WANLI

- Collaboration for NLI Dataset Creation

Alisa Liu

Swabha Swayamdipta

Noah Smith

Yejin Choi

Demo

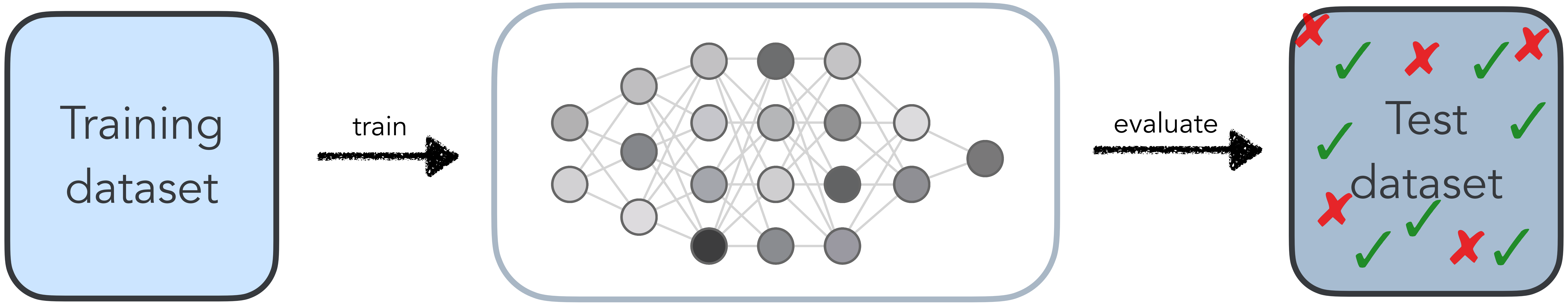
https://wanli.apps.allenai.org/

Datasets in natural language processing

Datasets are the backbone of machine learning

good training sets teach
our model the task

good test sets
evaluate progress



*How can we distill **human language understanding** into datasets that models can **learn from** and be **evaluated** on?*

Limitations of crowdsourcing

Idea: ask people to write down examples of what they know

Premise

A cat sat on the mat.

entails

contradicts

$\Rightarrow \; \Leftarrow$

is neutral w.r.t.

Hypotheses

An animal sat on the mat.

No one sat on the mat.

A fluffy cat sat on the mat.

some
instrument
animal

Replace specific words with general ones

something
outdoors
person

no
never

Just negate it!

nothing
sleeping
naked
nobody

sad

Add a plausible adjective

popular
tall

(Geva et al., 2021; Gururangan et al., 2018)

train

Models overfit to these patterns and don't produce the ***right answer for the right reasons***

Idea: our linguistic knowledge is largely subconscious

Humans are not good at painting a complete picture of what they can do under the task

But we are good at evaluating what's right and what's wrong!

*We want to use humans to **revise + evaluate** examples… but where can we get decent examples to start with?*

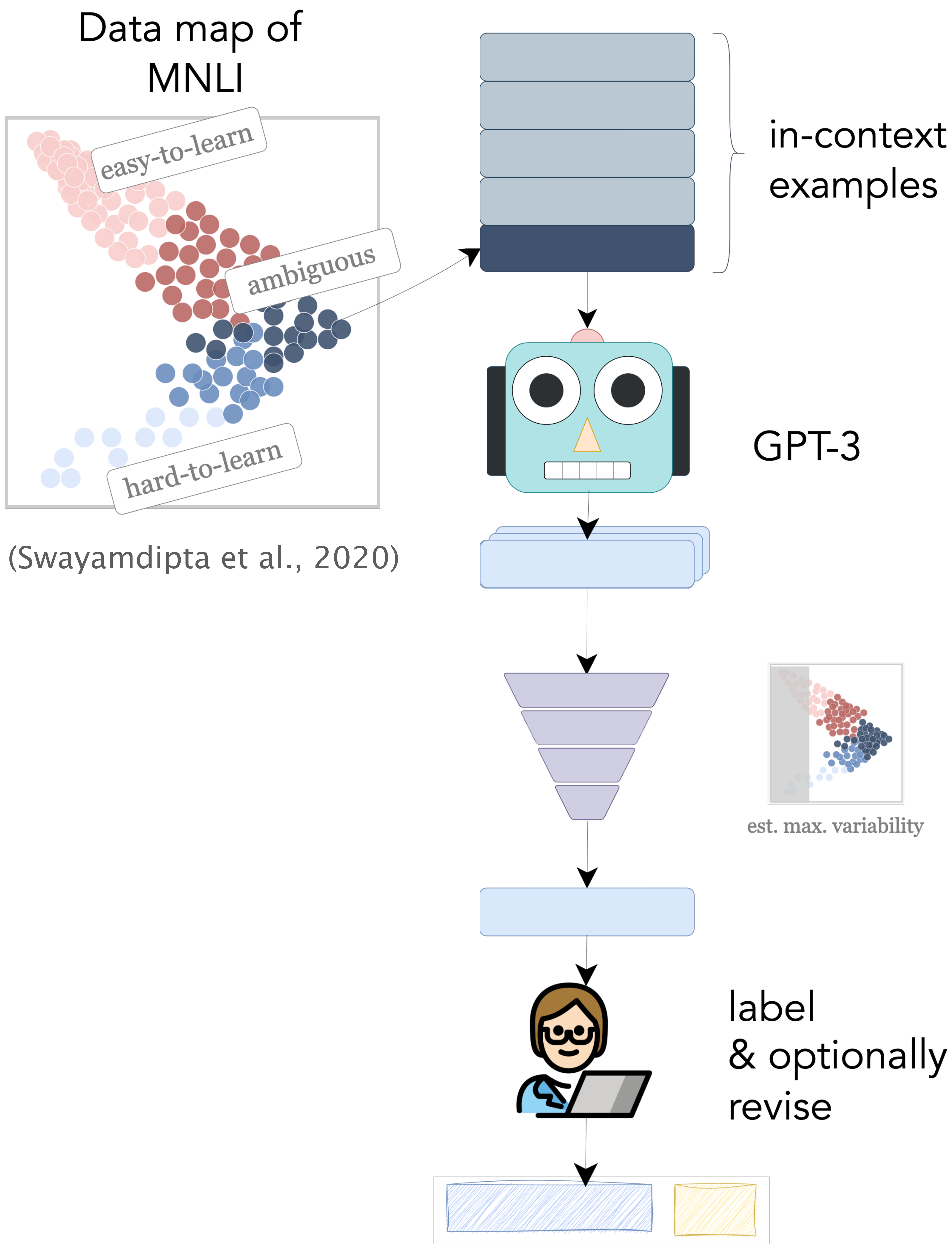
This is where AI comes in!

Worker-AI Collaboration

Leverage the **generative strength** of LMs and **evaluative strength** of humans

LMs create new examples by replicating valuable reasoning patterns in an existing dataset

Humans revise and assign a label



1. ***Exemplar collection***: *automatically collect* ***groups*** *of examples that share a* ***challenging reasoning pattern***

2. ***Overgeneration***: *prompt GPT-3 to create* ***novel examples*** *with the* ***same reasoning pattern***

3. ***Filtering***: *filter with new metric based on* ***Data Maps***

4. ***Human annotation***: *humans optionally* ***revise*** *for clarity and fluency, and assign a* ***gold label***

Write a pair of sentences that have the same relationship as the previous examples. Examples:

1. ***In six states***, the federal investment represents almost the entire contribution for providing civil legal services to low-income individuals.
Implication: In ***44 states***, the federal investment does not represent the entire contribution for providing civil legal services for people of low income levels.

2. But if it's at all possible, plan your visit for the ***spring, autumn, or even the winter***, when the big sightseeing destinations are far less crowded.
Implication: This destination is most crowded in the ***summer***.

3. ***5 percent*** of the routes operating at a loss.
Implication: ***95 percent*** of routes are operating at either profit or break-even.

4. About ***10 percent of households*** did not
Implication: Roughly ***ninety percent of households*** did this thing.

5. ***5 percent probability*** that each part will be defect free.
Implication: Each part has a ***95 percent chance*** of having a defect.

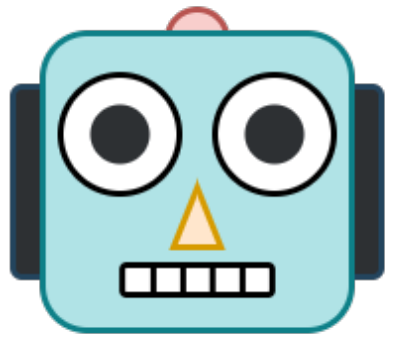
6.

* formatting added for clarity

entailment pattern: reasoning about set complements

nearest neighbors under task model

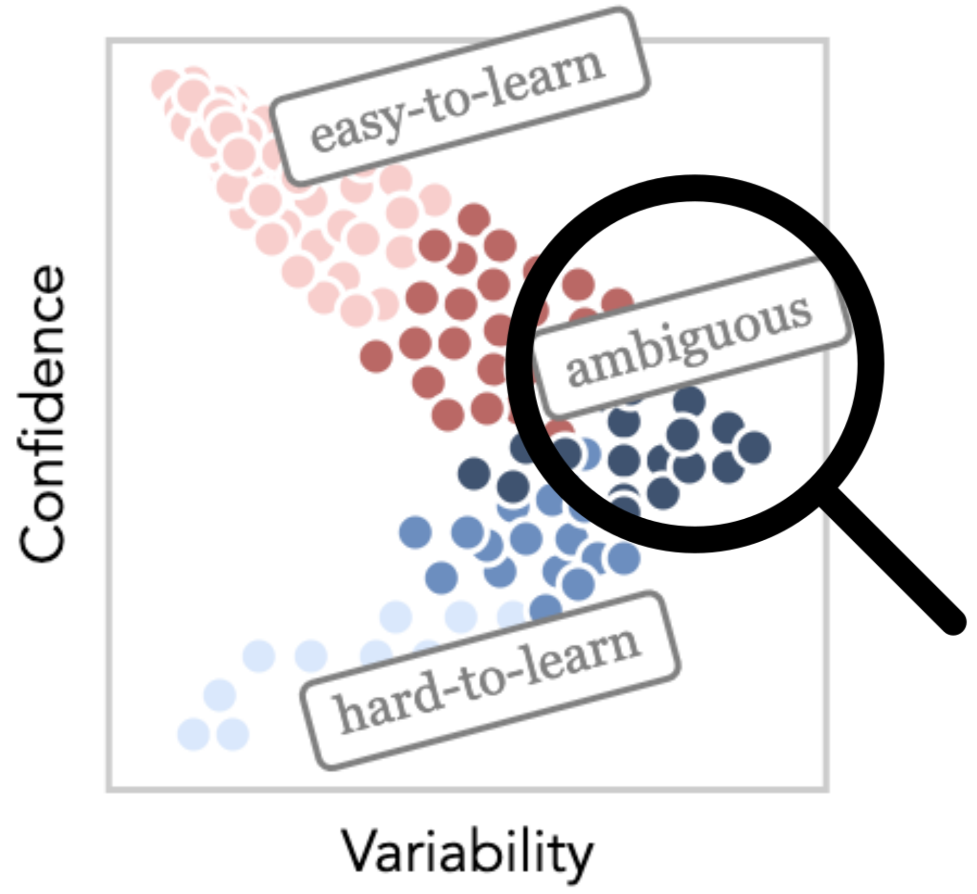
seed example



vanilla GPT-3 curie

1 percent of the seats were vacant.

Implication: ***99 percent*** of the seats were occupied.



Dataset Cartography (Swayamdipta et al., 2020)

Revision

1) Improve the fluency of the text

P: He had no idea that he was the only one in the room.

H: He was the only one in the room, ~~he was the only one in the room~~.

Entailment

P: There is a slight possibility that, if the same temperature data are used, the temperature of the Earth's surface in 1998 will be lower than the temperature of the Earth's surface ~~in 1998~~ now.

H: The Earth's surface in 1998 was lower than the Earth's surface in ~~in 1998~~ now.

Neutral

2) Improve the clarity of the relationship

P: As I climbed the mountain, I noticed that the clouds were parting, and the sun was shining through.

H: The sun ~~is~~ was shining through the clouds.

Entailment

P: This will be the first time the king has met the queen in person.

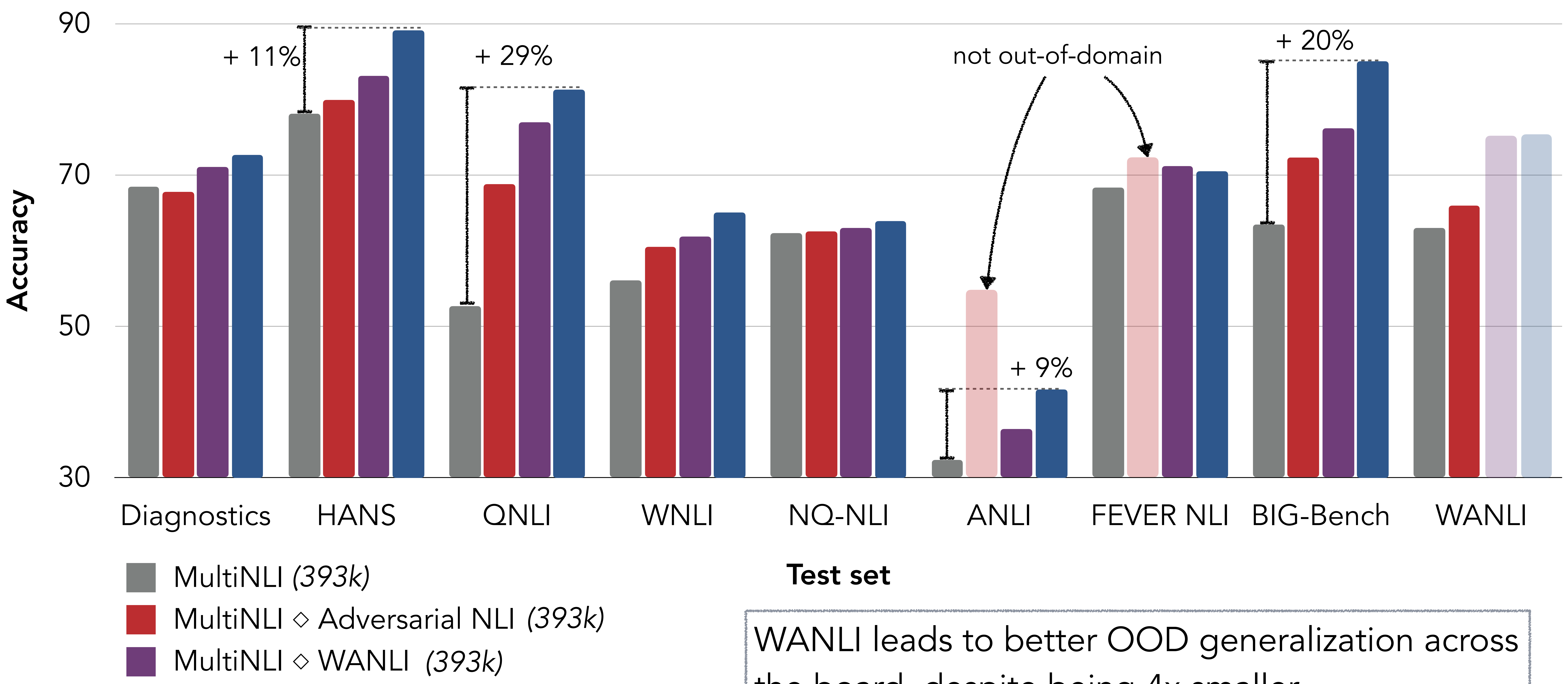
H: The king has met the queen in person before.

Contradiction

Worker and AI NLI (万理)

Split	Size	Label distribution (E/N/C)
Train	103,079	39K / 49K / 15K
Test	5,000	1.8K / 2.4K / 745

Does training on WANLI improve model robustness?



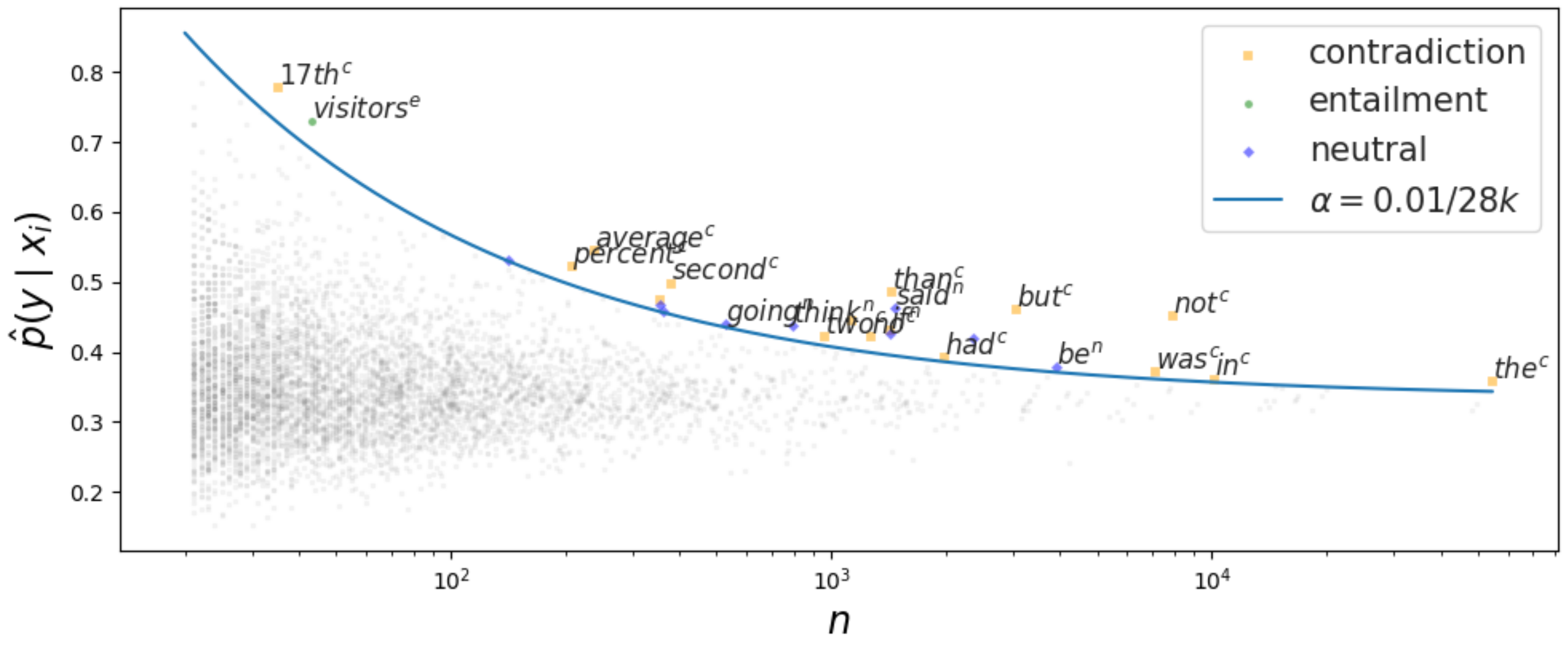
WANLI leads to better OOD generalization across the board, despite being 4x smaller

WANLI contains fewer known artifacts

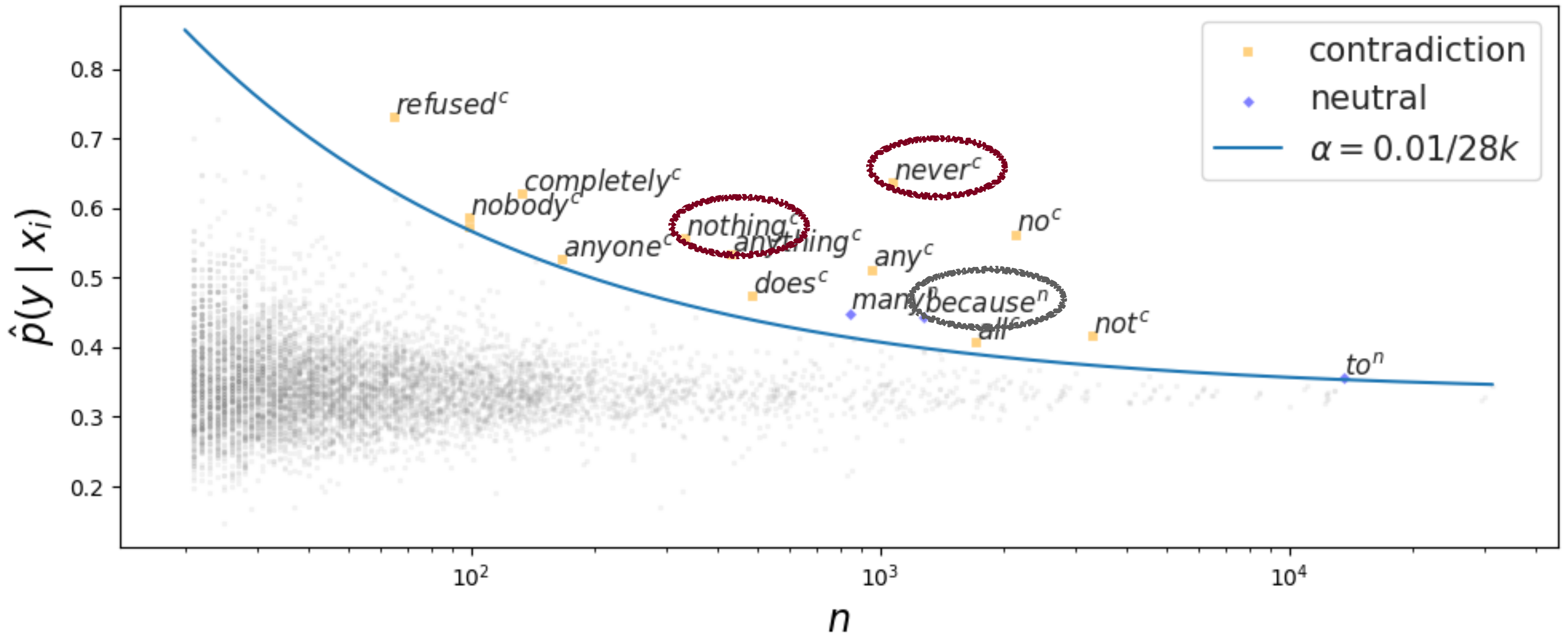
Compared to MultiNLI, WANLI has

- less information about the label contained in the hypothesis alone (Gururangan et al., 2018)
- fewer previously known lexical correlations (Gardner et al., 2021)

WANLI



MNLI



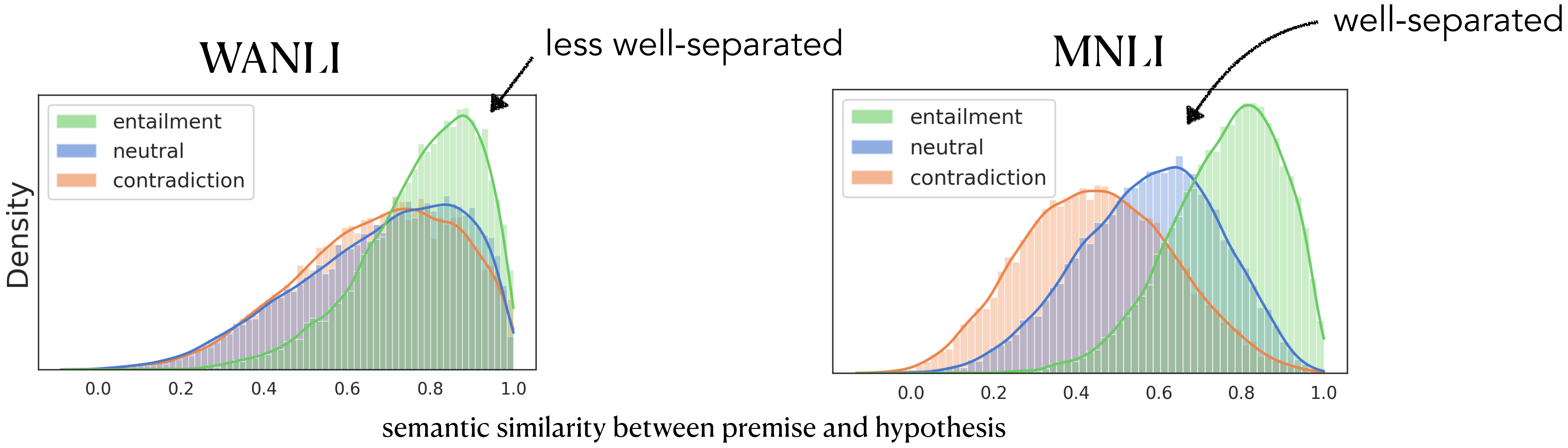
WANLI contains fewer known artifacts

Compared to MultiNLI, WANLI has

less information about the label contained in the hypothesis alone (Gururangan et al., 2018)

fewer previously known lexical correlations (Gardner et al., 2021)

less information about the label contained in the semantic similarity between the premise and hypothesis



Takeaways

Human-AI collaborative creation of NLP datasets!

Applied it to create a new dataset for NLI, which we showed leads to more robust models while avoiding known issues in existing NLI datasets

*How can we distill **human language understanding** into datasets that models can **learn from** and be **evaluated** on?*

This work: ask workers to **revise** and **evaluate content**, rather than write free-form examples

Demo

https://wanli.apps.allenai.org/