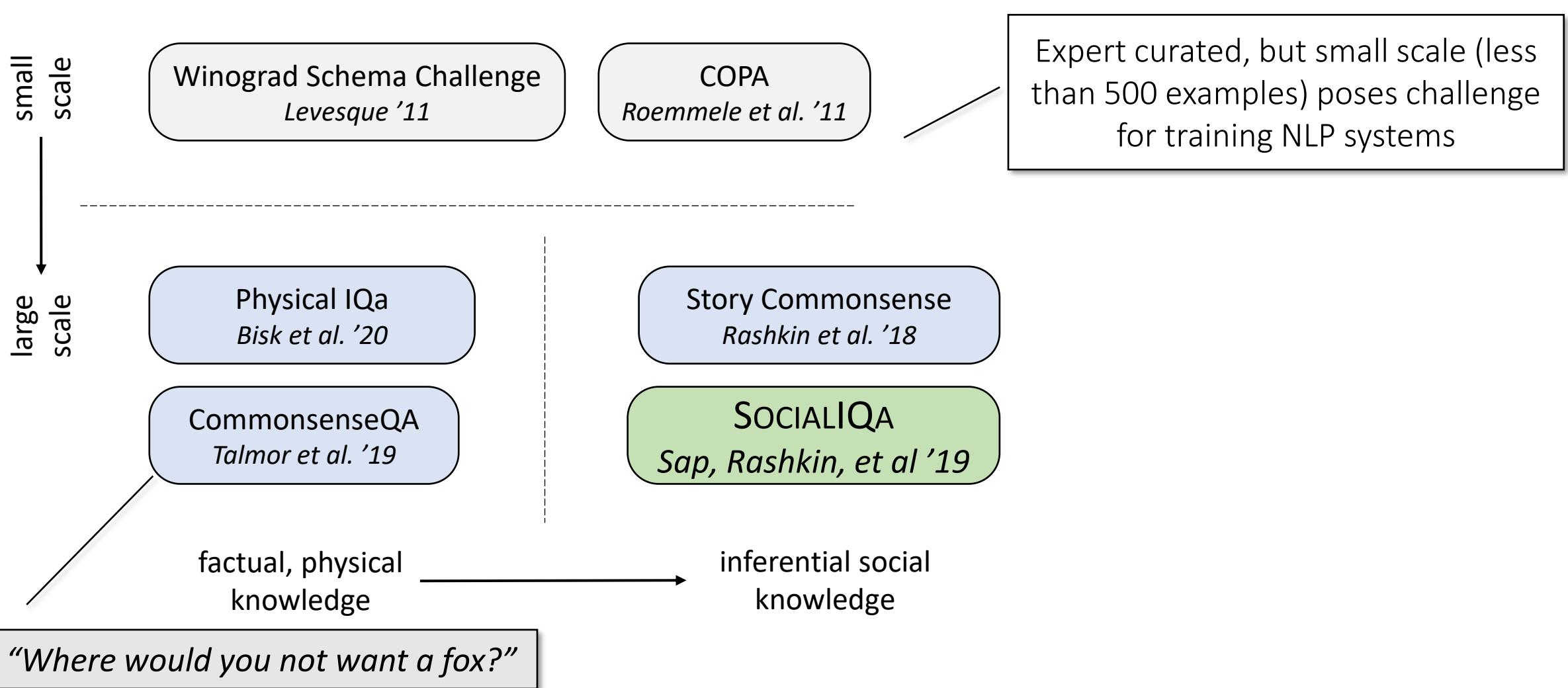
Physical IQa
Bisk et al. '20

Story Commonsense
Rashkin et al. '18

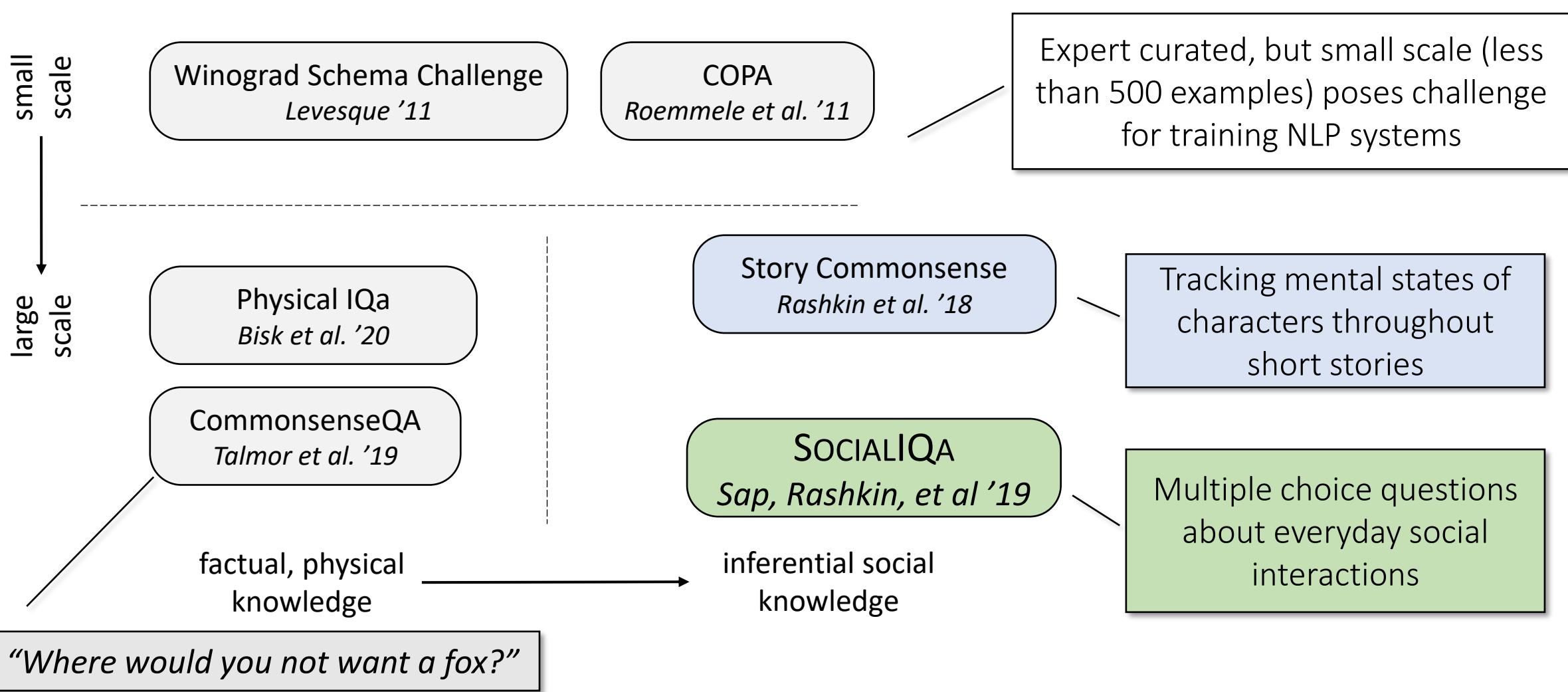
CommonsenseQA
Talmor et al. '19

SOCIALIQA
Sap, Rashkin, et al '19

Related commonsense benchmarks



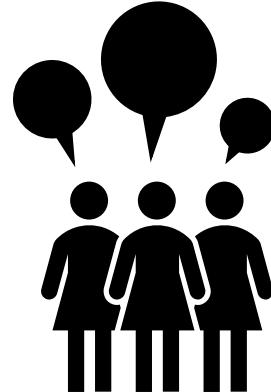
Related commonsense benchmarks



Outline

Creation of a large-scale benchmark

- How we overcome challenge of annotation artifacts



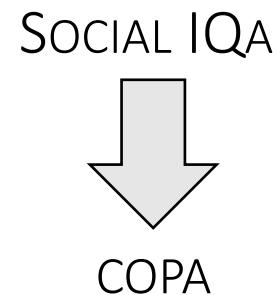
Modeling experiments on SOCIAL IQA

- GPT/BERT performance well below humans



SOCIAL IQA as transfer learning resource

- New SOTA on COPA and WSC



How do we create a benchmark like this?

Goals

M/C QA w/
leaderboard

Easy to
compare
models

Multi-stage adversarial
crowdsourcing pipeline

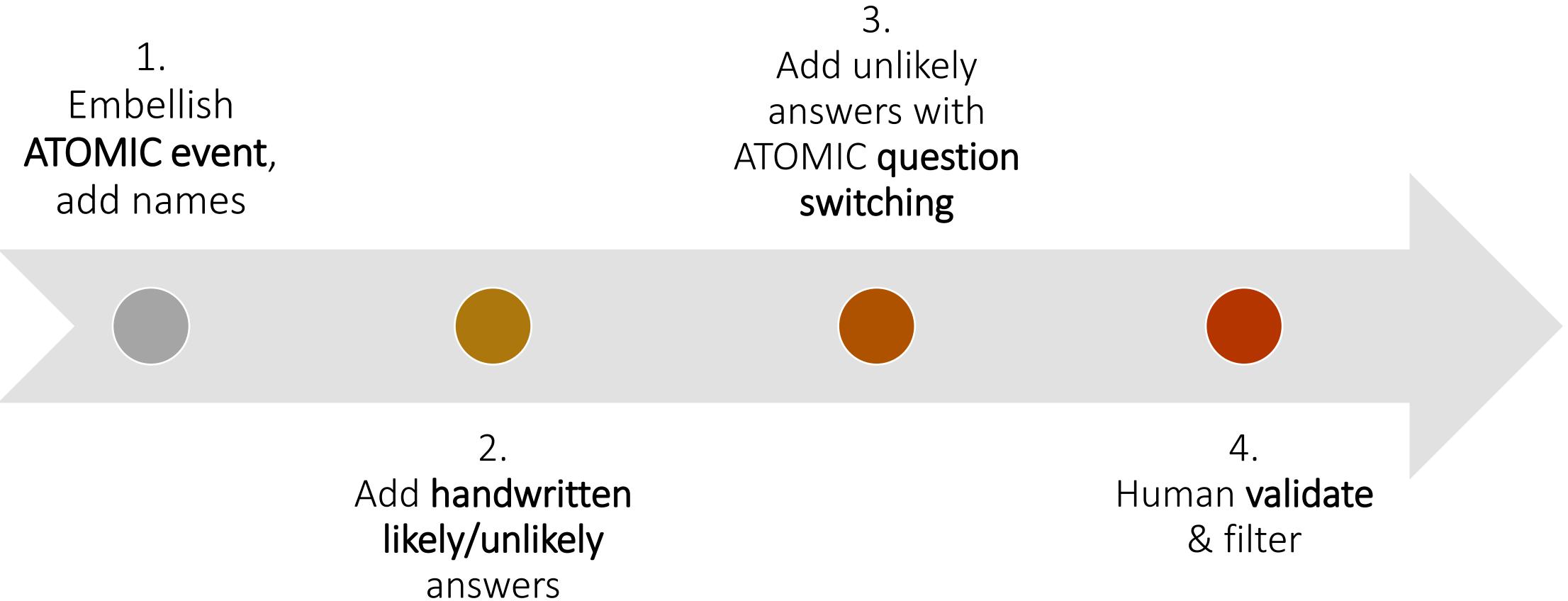
Challenging

Use ATOMIC
commonsense
resources to scale-up

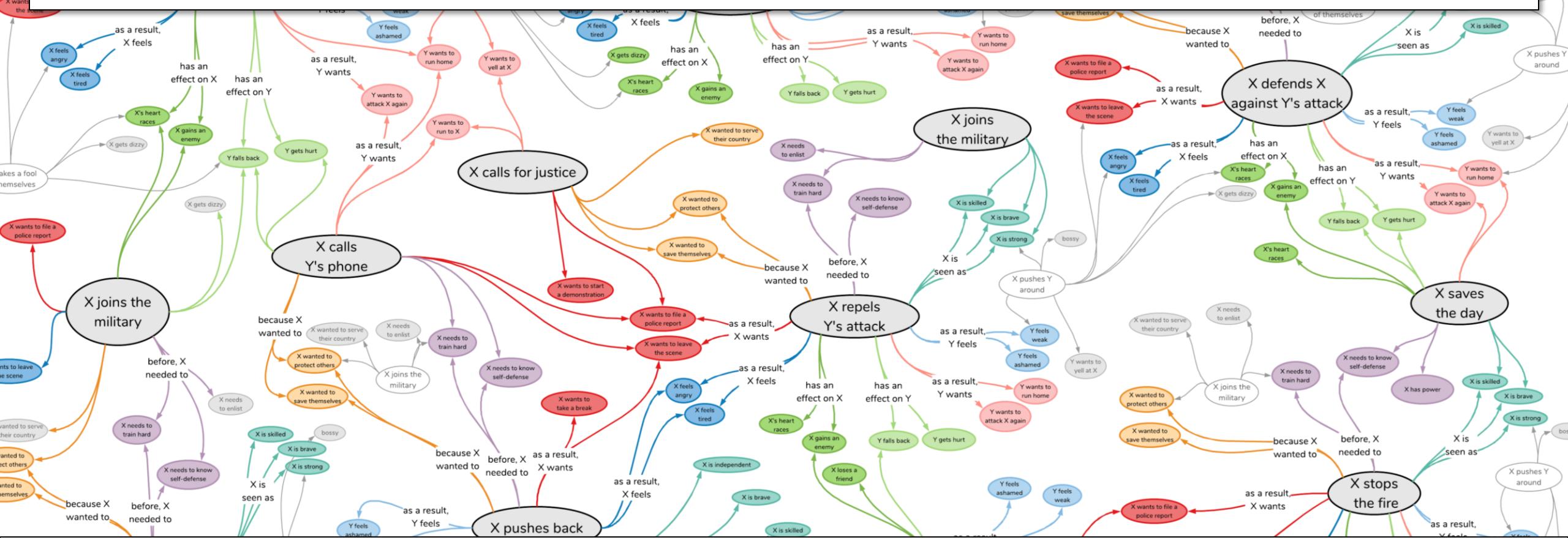
High
Coverage

Large-scale

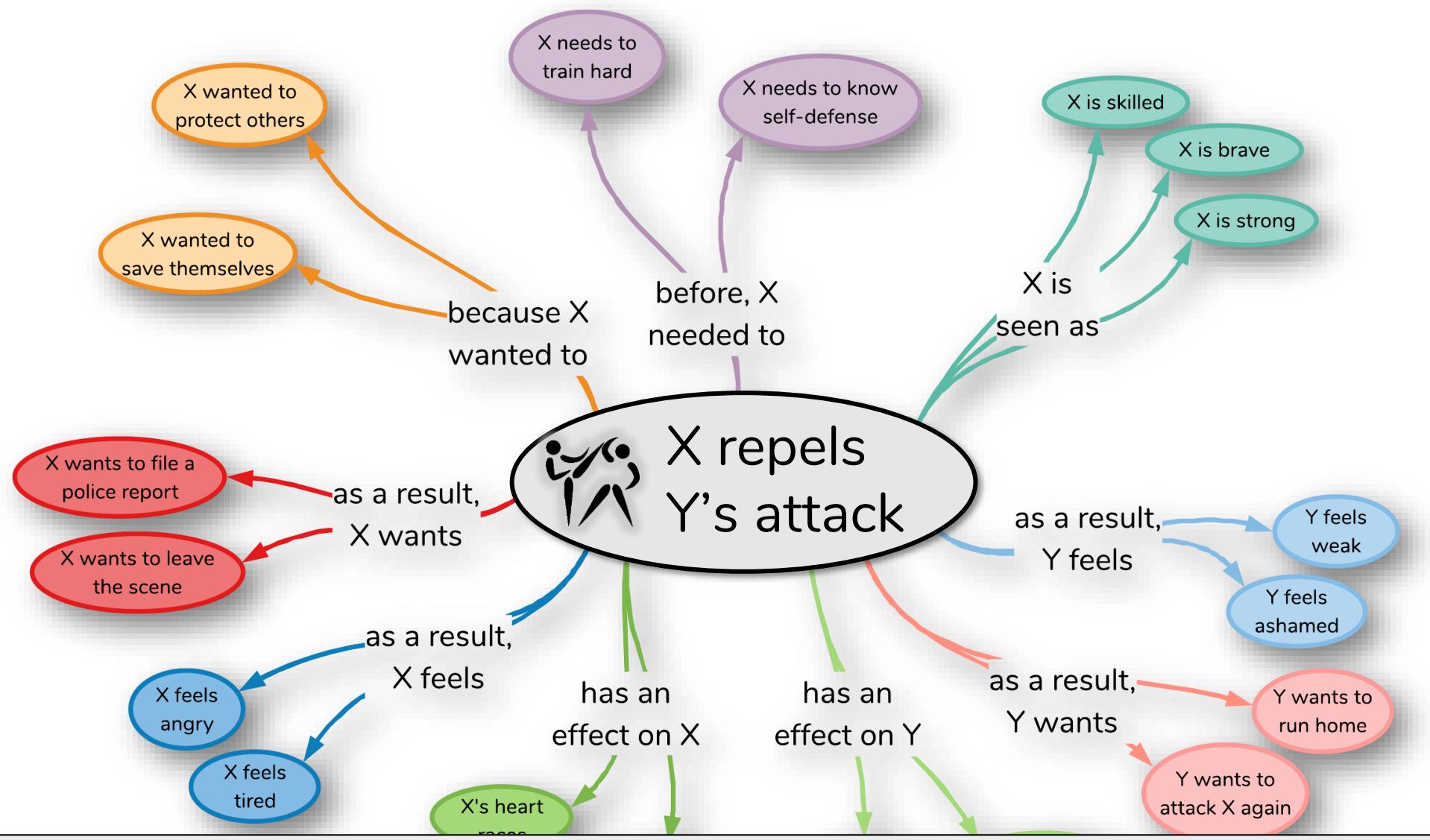
Crowdsourcing pipeline overview



ATOMIC: ATlas Of MachIne Commonsense

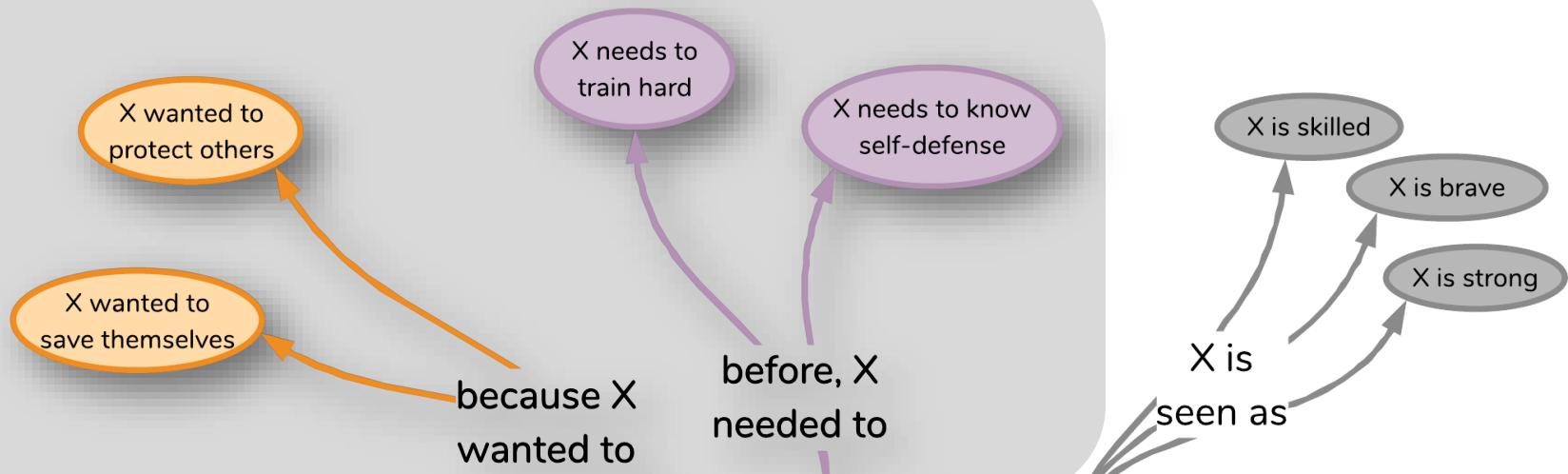


880,000 knowledge triples for AI systems to reason about the
causes and *effects* of everyday situations [Sap et al. '19]

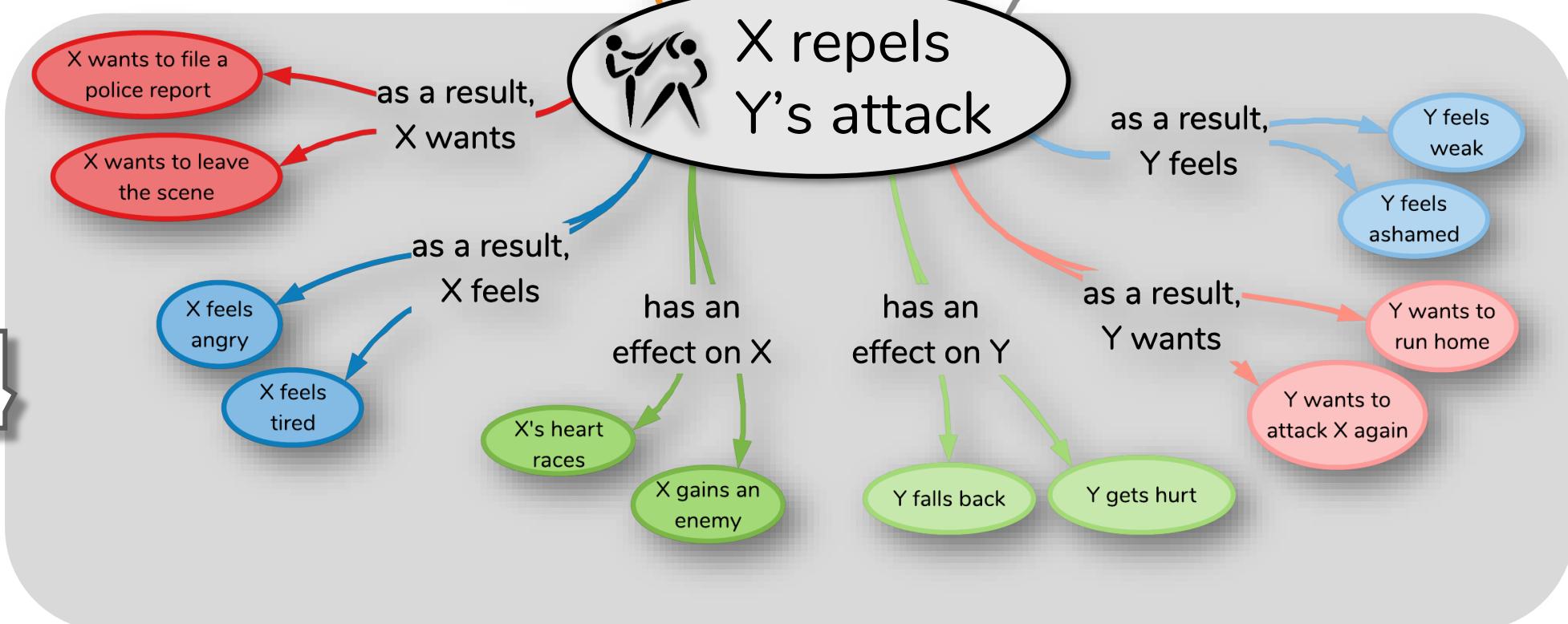


Knowledge structure: event triples with nine inference dimensions

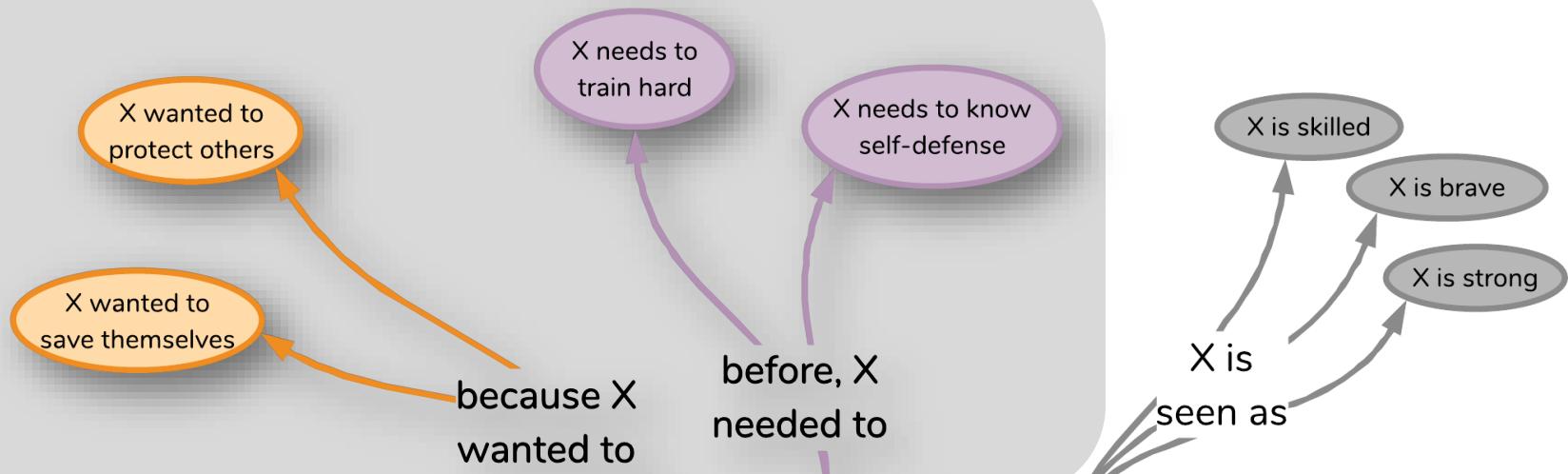
Causes



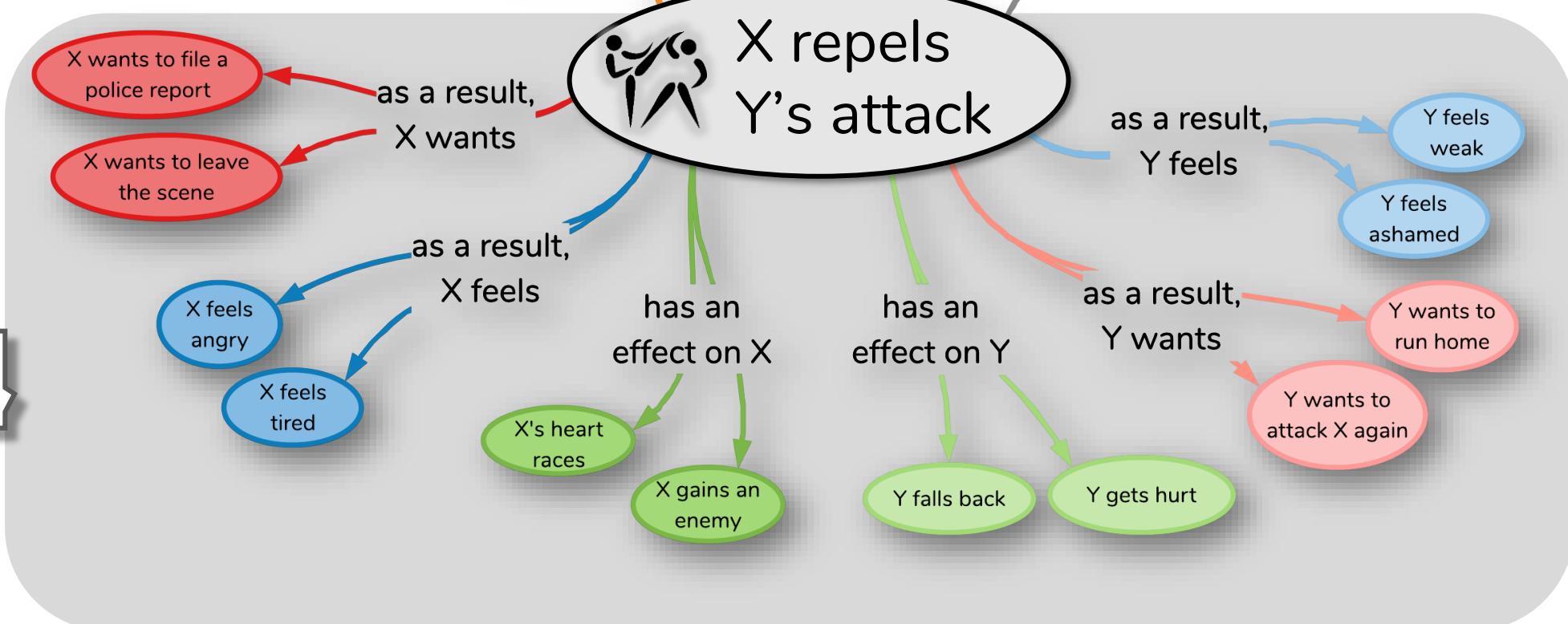
Effects



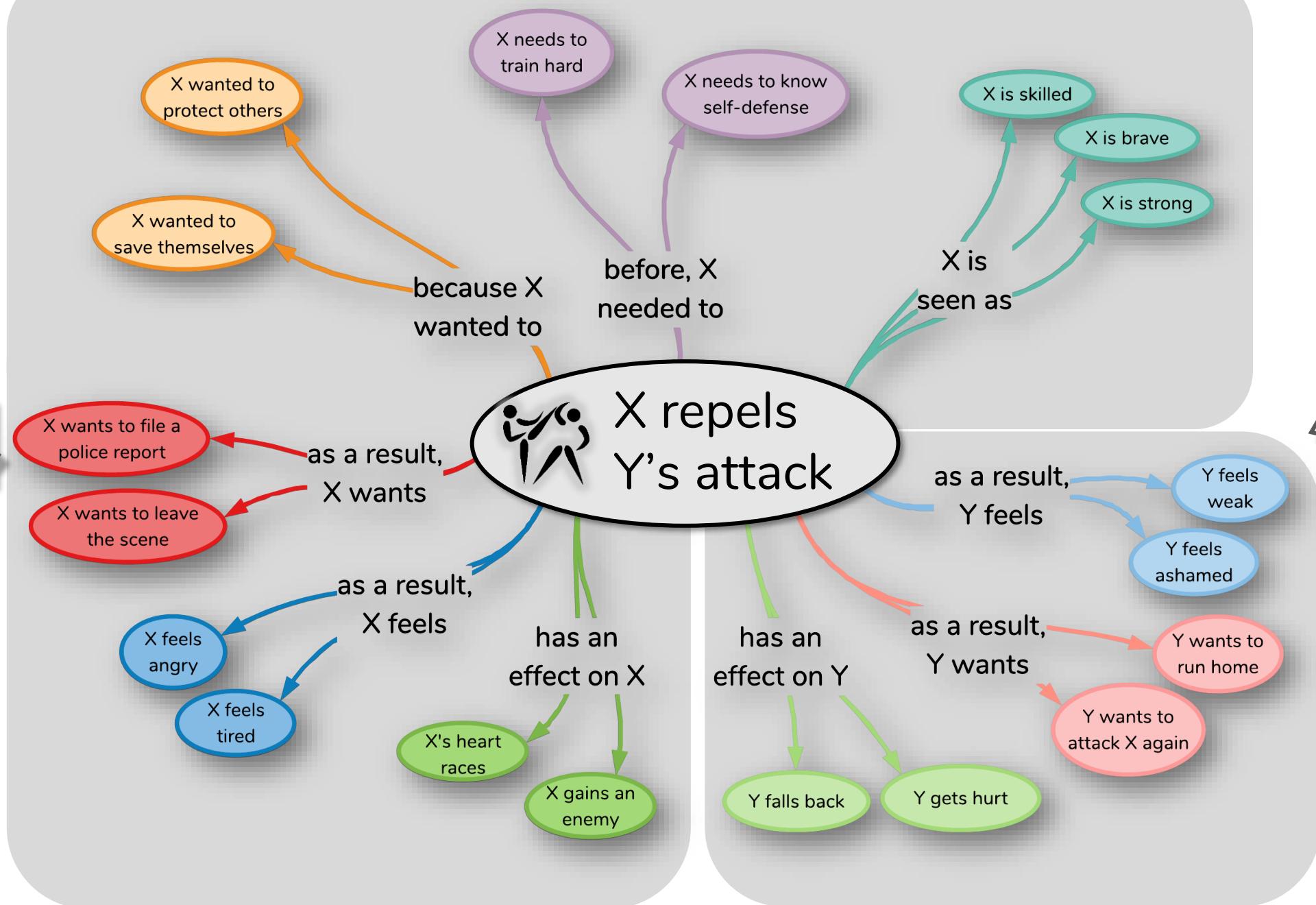
Causes



Effects

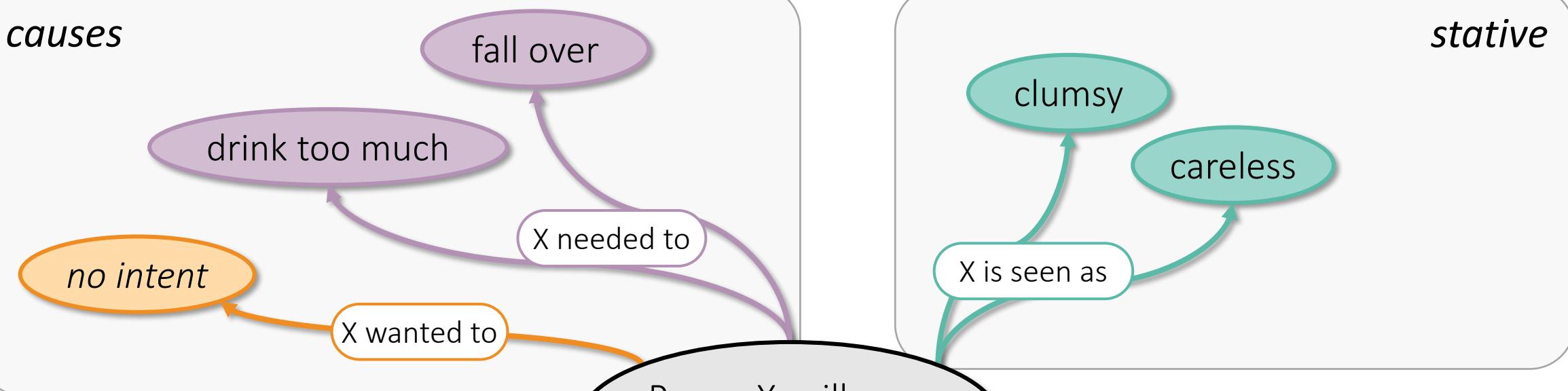


Agent



causes

stative

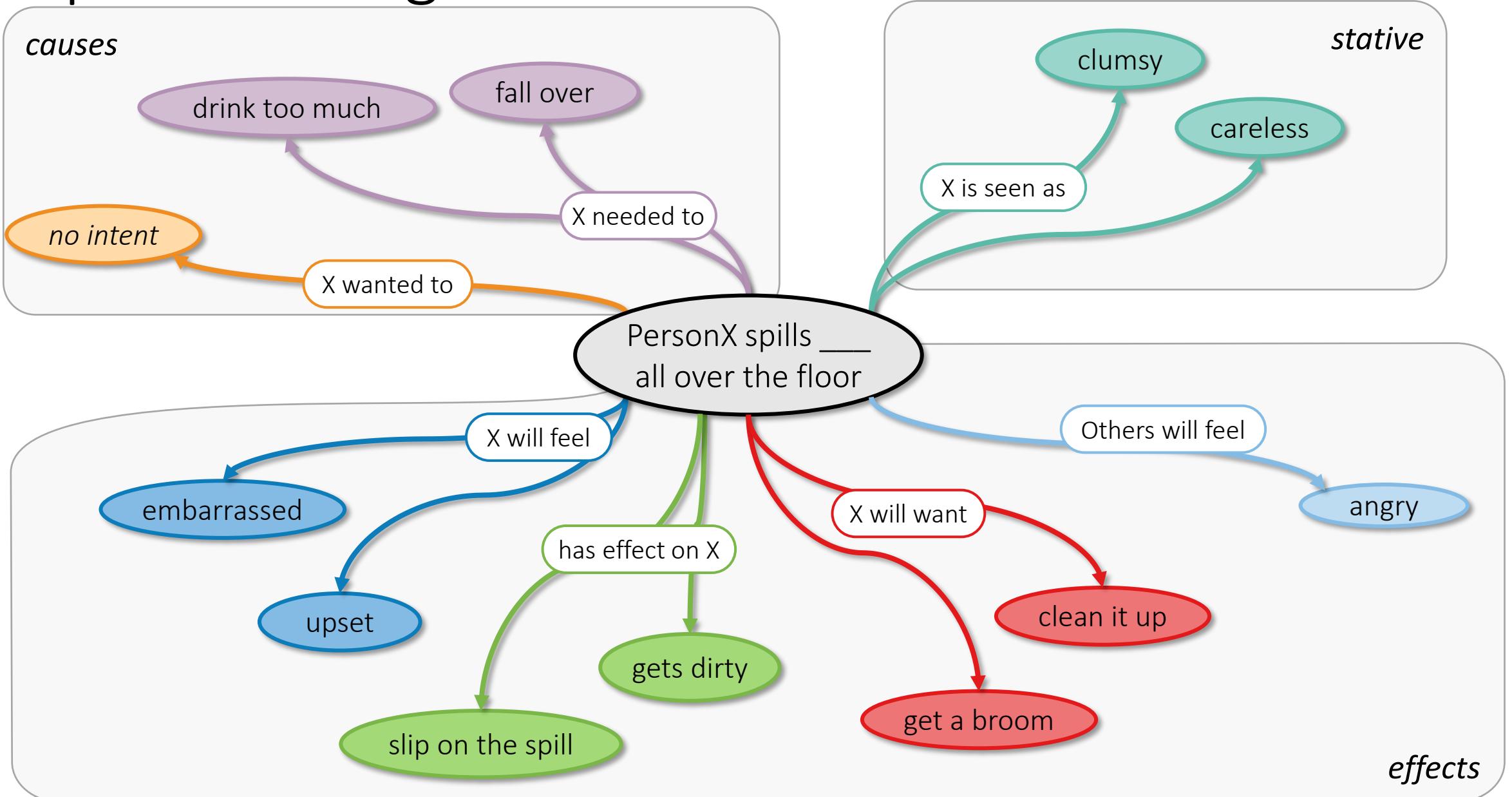


effects

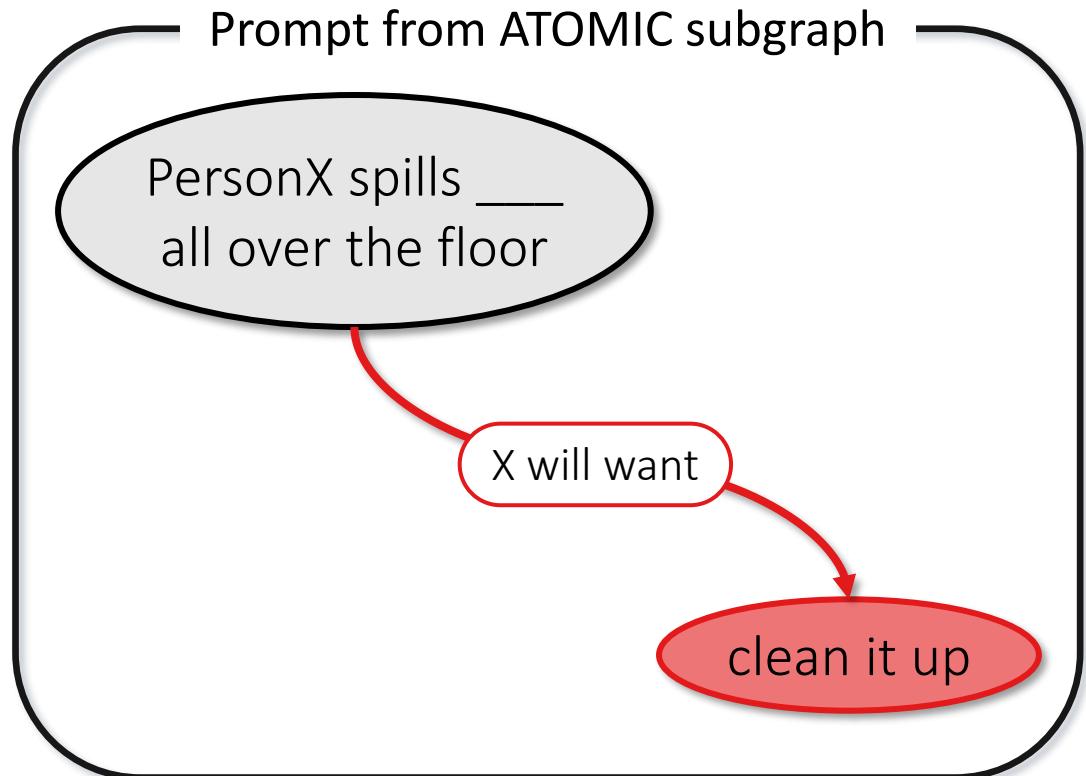
Step 1: Building from ATOMIC

PersonX spills __
all over the floor

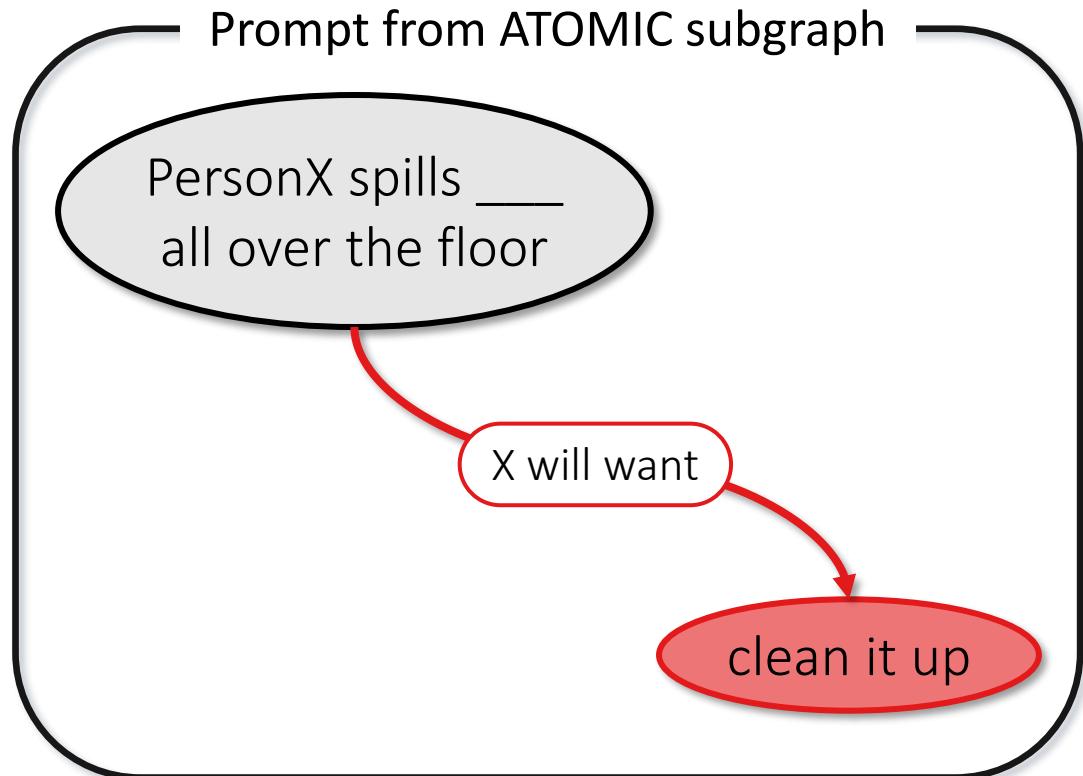
Step 1: Building from ATOMIC



Step 1: Building from ATOMIC



Step 1: Building from ATOMIC

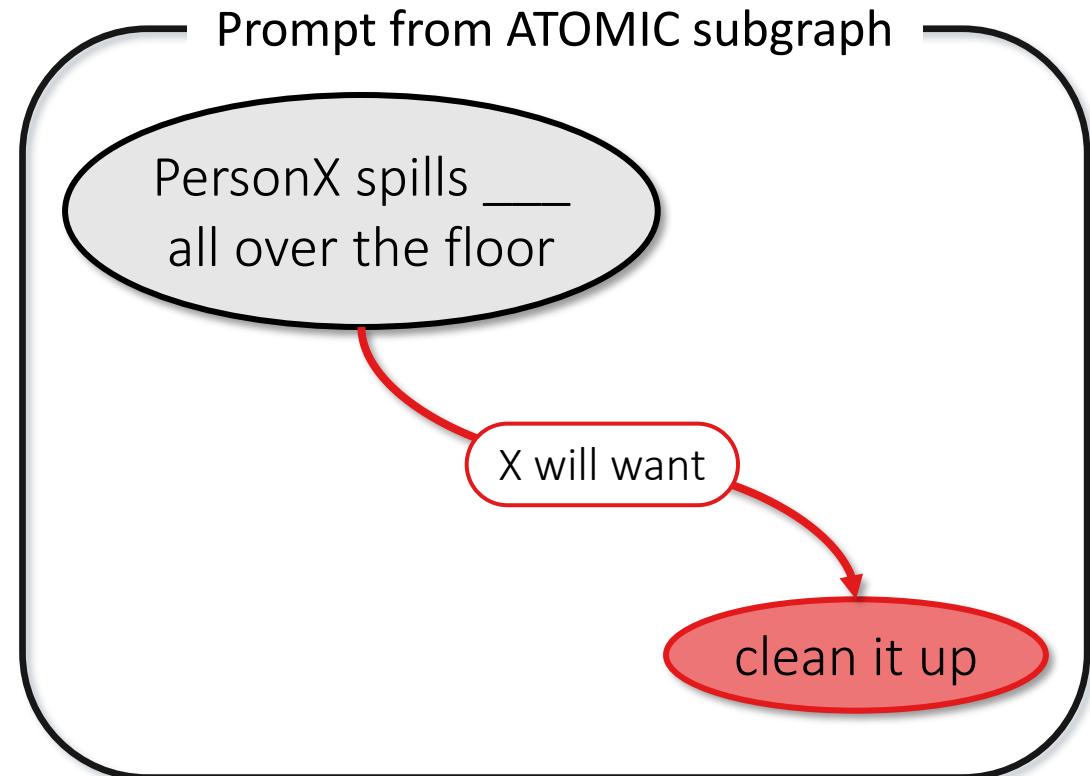


Names & question templates

PersonX xWant

Alex	What will Alex want to do next?
------	---------------------------------

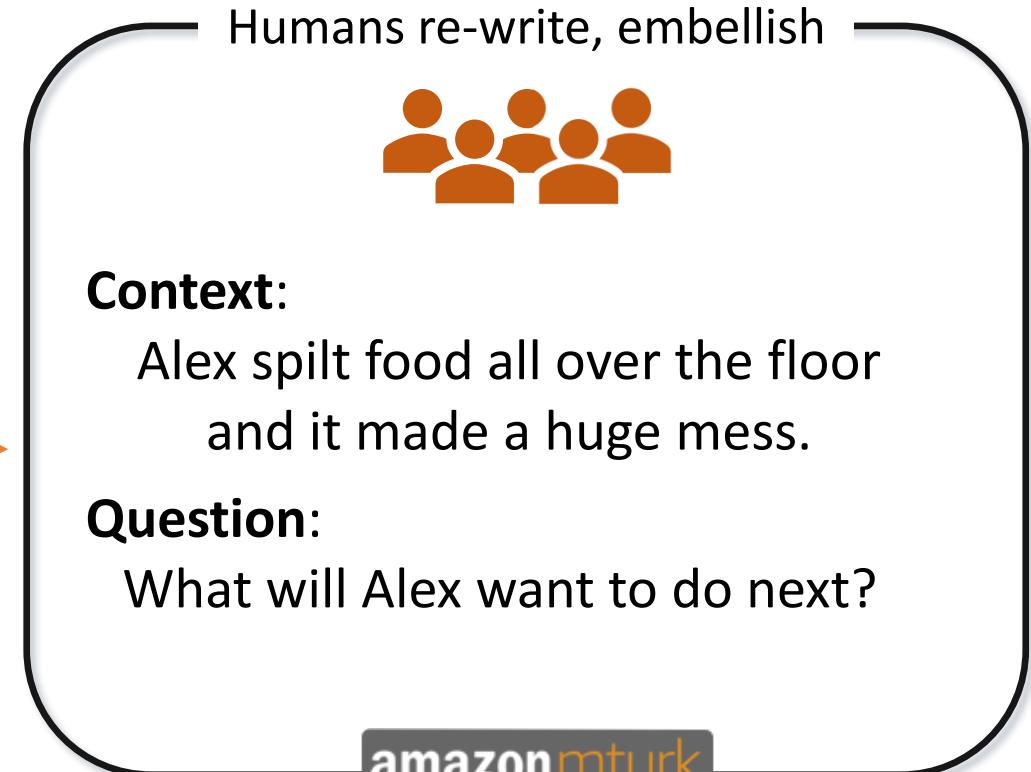
Step 1: Building from ATOMIC



Names & question templates

PersonX	xWant
Alex	What will Alex want to do next?

This diagram shows the names and question templates used in the process. It consists of a table with two columns: "PersonX" and "xWant". The first row contains "PersonX" and "xWant" as column headers. The second row contains the specific values "Alex" and "What will Alex want to do next?". A curved orange arrow originates from the bottom right of the left panel and points to the right panel.



How to collect answers that are plausible and likely / unlikely?

Step 2: Collecting Plausible Answers

Context and Question

Alex spilt food all over the floor
and it made a huge mess.

WHAT HAPPENS NEXT

What will Alex want to
do next?

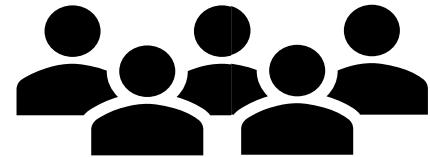
Step 2: Collecting Plausible Answers

Context and Question

Alex spilt food all over the floor
and it made a huge mess.

WHAT HAPPENS NEXT

What will Alex want to
do next?



Handwritten ✓ and ✗ Answers

- ✓ mop up
- ✓ give up and order take out
- ✗ leave the mess
- ✗ run around in the mess

Step 2: Collecting Plausible Answers

Context and Question



amazon mturk
Requester

Our benchmark is ready?

Handwritten ✓ and ✗ Answers

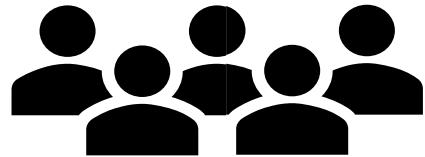
WHAT HAPPENS NEXT

What will Alex want to do next?

- ✓ mop up
- ✓ give up and order take out
- ✗ leave the mess
- ✗ run around in the mess

Step 2: Collecting Plausible Answers

Context and Question



Our benchmark is ready?

Handwritten ✓ and ✗ Answers

No! Problem: handwritten unlikely answers are **too easy to detect**

WHAT HAPPENS NEXT

What will Alex want to do next?

✓ go to the bar and order take out

- ✗ leave the mess
- ✗ run around in the mess

Problem: annotation artifacts

- Models can exploit spurious correlations, annotation artifacts in handwritten incorrect/unlikely answers
 - Exaggerations, off-topic, overly emotional, etc.
- Stem from **cognitive biases** of crowdworkers
[Schwartz et al. '17, Gururangan et al. '18]
- Seemingly “super-human” performance by large pretrained LMs (BERT, GPT, etc.)
 - “*Models solve the dataset not the task*”



Q: How to make unlikely answers **robust to annotation artifacts?**

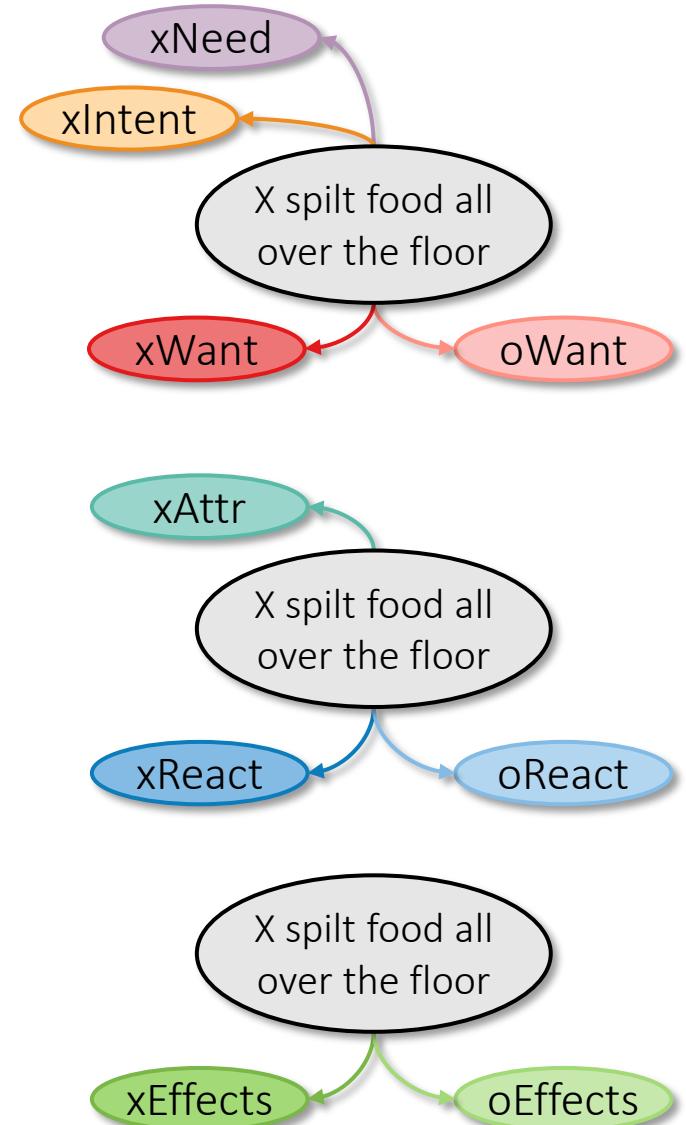
Q: How to make unlikely answers robust to annotation artifacts?

A: Collect the right answers but to a different question

Step 3: Question-Switching setup

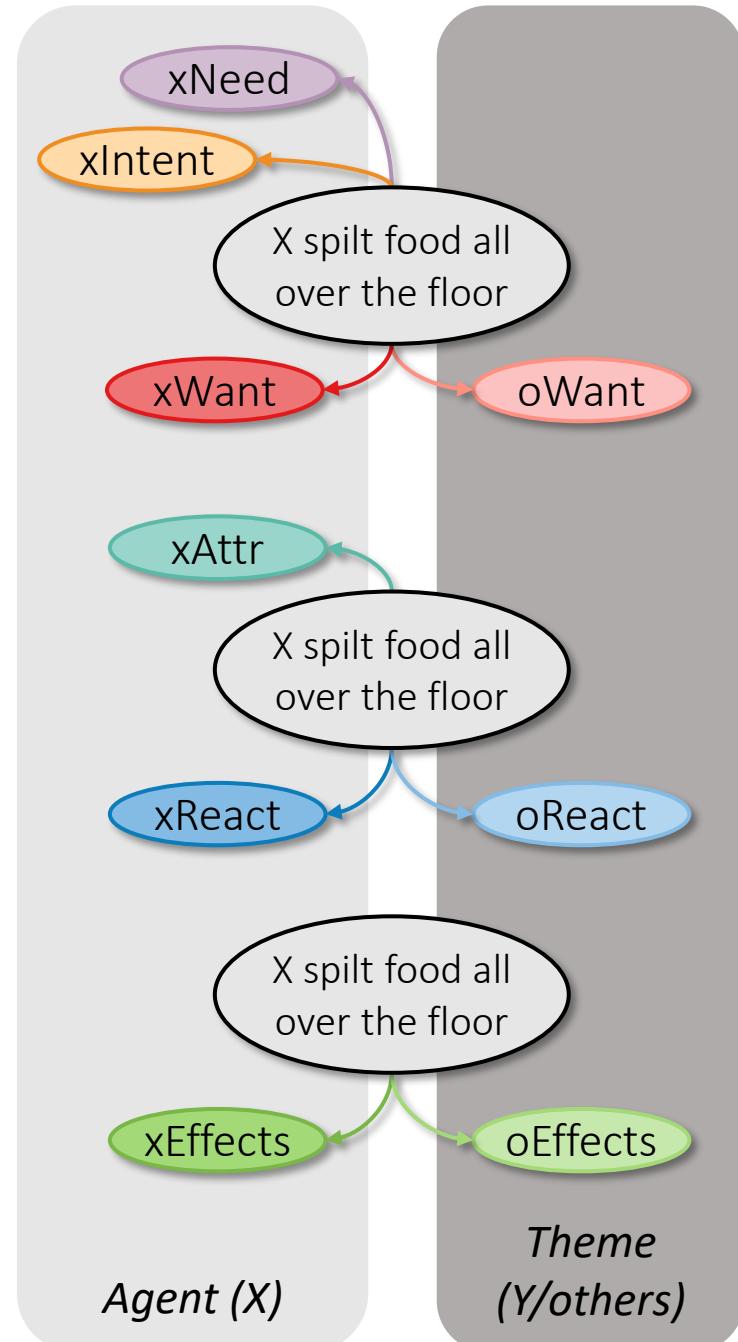
Step 3: Question-Switching setup

- Goal: find questions/answers that...
 - have similar phrasings
 - but are clearly answers to a different question
- Switch out using ATOMIC dimensions
 - Three different clusters of dimensions



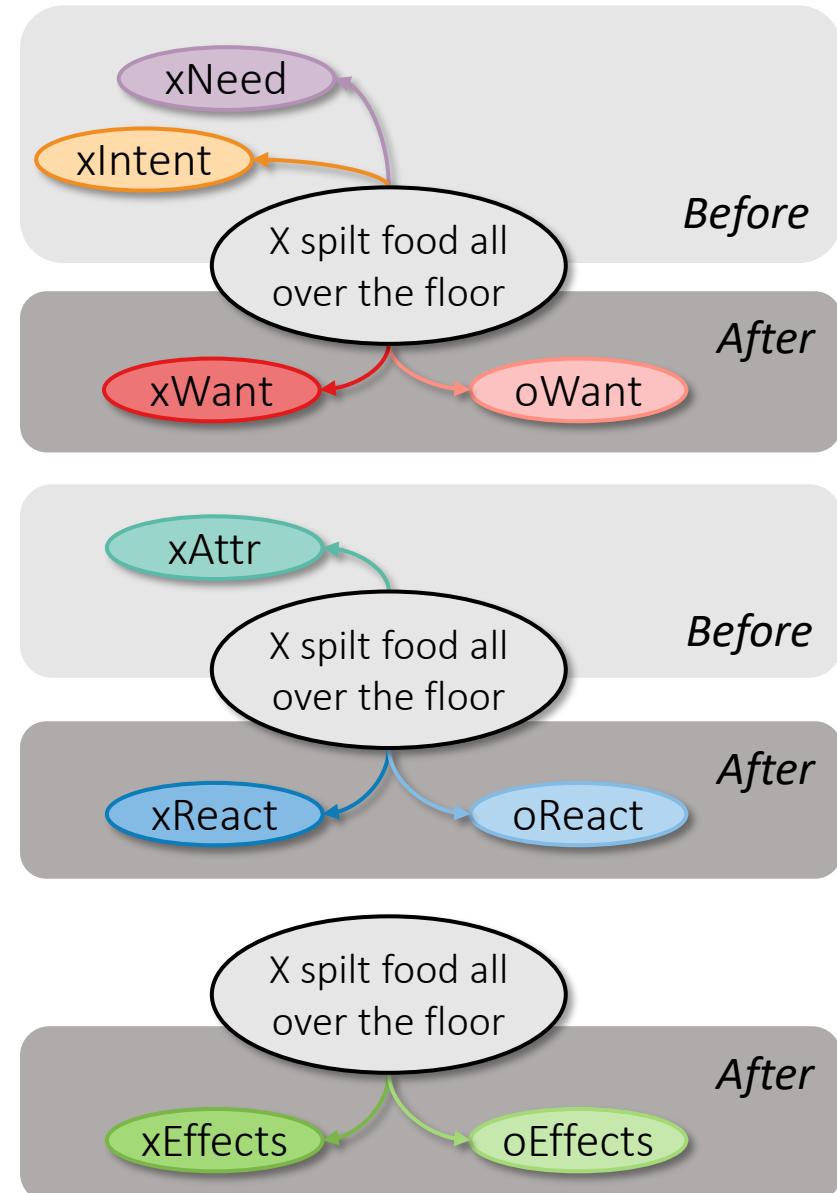
Step 3: Question-Switching setup

- Goal: find questions/answers that...
 - have similar phrasings
 - but are clearly answers to a different question
- Switch out using ATOMIC dimensions
 - Three different clusters of dimensions
- Adversarial question switching
 - Switch who the question is about
(agent vs. theme)



Step 3: Question-Switching setup

- Goal: find questions/answers that
 - have similar phrasings
 - but are clearly answers to a different question
- Switch out using ATOMIC dimensions
 - Three different clusters of dimensions
- Adversarial question switching
 - Switch who the question is about
(agent vs. theme)
 - Switch the temporal ordering of the question
(before vs. after)



Step 3: Question-Switching Answers

Original Question

Alex spilt food all over the floor
and it made a huge mess.

WHAT HAPPENS NEXT

What will Alex want to do
next?

- mop up
- give up and order take out

X

X

Step 3: Question-Switching Answers

Original Question

Alex spilt food all over the floor
and it made a huge mess.

WHAT HAPPENS NEXT

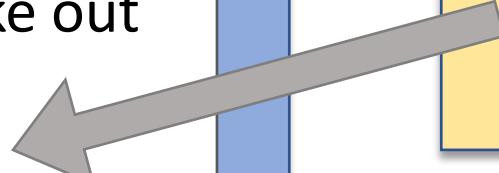
What will Alex want to do
next?

- ✓ mop up
- ✓ give up and order take out
- ✗ have slippery hands
- ✗ get ready to eat

Question-Switching Answer

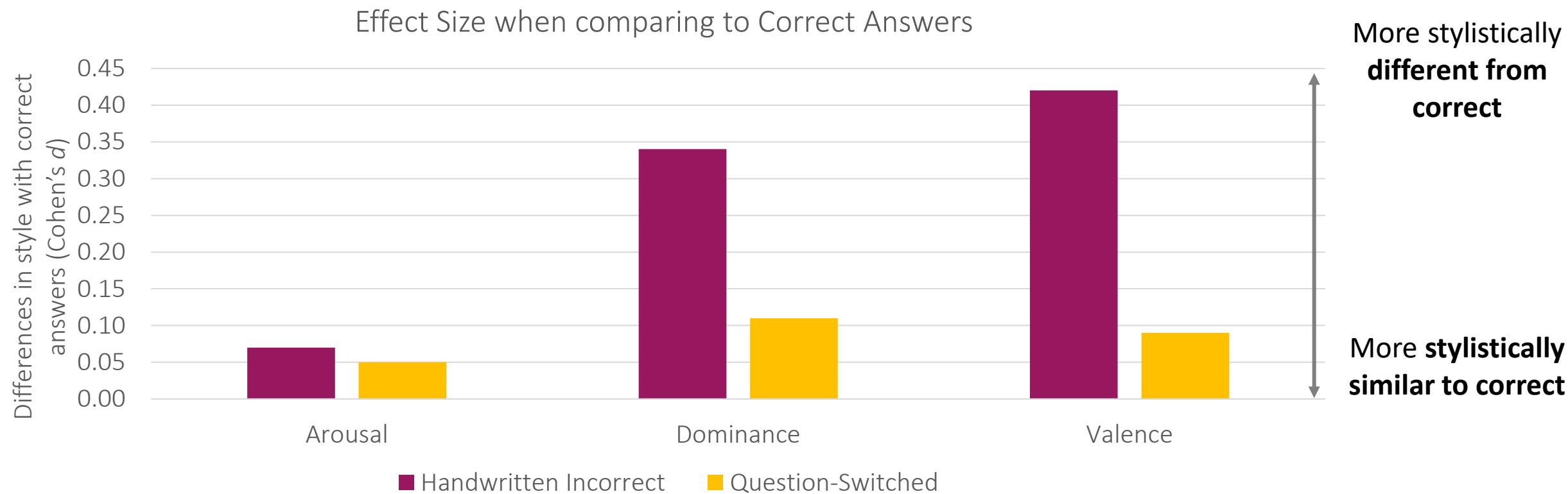
WHAT HAPPENED BEFORE
What did Alex need to do
before this?

- ✓ have slippery hands
- ✓ get ready to eat



Comparing incorrect/correct answers' styles

Using NRC Canada's VAD lexicon [Mohammad et al. '18]



Question switching answers **more stylistically similar** to correct answers

Q: The data is robust to artifacts now?

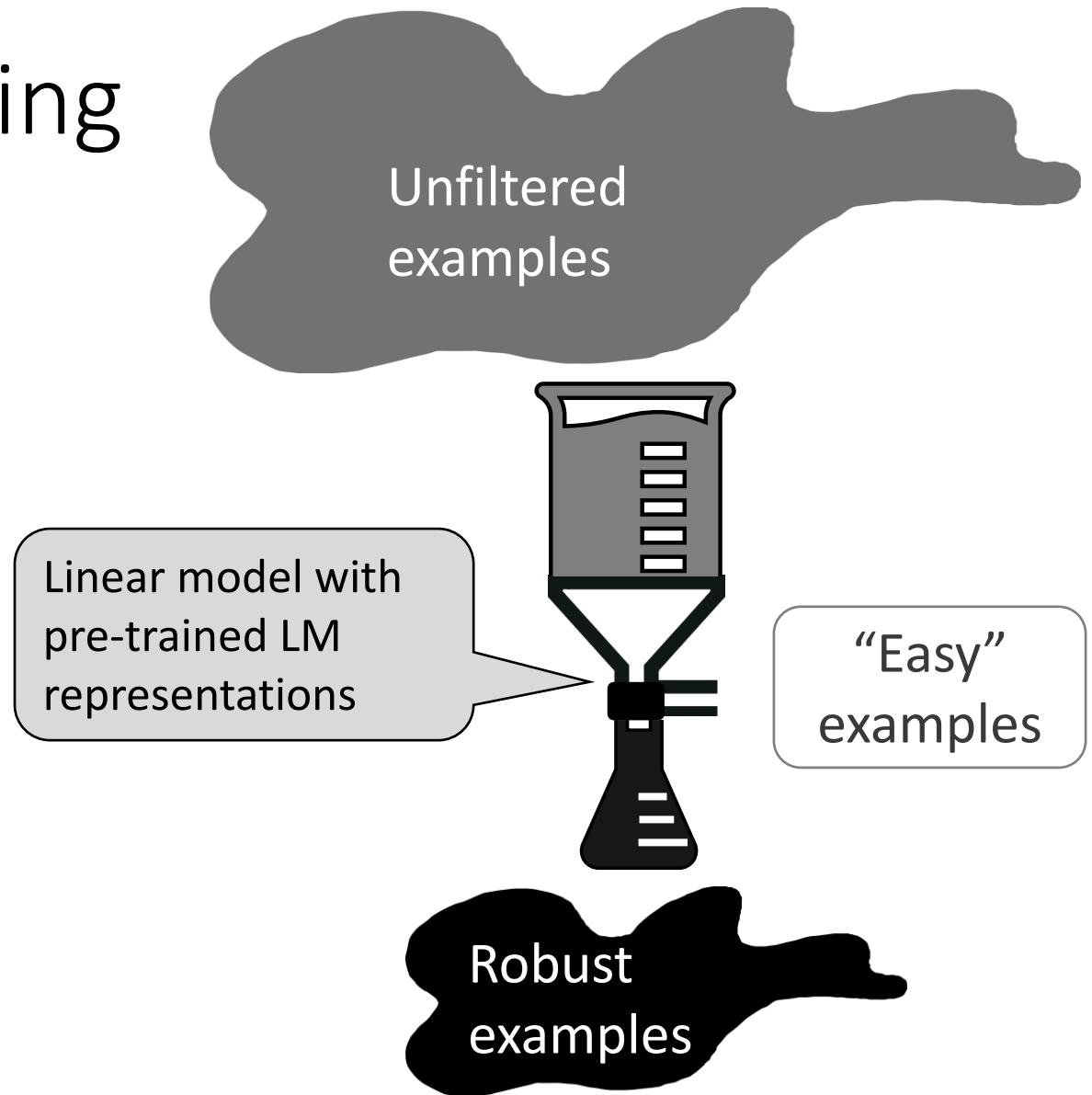
Q: The data is robust to artifacts now?

A: Almost, but not fully!

Step 4: Validation & filtering

Choose examples with **robust** and **diverse answer** options:

- Select **1 likely & 2 unlikely answers** as m/c candidates using NLI entailment scores [Zellers et al. '19]
- Human validate all 3-way m/c QA tuples using crowdsourcing 
- **AF-Lite**: lightweight adversarial filtering of spurious correlations [Sakaguchi et al., 2019]



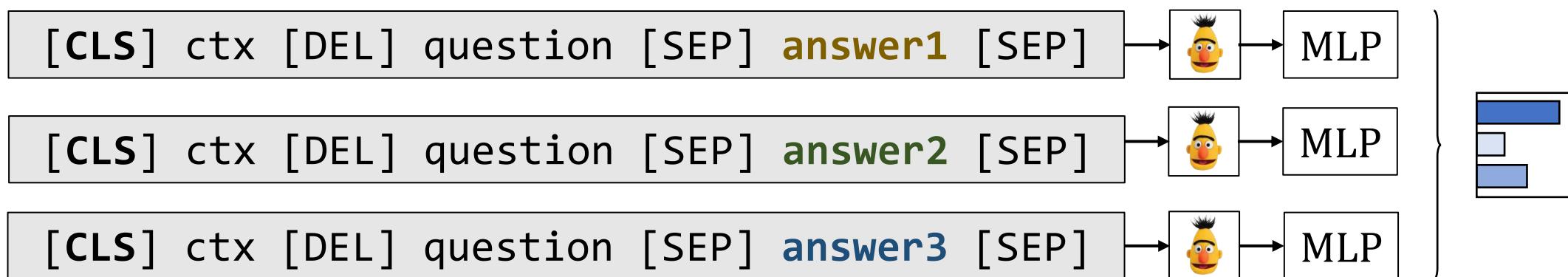
How do computational models hold up against SOCIAL IQA

Experimental Set-up

Formulate as M/C questions
with 3 answer options
Over 38k total questions

	# m/c questions
train	33,410
dev	1,954
test	2,224

Fine-tune large pretrained LM (BERT, OpenAI-GPT, etc.)



Finetuned model performance

Humans



Bert-large



Bert-base



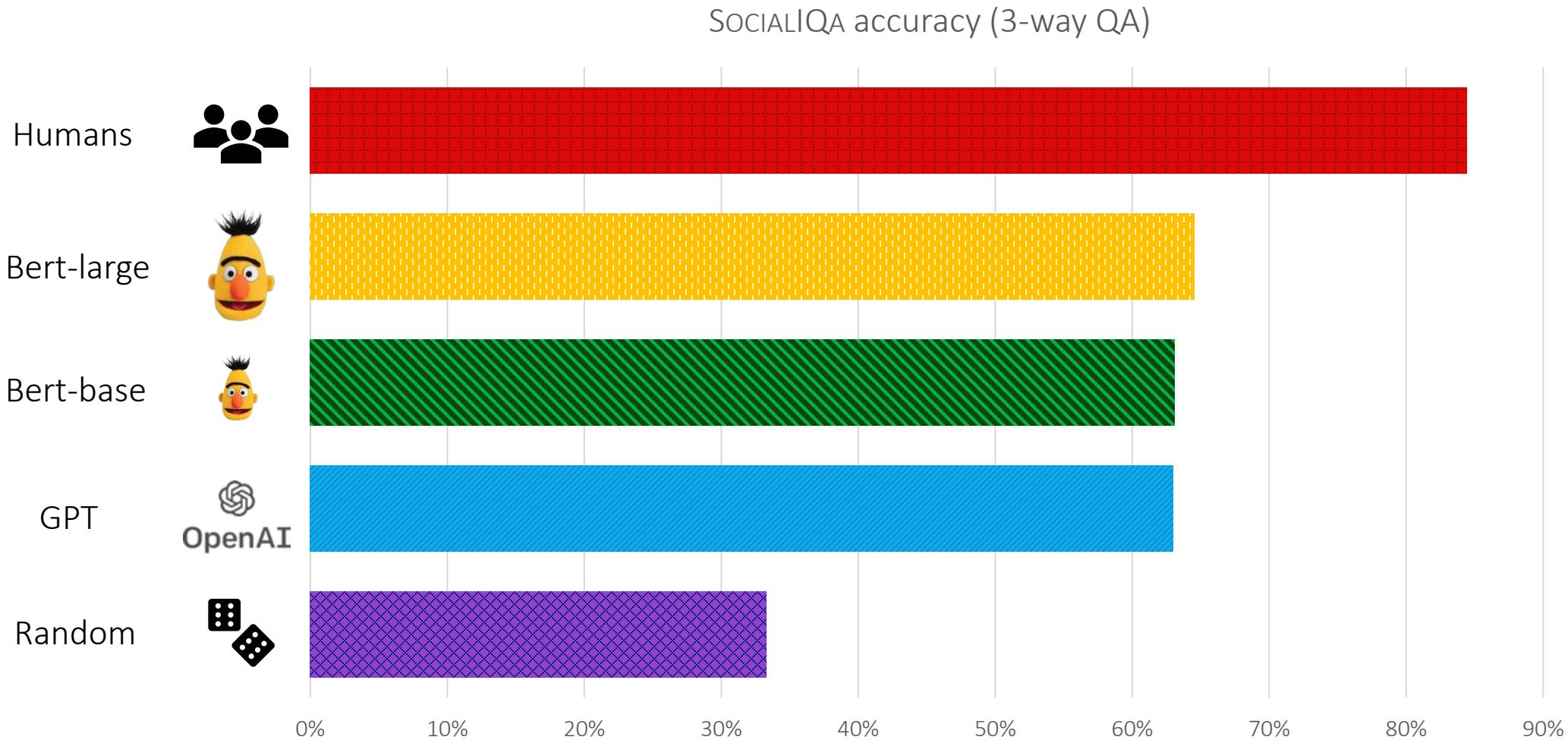
GPT



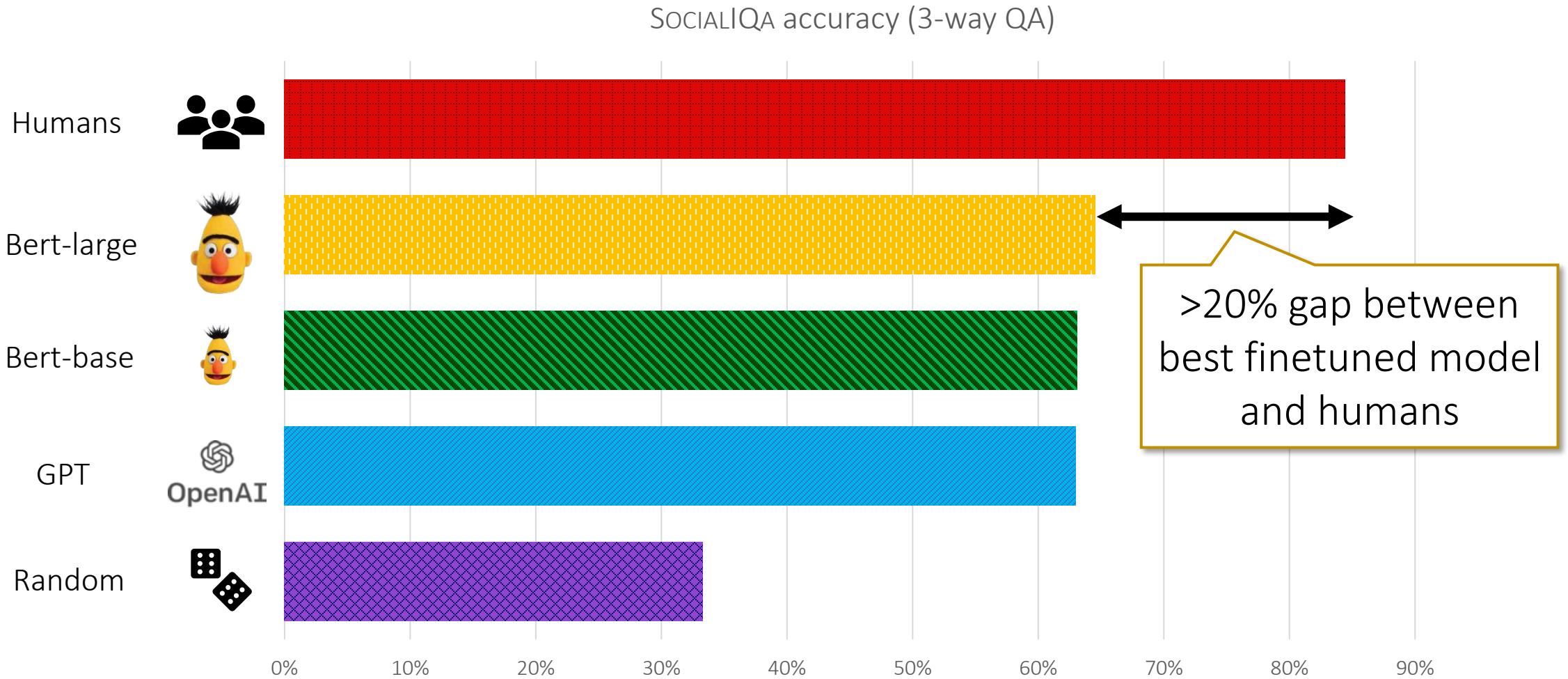
Random



Finetuned model performance

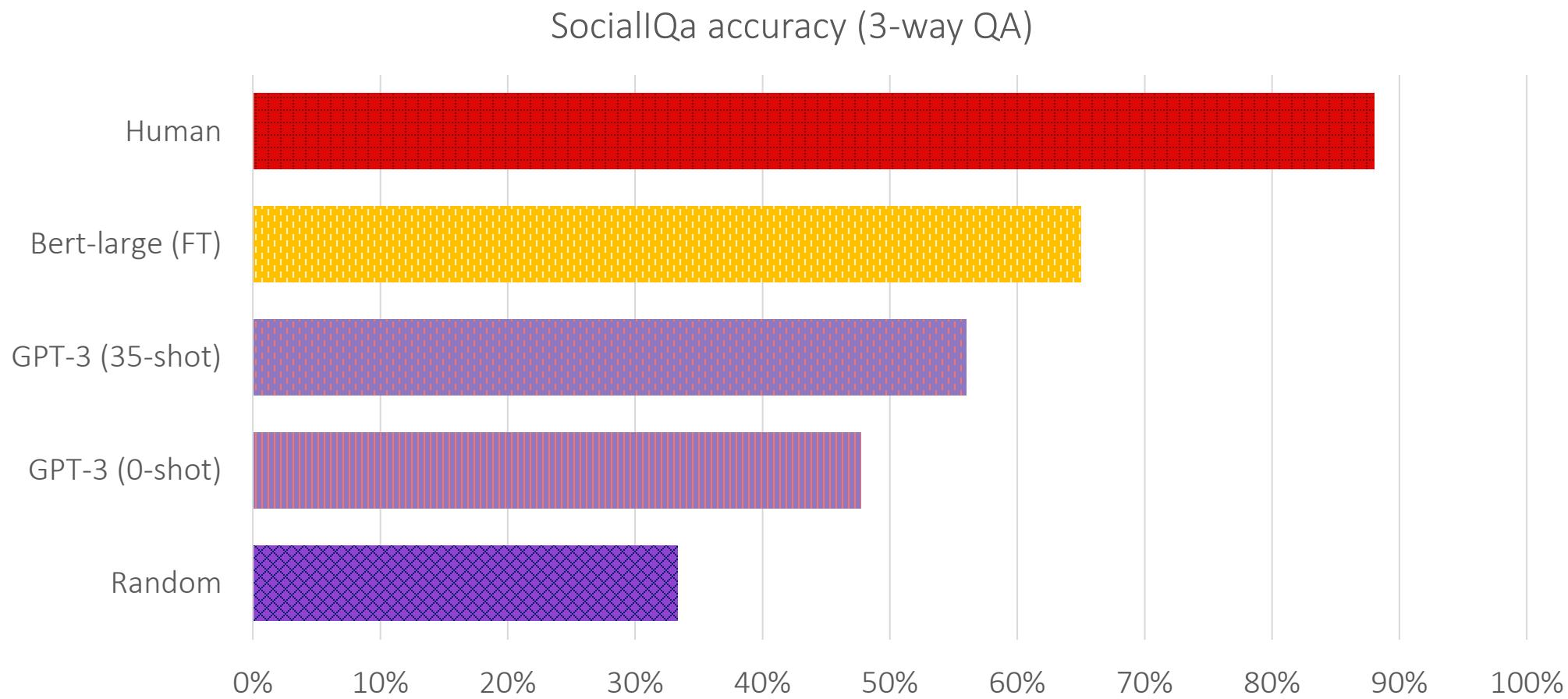


Finetuned model performance



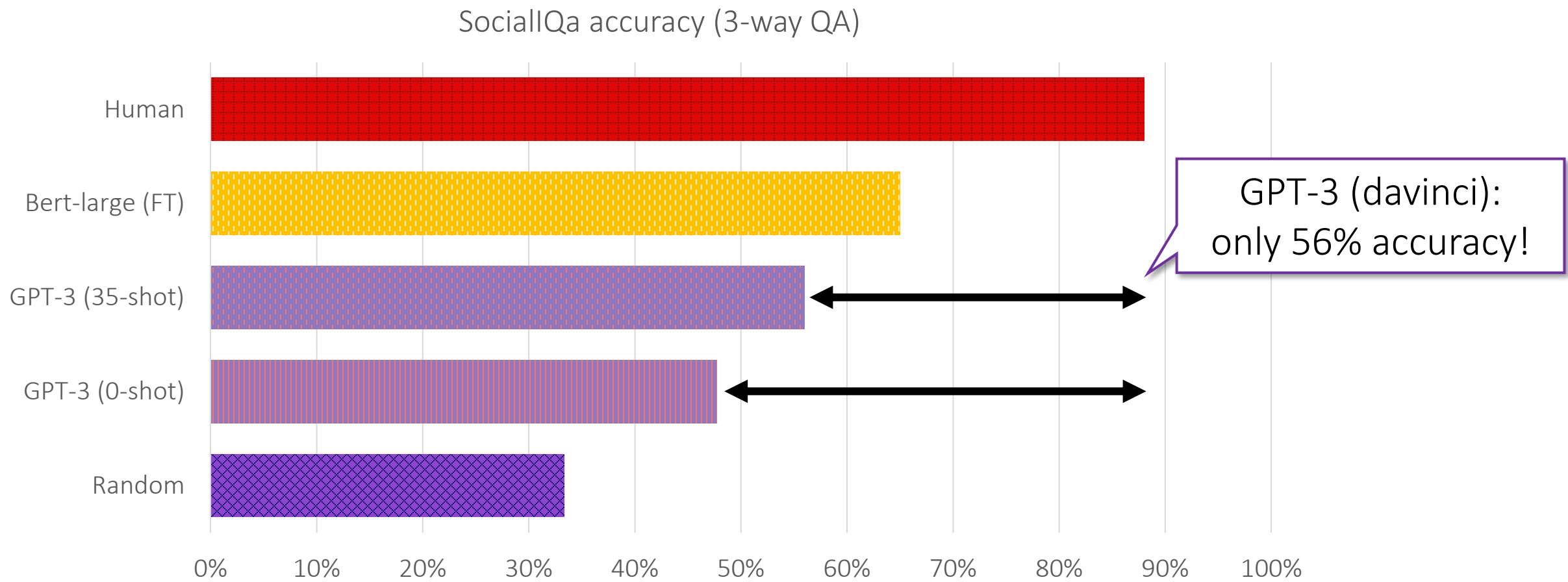
GPT-3 results – new 2021 results

Using LM-probing setup in **zero- and few-shot settings** as in GPT-3 paper [Brown et al. '20]



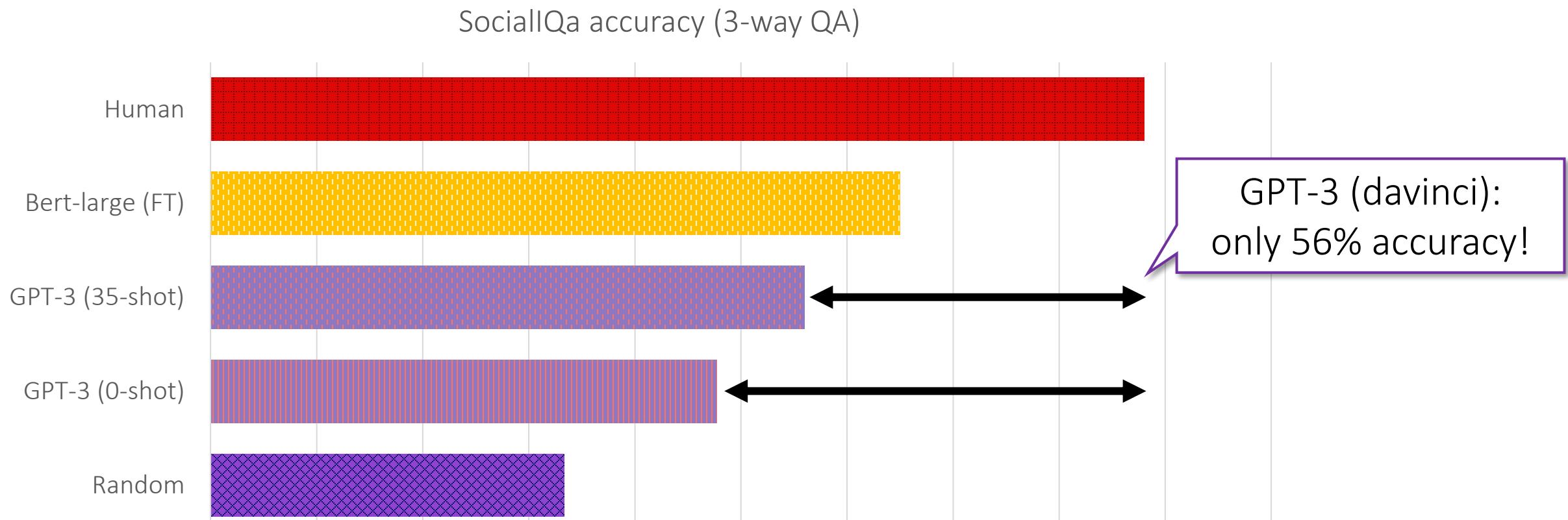
GPT-3 results – new 2021 results

Using LM-probing setup in **zero- and few-shot settings** as in GPT-3 paper [Brown et al. '20]



GPT-3 results – new 2021 results

Using LM-probing setup in **zero- and few-shot settings** as in GPT-3 paper [Brown et al. '20]



SOCIAL IQA is challenging in both finetuned and probing setups

Why kinds of mistakes do models make on Social IQa

Challenging examples for finetuned BERT-large



Although Aubrey was older and stronger,
they lost to Alex in arm wrestling.

How would Alex feel as a result?



ashamed

how **Aubrey** would
feel, not Alex



boastful

they need to practice more

Challenging examples for finetuned BERT-large



Although Aubrey was older and stronger, they lost to Alex in arm wrestling.

How would Alex feel as a result?



ashamed

how **Aubrey** would
feel, not Alex



boastful

they need to practice more

Remy gave Skylar, the concierge, her account so that she could check into the hotel.



What will Remy want to do next?

lose her credit card



arrive at a hotel

what Remy
did **before**



get the key from Skylar

Challenging examples for finetuned BERT-large



Although Aubrey was older and stronger, they lost to Alex in arm wrestling.

How would Alex feel as a result?



ashamed



boastful

they need to practice more

how **Aubrey** would feel, not Alex

Remy gave Skylar, the concierge, her account so that she could check into the hotel.



What will Remy want to do next?

lose her credit card



arrive at a hotel

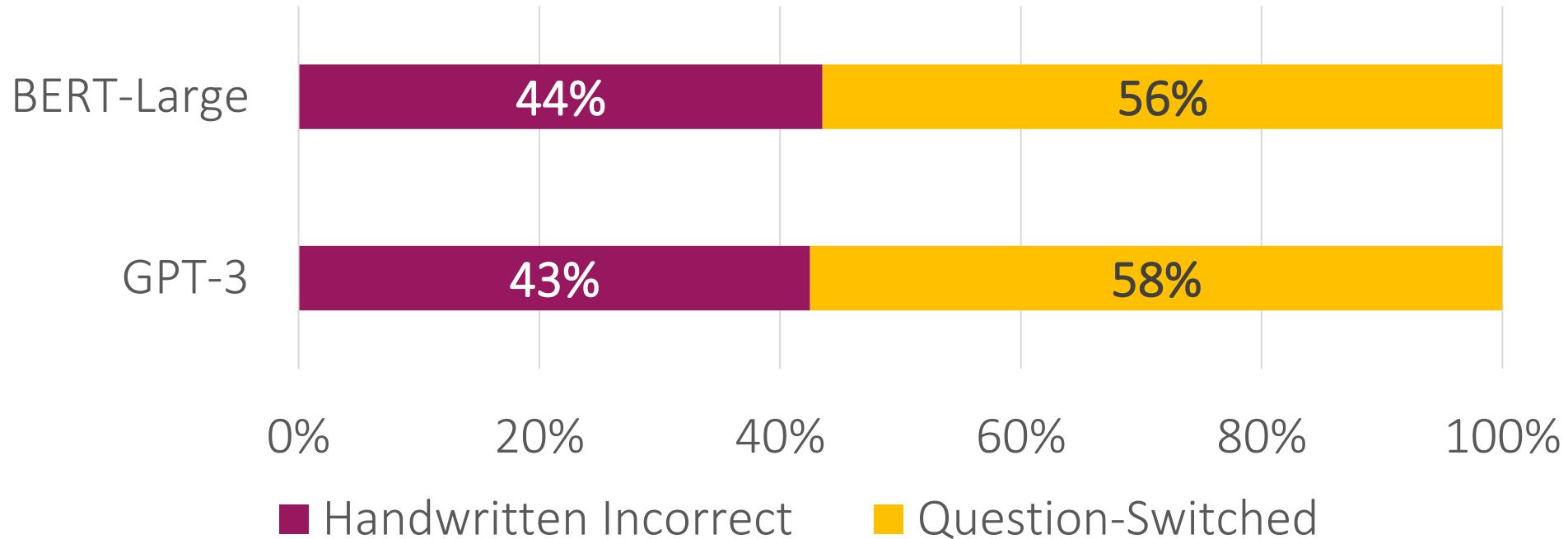
what Remy did **before**



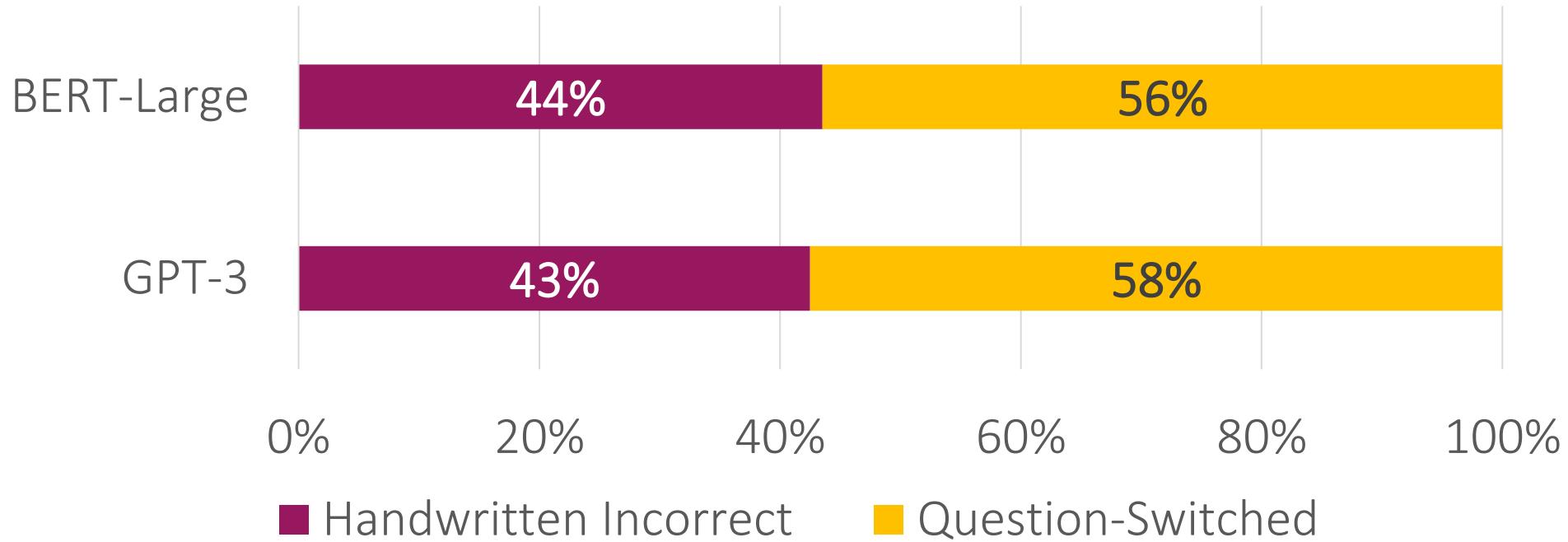
get the key from Skylar

- Need better **person-centric reasoning**
- Better distinguishing of **causes vs. effects**
- Mistakes seem to align with our **question switching...**

Rates of HIA vs. QSA mistakes



Rates of HIA vs. QSA mistakes



Question-switched answers are often better distractors for models

Can BERT be taught social commonsense knowledge?

SOCIAL IQA for transfer learning

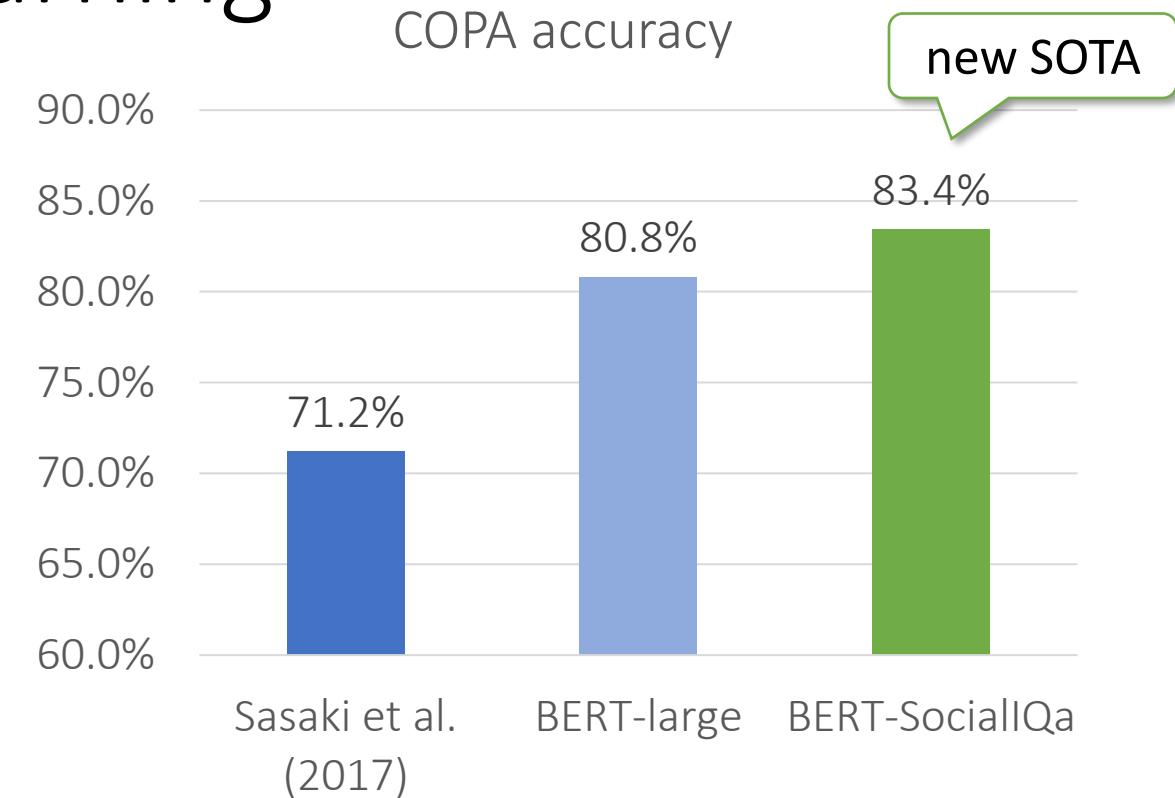
Sequential finetuning:

similar to Phang et al. '18, Talmor & Berant '19



End tasks:

- Choice of plausible alternatives (**COPA**)
- Winograd Schema Challenge (**WSC**)



- Similar improvements on Winograd Schema Challenge
- SOCIALIQA endows BERT with some social reasoning skills

Takeaways

- Introduced SOCIALIQA, the first large scale benchmark for **social commonsense reasoning**
- Collected using a framework that **minimizes annotation artifacts** using question-switching
- Remains **challenging for computational models**
 - Even for GPT-3!
- Show usefulness as a **resource for transfer learning**, on COPA and WSC

Co-authors



maartensap.github.io/social-iqa/



leaderboard.allenai.org/socialiqa/

EMNLP 2021 Tutorial

Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection

Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar,
Sam Bowman, and Yoav Artzi

Summary

Presented by Yoav Artzi

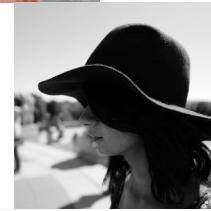


Summary

- Discussed five very different case studies
- Slides are available at our website:
<https://nlp-crowdsourcing.github.io/>
- During EMNLP, we are looking forward to meeting you at our clinic!



Alane



Nikita



Clara



Mark



Sam



Yoav



Maarten