(万理) WANLI: Worker and AI Collaboration for NLI Dataset Creation

Alisa Liu[1] Swabha Swayamdipta[2]
Noah A. Smith[1,2] Yejin Choi[1,2]
[1] University of Washington [2] Allen AI

wanli.apps.allenai.org

@alisawuffles

What should be the role of humans in data creation?

Crowdworkers often rely on simple writing strategies when creating examples from scratch, leading to datasets flooded with **repetitive & spurious** examples, and therefore **brittle models**.

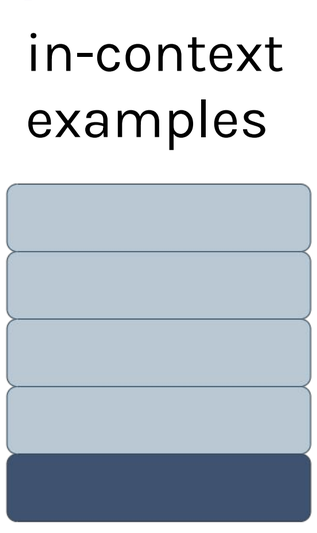
While being creative at scale is challenging, **evaluating** examples is easy! And there has been remarkable progress in **open-ended text generation**. Can we leverage the **generative strength of LMs** and the **evaluative strength of humans** for dataset curation?

Key Takeaways

→ Introduce new approach for dataset creation based on LM generation and human revision

→ Created WANLI, a new dataset of 108K NLI examples, which leads to better OOD performance across diverse test sets

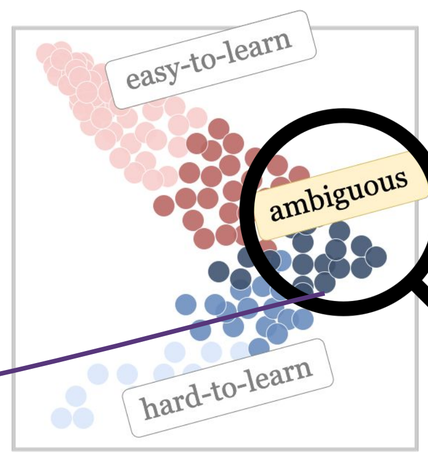
Pipeline for collaborative dataset creation



in-context examples

1) Collection: Collect **groups** of examples that share **the same reasoning pattern**

Data map of MNLI *(Swayamdipta et al., 2020)*



Prior work showed **ambiguous** examples lead to better OOD generalization!

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

1. {premise}
Implication: {hypothesis}
…
5. {premise}
Implication: {hypothesis}
6.
```

We create diverse new examples exemplifying under-represented **reasoning patterns under the task**

GPT-3

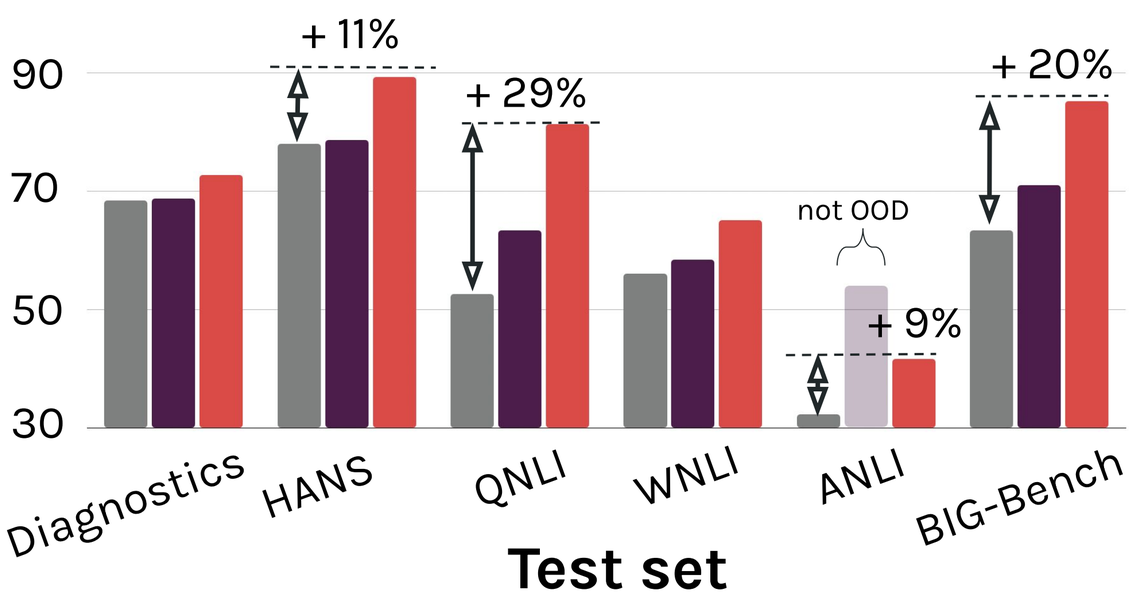
2) Generation: Prompt GPT-3 to create **novel examples** with the **same reasoning pattern**

3) Filtering: Filter with estimated **ambiguity** metric

4) Human review: Humans optionally **revise** for clarity and fluency, and assign a **gold label**

	Ambiguous **MNLI** Ex	Generated **WANLI** Ex	Reasoning
Entailment	P: Salinger ***wrote*** … ***letters to*** … young female writers. H: … young female writers ***received*** … ***letters from*** Salinger…	P: The… ***schools have*** … ***students*** … ***from families*** with no… financial difficulties. H: ***Families*** with no… financial difficulties ***send*** their ***children to*** the… ***schools***.	Describing the same situation with a different predicate & arrangement of arguments
Neutral	P: … Republicans sold big political ***donors meals*** … H: ***It is illegal*** for a party to solicit products to donors.	P: … students… tried to ***organize a union***. H: ***It was illegal*** for the students to organize a union.	Illegal things can happen

Training on WANLI improves OOD test performance across the board



Model:
Roberta-large

Training set *(size)*

- MNLI *(393K)*
- MNLI + SNLI + Adv NLI *(943K)*
- WANLI ***(103K)***

WANLI has fewer artifacts

- Lower performance from **hypothesis-only** model
- Fewer (and different) **lexical correlations**
- Less correlation between label and **semantic similarity** of P and H