

Ambiguity and disagreement

Presenter:

Jiafei Duan:

Jury Learning: Integrating Dissenting Voices into Machine Learning

Weikai Huang:

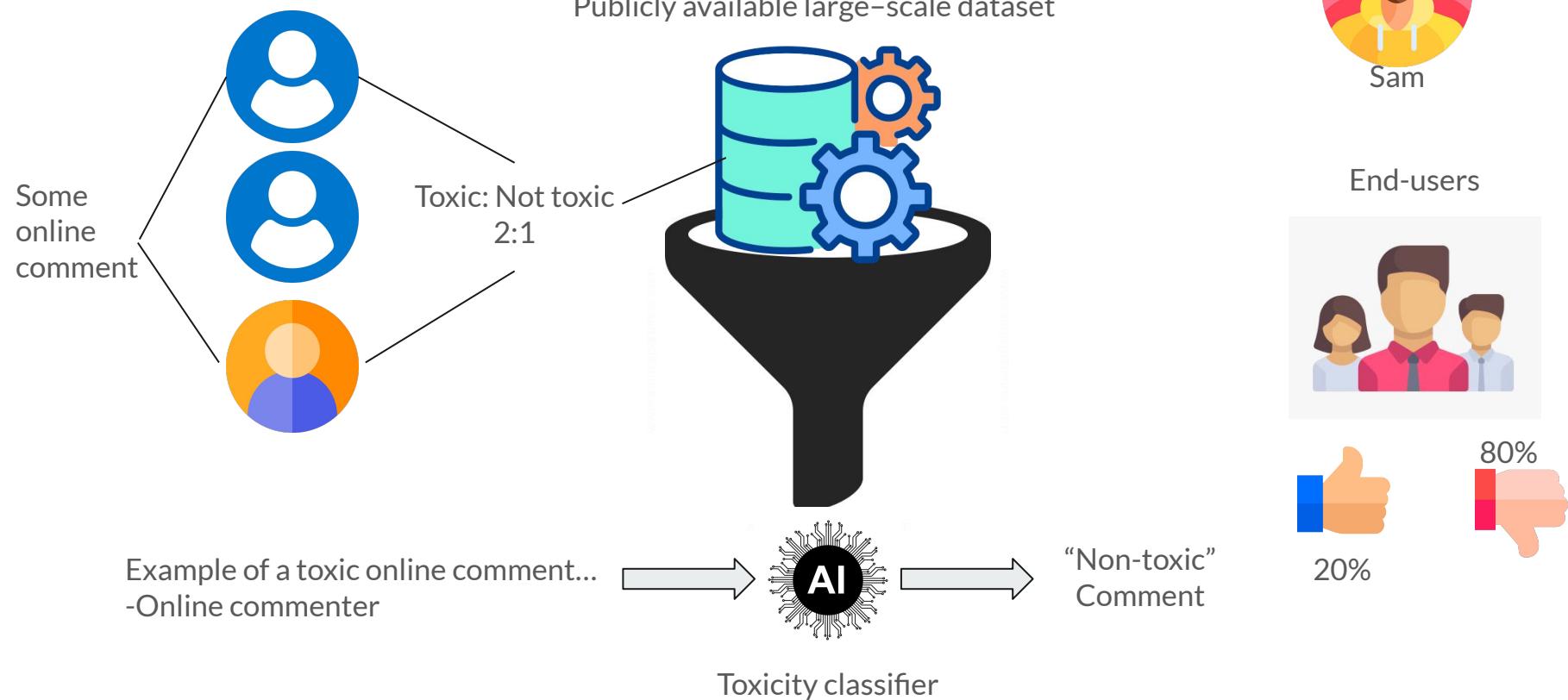
We're Afraid Language Models Aren't Modeling Ambiguity

Jury Learning

Integrating Dissenting Voices Into Machine Learning Models



Example scenario



Looking for the reason behind it all

Explicit composition



Seniors find more comments to be toxic, while 18-35 year olds find fewer comments to be toxic.



Men are more likely to rate borderline comments as not toxic, but also noting that there are many women on her platform

How can Jury Learning help Sam with his problem?

Toxicity dataset



Jury config 1



Jury config 2



Jury config 3



Jury config N

.....

End-users



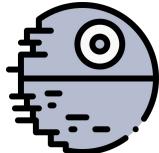
Voice matters



Healthy spaces and communities should have their own distinct norms/values and that should be encapsulated into the data we train our models



Non-parents shouldn't decide which topics are fair game in a parenting forum



Different standards around political issues in r/StarWars should be different r/USA

One model trained from one dataset won't fit all, nor provide opportunity to determine what will fit the context

Disagreements are in more than just one domain



Poster design



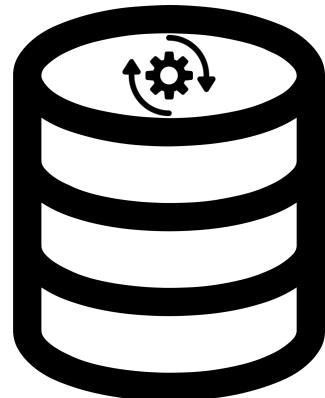
Political issues



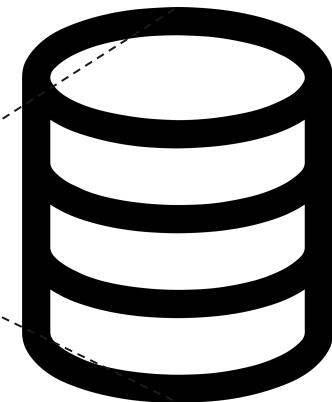
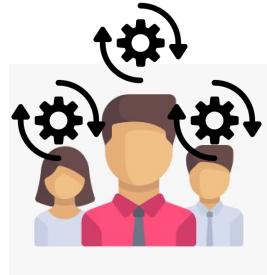
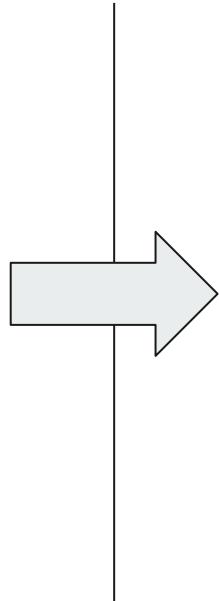
Medical diagnostic

Intuition behind Jury Learning

Model individual people, and not an aggregated pseudo-human.
Every user/annotator isn't just a number on a statistic.

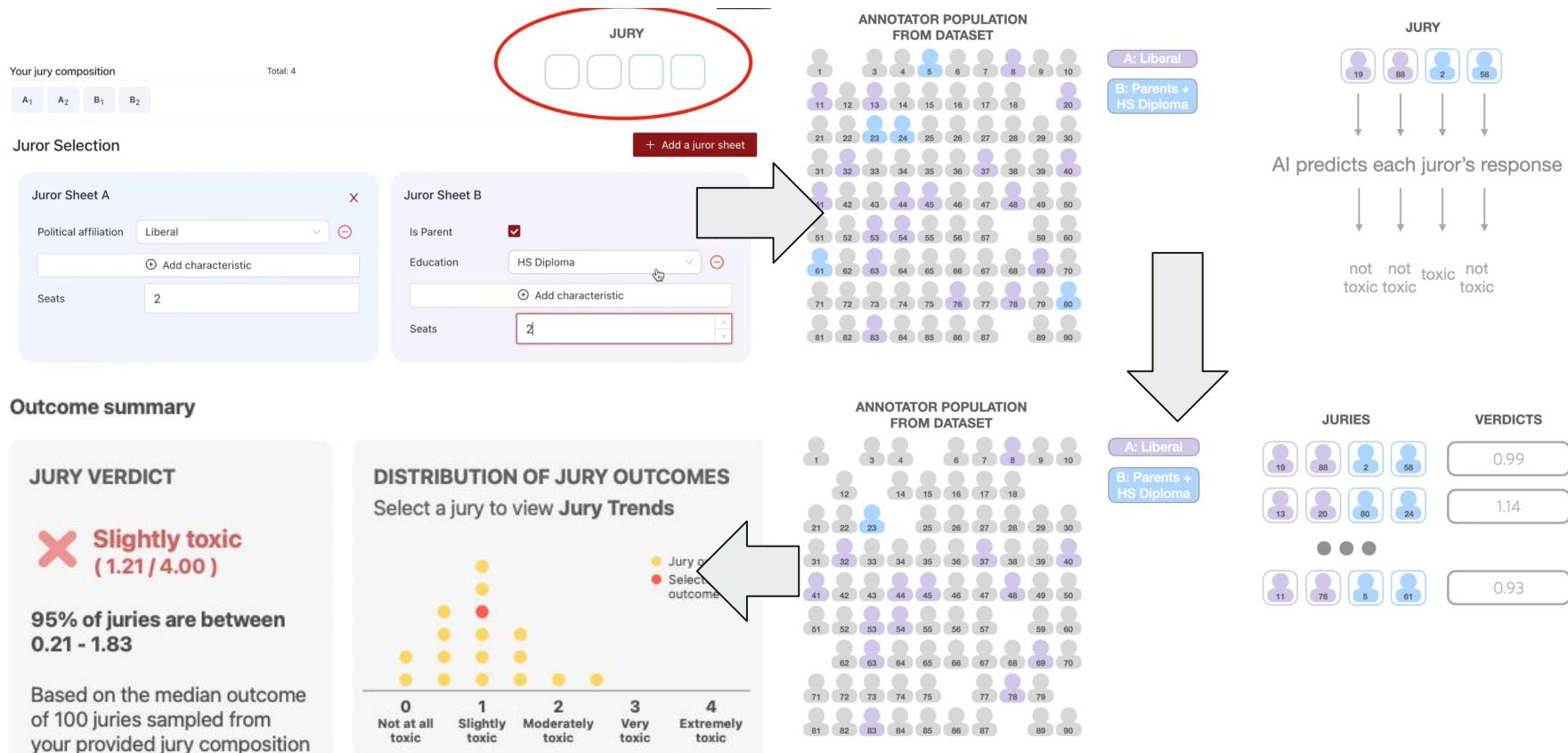


Dataset



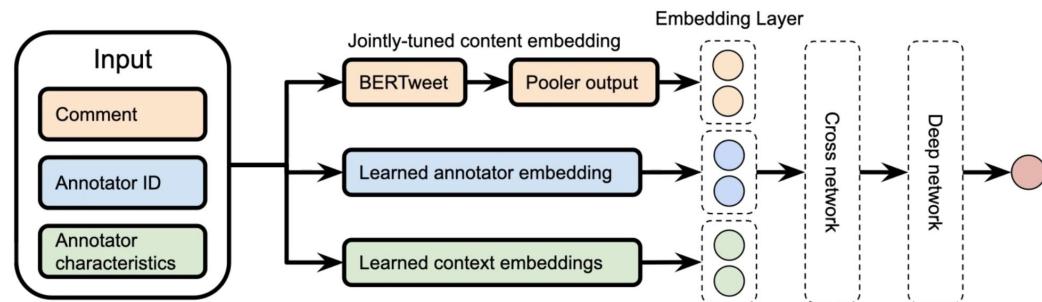
Model the annotators

How does Jury Learning works in detail?



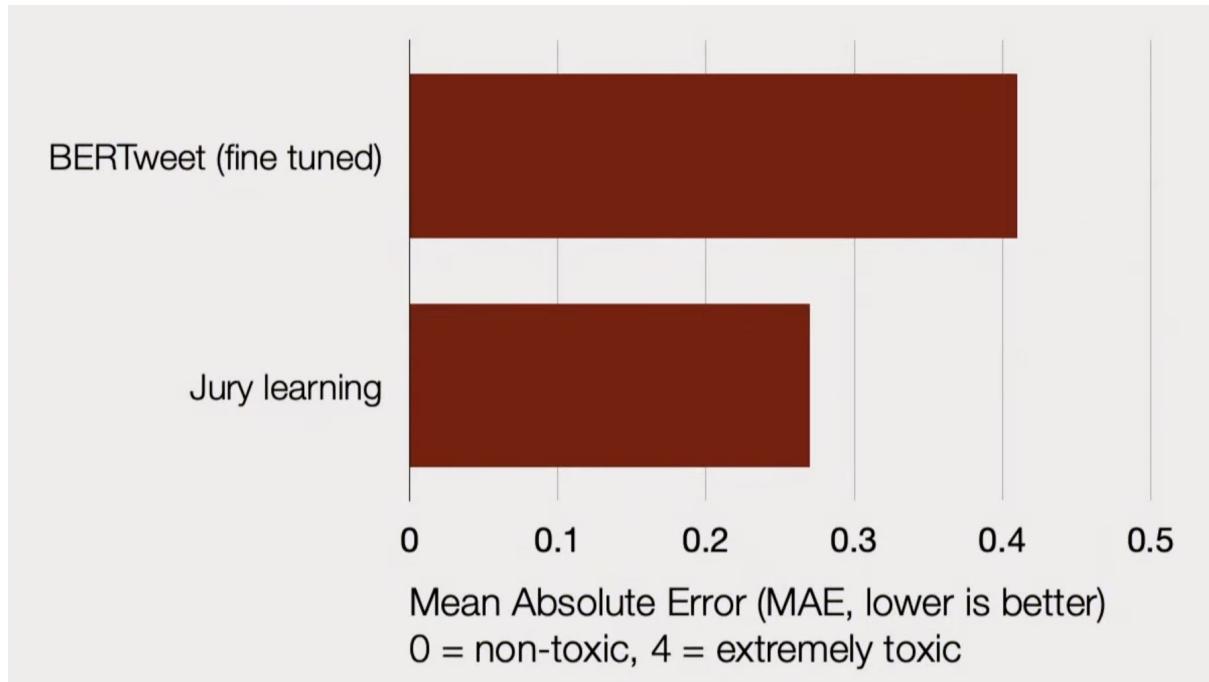
Training strategy for Jury Learning

	19	88	2	58	characteristic 1 characteristic 2 characteristic 3 characteristic 4 ...	?	?	✓	?	✗
training example 0	✓	?	?	✓	characteristic 1 characteristic 2 characteristic 3 characteristic 4 ...	?	?	✓	?	✗
training example 1	?	✗	✓	?	?	✗	✓	?	✓	?
training example 2	?	?	?	?	...	?	?	?	?	?
training example 3	✗	?	✓	?	?	✓	?	✓	?	✗
training example 5	?	?	?	?	?	?	?	?	?	?
•••	•••									
training example 9999	✓	?	✗	✓	?	✓	?	✗	✓	✗
↓ ↓ ↓ ↓										
previously unseen example	AI predicts each juror's response	↓	↓	↓	↓	✗	✓	✓	✓	✓



Technical evaluation

They design a specific test set with 5,000 comments and 24,545 annotations

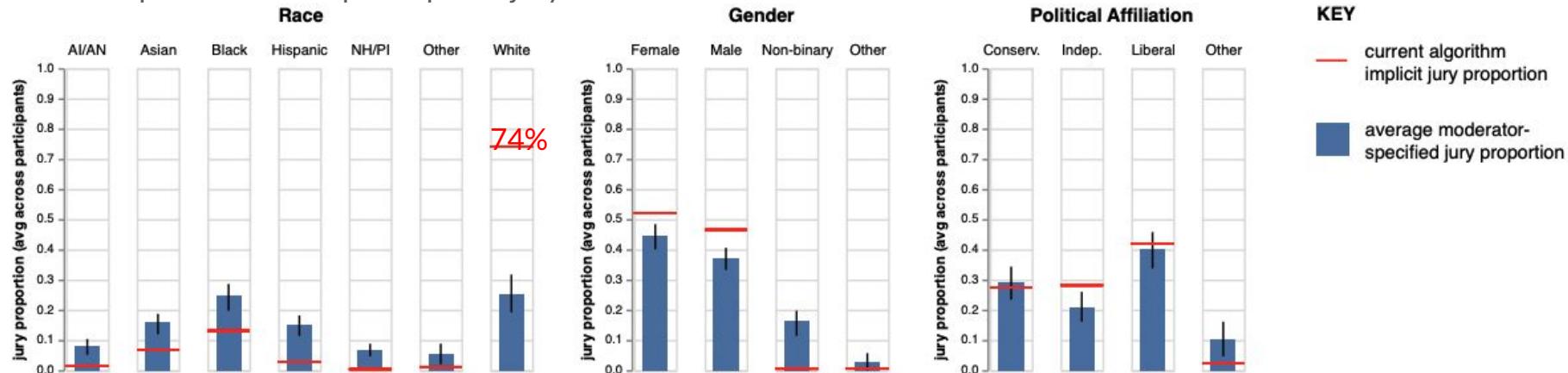


Modelling annotators leads to better performance, even on the traditional aggregated task

Can Jury Learning effectively address the inherent implicit disagreements embedded in the original dataset?

Red = Representation in state of the art training dataset

Blue = Representation in participants' jury classifiers



Study: They recruited 18 moderators to author juries comparing with current algorithm proportion.

Result: People tend to have a more diverse racial representation than today's dataset.

Question to ask: How many of these datasets are still out there?



Do user-specified juries result in different prediction outcome than a standard classifier?

Participants' juries change 14% of classifications vs SoTA classifier

Most likely to flip: contentious, divisive issues

- Racism
- Death/suicide
- LGBTQ+
- Mental illness/disorders
- Cops



Least likely to flip: uncontroversial issues (good and bad)

- ### Largely innocuous topics
- Thank-yous
 - Happiness
 - Hugs
 - Weddings



- ### Largely offensive topics
- Human trafficking
 - R-word
 - Racial/ethnic slurs

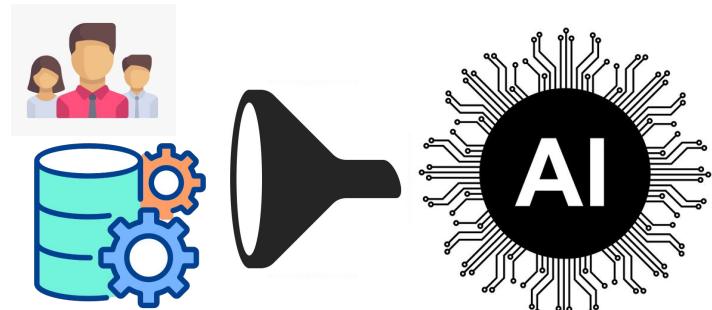


Participants' juries change 14% of classifications vs SoTA classifier

Behaviors on the edge case often becomes de facto platform policy, hence they are important.

Different angle on improving ML system's output biases.

[Scheuerman et al, CSCW 2021; Suresh et al, EAAMO 2021]



Select or train the appropriate voices for the given context

Takeaways

- Many dataset contains an inherent ‘hidden’ jury!
- Increasing or decreasing different representation affects classifier performances.
- Disagreements between annotators are inevitable, but it is important for us to take that into consideration when curating training data.

Bio: Jury Learning: Integrating Dissenting Voices into Machine Learning Models



Mitchell Gordon

Ph.D Student, Computer Science
Stanford University

Advisors: Michael Bernstein &
James Landay

Areas of Interest: Human-Computer
Interaction, Social Computing



Michelle Lam

Ph.D Student, Computer Science
Stanford University

Advisors: Michael Bernstein &
James Landay

Areas of Interest: Human-Computer
Interaction, Social Computing,
Human-Centered AI, Algorithmic
Fairness



Joon Sung Park

Ph.D Student, Computer Science
Stanford University

Advisors: Michael Bernstein & Percy
Liang

Areas of Interest: Human-computer
interaction, natural language interface,
social computing, algorithm audits,
generative agents, human-centered AI

Jury Learning: Integrating Dissenting Voices into Machine Learning Models



Kayur Patel

Usable ML Hipster, Apple since June 2017

Previously:
Researcher at Google AI
(2012-2017)

Ph.D in Computer Science,
University of Washington (2012)
Advisors: James Fogarty & James Landay

Areas of Interest:
Human-Computer Interaction,
Machine Learning



Jeffrey Hancock

Professor of Communication,
Stanford University (2015)

Previously:
Professor of Info Science &
Communication at Cornell
(2002-2015)

Ph.D in Psychology, Dalhousie
University (2002)
Advisor: Philip J. Dunham

Areas of Interest: Social Media,
AI-mediated Communication,
Deception, Trust,
Computer-mediated communication



Tatsunori Hashimoto

Assistant Professor, Computer Science
Stanford University (2020)

Previously:
Researcher, Microsoft (2019-20)
Postdoc, Stanford w/ Percy Liang &
John Duchi (2016-19)

Ph.D, MIT CSAIL (2016)
Advisors: Tommi Jaakkola & David Gifford

Areas of Interest: Long-tail behavior,
Fairness, Understanding, OOD
Generalization, NLP, Comp Bio



Michael Bernstein

Associate Professor, Computer Science
Stanford University (2013)
STMicroelectronic Faculty Scholar (2019)

Previously:
Postdoc, Facebook Data Sci (2012)

Ph.D, MIT EECS (2016)
Advisors: David R. Kriger & Robert Miller

Areas of Interest: Social computing
systems, Human-Computer Interaction,
Social Computing

Ambiguity

We're Afraid Language Models aren't Modeling Ambiguity

EMNLP 2023

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West,
Alexander Koller, Swabha Swayamdipta, Noah A. Smith, Yejin Choi

Why is ambiguity important?

- Ambiguity is the presence of multiple interpretations
- One of the essential features of language is balancing between clarity and ease
- Part of language understanding is recognizing multiple interpretations

This is not my favorite

I HATE it!

It is fine!

Background: How to represent ambiguity

AmbigQA (Min et al 2020)

- Propose a benchmark(AmbigQA) and a dataset(AmbigNQ) over 10000 questions
- Represent the ambiguity by splitting the ambiguous text to several unambiguous QA pairs



AMBIGQA Input

When did harry potter and the sorcerer's stone movie come out?

Harry Potter and the Philosopher's Stone (film)
From Wikipedia, the free encyclopedia

The film had its world premiere at the Odeon Leicester Square in London on **4 November 2001**, with the cinema arranged to resemble Hogwarts School. (...) The film was released to cinemas in the United Kingdom and United States on **16 November 2001**.

AMBIGQA Output

Q: When did harry potter and the sorcerer's stone movie come out at the Odeon Leicester Square?
A: 4 November 2001

Q: When did harry potter and the sorcerer's stone movie come out in cinemas?
A: 16 November 2001



Figure 1: An AMBIGNQ example where the prompt question (top) appears to have a single clear answer, but is actually ambiguous upon reading Wikipedia. AMBIGQA requires producing the full set of acceptable answers while differentiating them from each other using disambiguated rewrites of the question.

Background: The SOTA of AmbigQA

Answering Ambiguous Questions through Generative Evidence Fusion and Round-Trip Prediction (Gao et al 2020)

Model	F1 _{ans} (all)		F1 _{ans} (multi)		F1 _{BLEU}		F1 _{EDIT-F1}		Comb.	
	dev	test	dev	test	dev	test	dev	test	dev	test
DISAMBIG-FIRST (Min et al., 2020)	28.1	24.8	21.9	18.8	4.2	4.0	2.7	2.2	30.8	27.0
DPR Reader (Min et al., 2020)	37.1	32.3	28.4	24.8	13.4	11.3	6.6	5.5	43.7	37.8
SPANSEQGEN (Min et al., 2020)	39.7	33.5	29.3	24.5	13.4	11.4	7.2	5.8	46.9	39.3
SPANSEQGEN (ensemble)	41.2	35.2	29.8	24.5	13.6	10.6	7.4	5.7	48.6	40.9
SPANSEQGEN (ensemble + co-training)	42.3	35.9	31.7	26.0	14.3	11.5	8.0	6.3	50.3	42.2
REFUEL (single model)	48.4	41.7	37.0	32.7	16.0	14.8	11.2	9.0	59.6	50.7
+ Round-Trip Prediction	48.3	42.1	37.3	33.3	16.2	15.3	11.8	9.6	60.1	51.7

Background: Can LM handle ambiguity naturally?

The Language Model Understood the Prompt was Ambiguous (Aina et al 2021)

- LMs are capable of tracking multiple interpretations simultaneously.
- The degree of uncertainty varies across different constructions and contexts.
- Disambiguating cues often lead to the selection of the correct interpretation, which is similar to humans.

Language models might be capable to handle ambiguity with its underlying mechanisms!

What is the upcoming challenge?

- Can we represent ambiguity with greater granularity and in a more quantitative manner?
- Can the current cutting-edge models (we have gpt-4, Llama, etc) handle ambiguity as good as human?
- Calling for high quality data (with human annotations)

These 3 questions lead to the paper we are gonna talk about today!

We're Afraid Language Models aren't Modeling Ambiguity

EMNLP 2023

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West,
Alexander Koller, Swabha Swayamdipta, Noah A. Smith, Yejin Choi

How to represent ambiguity?

Represent Ambiguity

He always ignores his mother's advice to follow his own dreams.

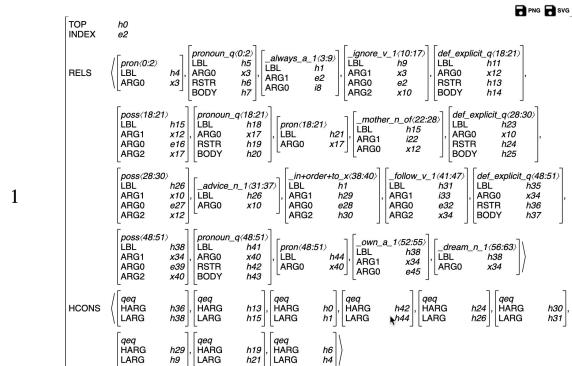
Represent Ambiguity

He always ignores his mother's advice to follow his own dreams.

Showing 231 of 231 analyses.

Showing 231 of 231 analyses.

He always ignores his mother's advice to follow his own dreams.



How can we capture ambiguities in a natural way?

Represent Ambiguity

Premise He always ignores his mother's advice to follow his own dreams.

... in order to follow his own dreams.

... which is to follow his own dreams.

entails



disambiguate

contradictive

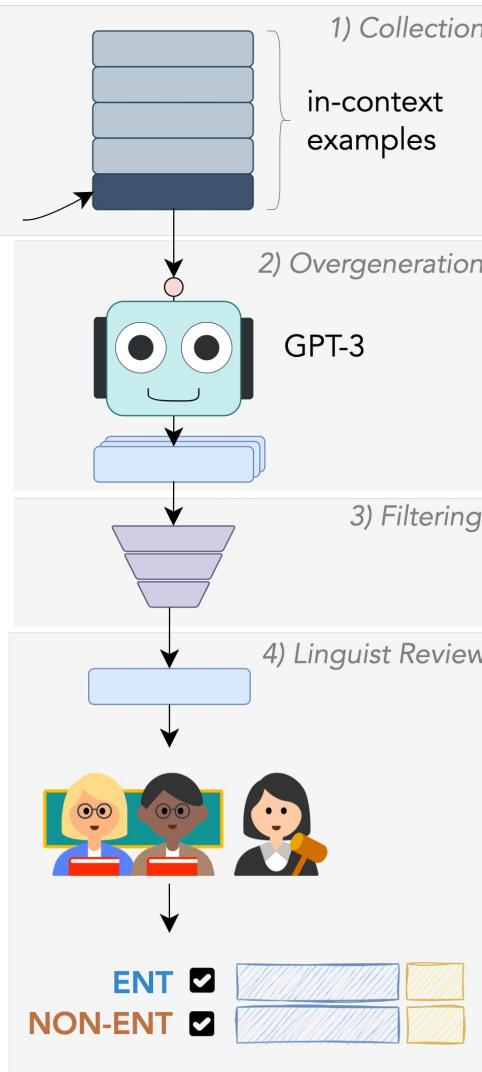
Hypothesis He follows his dreams.

Capture ambiguity through its effect on entailment relations with hypothesis

How to get high quality data?



Examples with
Annotator Disagreement



collect examples likely to be ambiguous from previous dataset

GPT-3 generates new examples with the same pattern

automatically filter generations

human annotations with the SET of possible interpretations

Data Creation Pipeline

Example	Disambiguation 1	Disambiguation 2	Type
P: I'm afraid the cat was hit by a car. H: The cat was not hit by a car. [?NEUTRAL, CONTRADICT] ↗: [7 N, 2 C]	P: I'm worried... NEUTRAL ↗: [7 N]	P: I'm sorry to share that... CONTRADICT ↗: [9 C]	Pragmatic (44.8%)
P: John and Anna are married. H: John and Anna are not a couple. [?NEUTRAL, CONTRADICT] ↗: [5 N, 4 C]	P: ... are both married. NEUTRAL ↗: [7 N, 2 E]	P: ... are married to each other. CONTRADICT ↗: [9 C]	Lexical (20.0%)
P: This seminar is full now, but interesting seminars are being offered next quarter too. H: There will be more interesting seminars... [?ENTAIL, NEUTRAL] ↗: [7 E, 2 N]	H: There will be more seminars... that are interesting. ENTAIL ↗: [9 E]	H: There will be seminars... that are more interesting. NEUTRAL ↗: [9 N]	Syntactic (8.6%)
P: The novel has been banned in many schools because of its explicit language. H: The novel has not been banned in many schools. [?NEUTRAL, CONTRADICT] ↗: [4 N, 5 C]	H: There are many schools where the novel has <u>not been</u> banned. NEUTRAL ↗: [9 N]	H: It is not the case that the novel has been banned in many schools. CONTRADICT ↗: [9 C]	Spatial (7.6%)
P: It is currently March, and they plan to schedule their wedding for <u>next</u> December. H: They plan to schedule... for next year. [?ENTAIL, CONTRADICT] ↗: [3 E, 2 N, 4 C]	P: ... for December next year. ENTAIL ↗: [9 E]	P: ... for the coming December. CONTRADICT ↗: [9 C]	Coreference (2.9%)
P: It is difficult to believe that the author of such a masterpiece could have been only 23 years old. H: The author of the masterpiece was only 23. [?ENTAIL, NEUTRAL] ↗: [3 E, 6 N]	P: It is shocking that... ENTAIL ↗: [9 E]	P: It is questionable that... NEUTRAL ↗: [9 N]	Figurative (1.9%)
P: A new study has found that nearly half of all Americans are in favor of gun control. H: The study found that half of all Americans are in favor of gun control. [?ENTAIL, CONTRADICT] ↗: [1 E, 2 N, 6 C]	H: ... that exactly half of all Americans... CONTRADICT ↗: [8 C, 1 N]	H: ... that about half of all Americans... ENTAIL ↗: [9 E]	Other (14.3%)

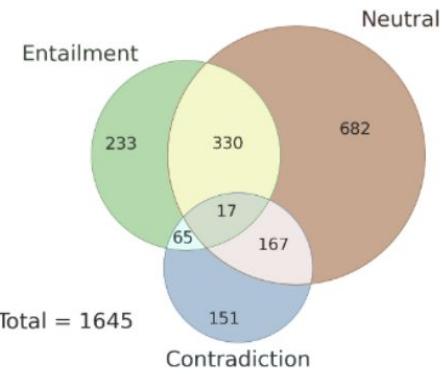


Figure 3: Distribution of set labels in AMBIENT.

combines curated and generated-then-annotated examples, consists of 1,645 examples

Statistic

*Does ambiguity explain
annotator disagreement?*

97% of disambiguations
marked plausible

Premise

Hypothesis

He always ignores his mother's advice to follow his own dreams.
He follows his dreams.



P ... in order to...

H He follows his dreams.



entail non-entail



Fleiss κ agreement on
ambiguous example: 0.12



P ... which is to...

H He follows his dreams.



entail non-entail

Fleiss κ agreement on
disambiguated examples: 0.67

Premise

He always ignores his mother's advice to follow his own dreams.

Hypothesis

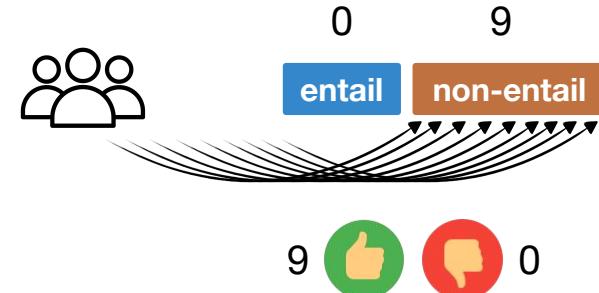
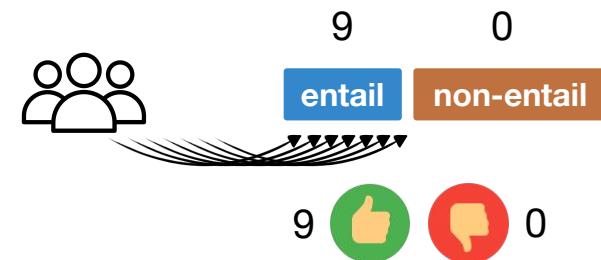
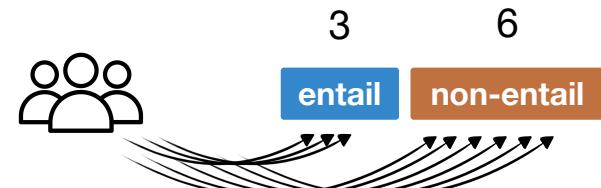
He follows his dreams.

Premise 1

He always ignores his mother's advice, in order to follow his own dreams.

Premise 2

He always ignores his mother's advice, which is to follow his own dreams.



Premise

Hypothesis

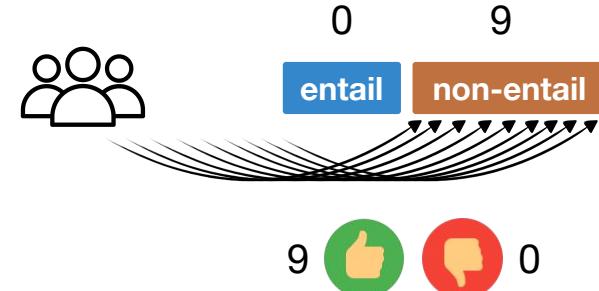
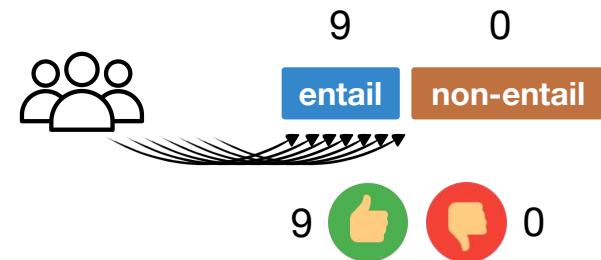
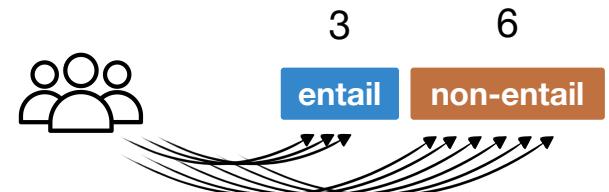
Premise 1

Premise 2

It is difficult to believe that the author of such a masterpiece could have been only 23 years old. The author of the masterpiece was only 23.

It is shocking that the author of such a masterpiece could have been only 23 years old.

It is questionable that the author of such a masterpiece could have been only 23 years old.



Ambiguity is a source of label variation!

*Can LMs handle ambiguity
as good as human?*

How do humans handle ambiguity?

1. Disambiguating by giving different interpretations
2. Identify ambiguity

Can LMs disambiguate by giving different interpretations

In each example, you will be given some **context** and a **claim**, where the correctness of the **claim** is affected by some ambiguity in the **context**.

Enumerate two interpretations of the **context** that lead to different judgments about the **claim**.

Context: He always ignores his mother's advice to follow his own dreams.

Claim: He follows his dreams. Given the **context** alone, is this **claim** true?

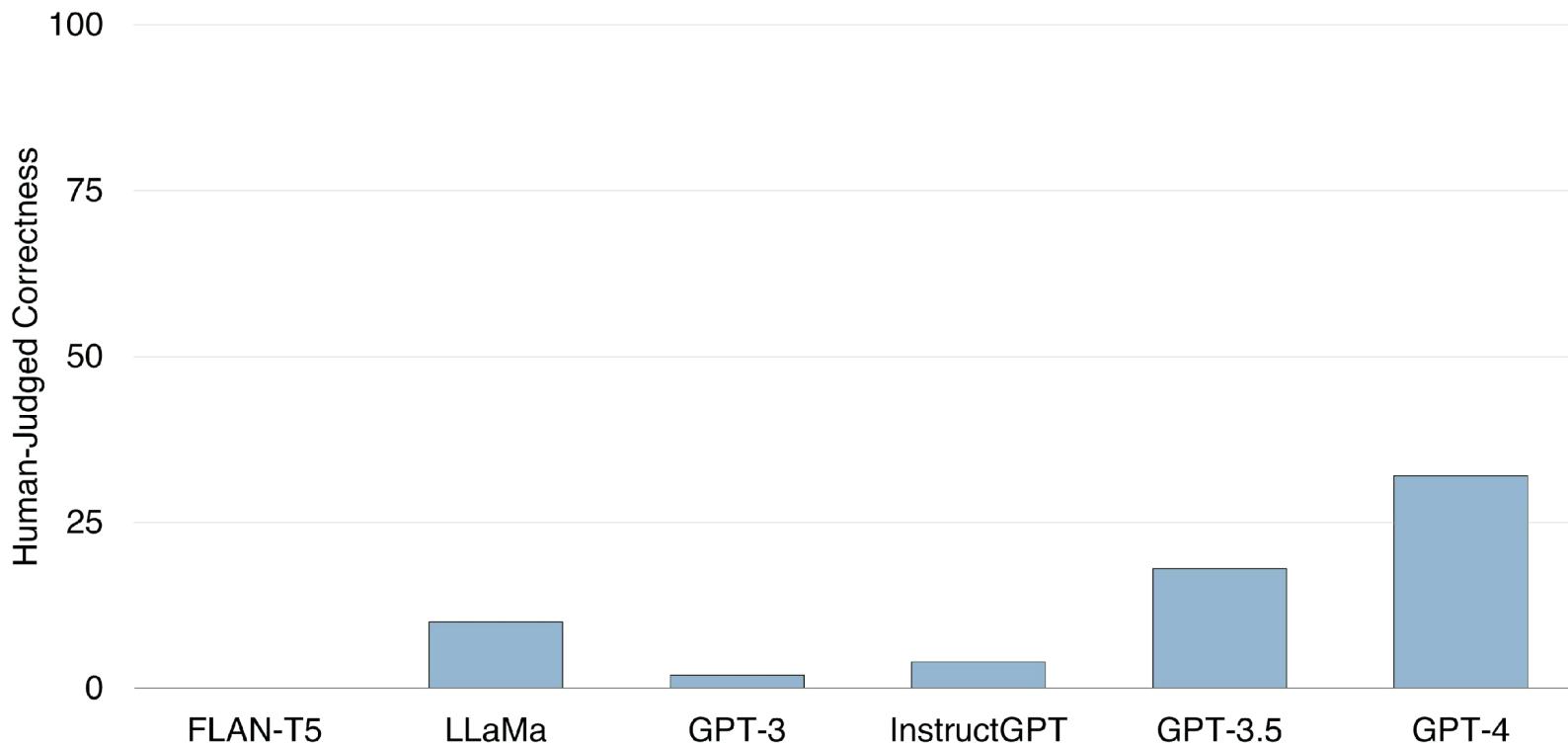
}

instruction

}

test example

Correctness of generated disambiguations



Can LMs identify ambiguity?

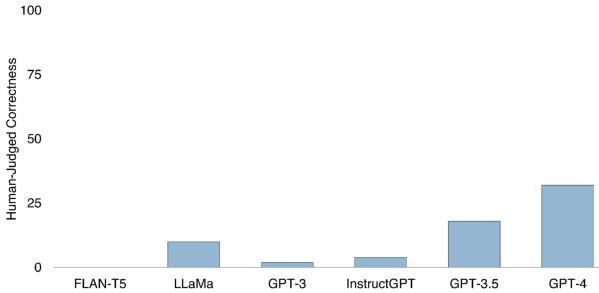
He always ignores his mother's advice to follow his own dreams.

- This may mean: He follows his dreams. TRUE
- This does not necessarily mean: He follows his dreams. TRUE
- This cannot mean: He follows his dreams. FALSE
- This can only mean: He follows his dreams. FALSE

Ambiguation recognition accuracy



Conclusion



Language models are still struggling with ambiguity!

Takeaways

- Language is ambiguous!
- Capture ambiguity through its effect on entailment relations.
- Ambiguity is something that language models still struggle with.

Reference

- Gao, Yifan et al. "Answering Ambiguous Questions through Generative Evidence Fusion and Round-Trip Prediction." Annual Meeting of the Association for Computational Linguistics(2020).
- Aina, Laura and Tal Linzen. "The Language Model Understood the Prompt was Ambiguous: Probing Syntactic Uncertainty Through Generation." ArXiv abs/2109.07848 (2021)
- Min, Sewon et al. "AmbigQA: Answering Ambiguous Open-domain Questions." Conference on Empirical Methods in Natural Language Processing (2020).
- Liu, Alisa et al. "We're Afraid Language Models Aren't Modeling Ambiguity." ArXiv abs/2304.14399 (2023)
- Kumar, Deepak, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. "Designing toxic content classification for a diversity of perspectives." In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pp. 299-318. 2021.
- Nguyen, Dat Quoc, Thanh Vu, and Anh Tuan Nguyen. "BERTweet: A pre-trained language model for English Tweets." *arXiv preprint arXiv:2005.10200* (2020).
- Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems. In Proceedings of the Web Conference 2021. 1785–1797.

Any questions?

Discuss Questions

Jury Learning

- Given the vast amount of data out there today, how can we implement Jury Learning given that it requires re-collecting annotators' information?
- How and who should be determining jury composition for any given dataset?
- Is Jury learning an universal approach that can work for all domains?
- Instead of collecting or having a dataset that has explicit information about the annotators, can there be future work that investigate unsupervised approach to reveal different voices in the dataset itself?

We're Afraid Language Models Aren't Modeling Ambiguity

- Is the current representation of ambiguity sufficient to capture all types of ambiguity, or are there better ways to represent it?
- Can prompting methods such as CoT (Chain of Thought), Self-Consistency, or ReAct that might enhance the performance of models in dealing with ambiguity?
- Can RLHF improve the performance of Language Models in handling ambiguity?
- Since the size of the data is limited due to human annotation, is there anyway to get more high-quality data without human annotations?

Any questions?