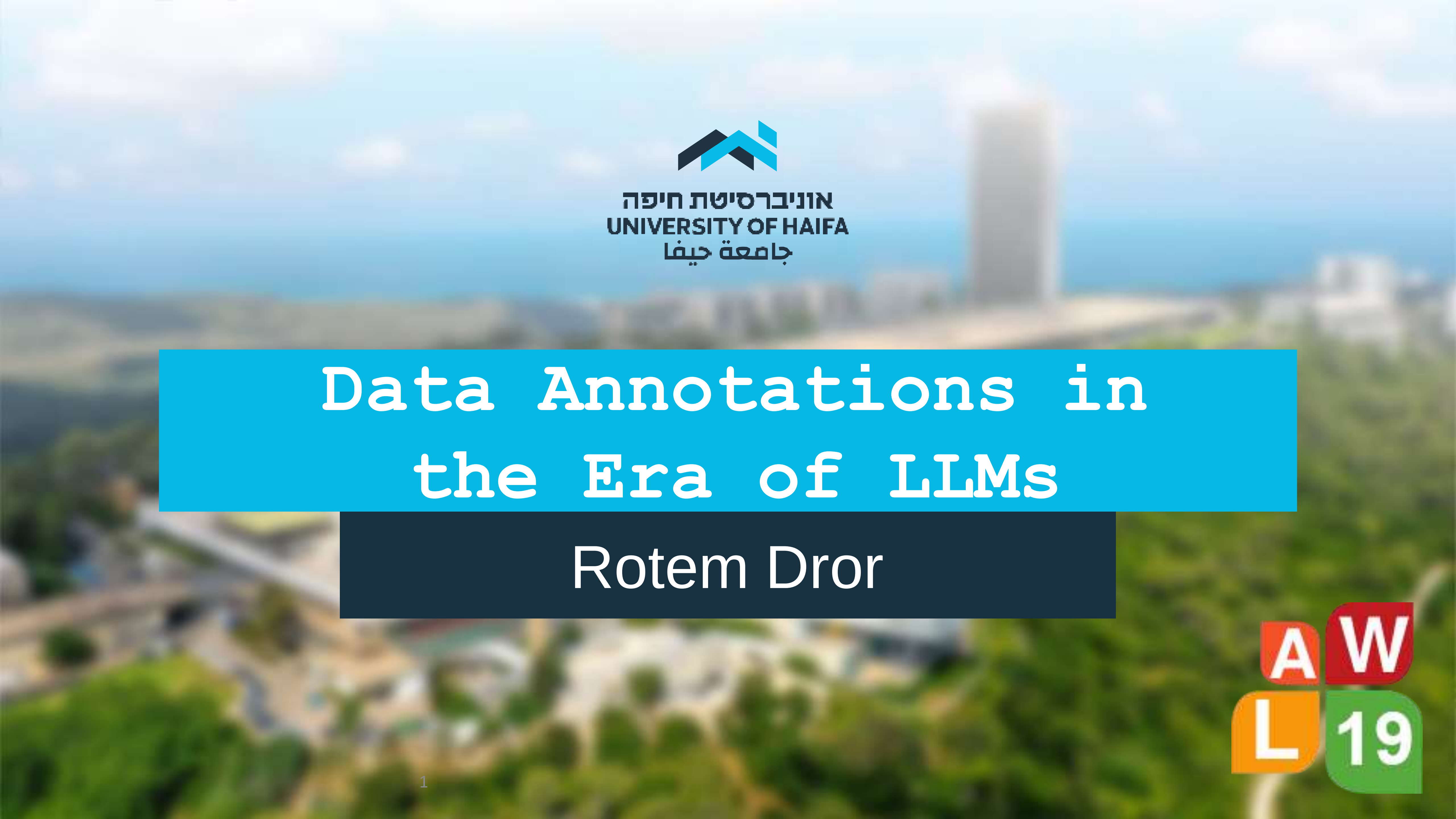


אוניברסיטת חיפה
UNIVERSITY OF HAIFA
جامعة حيفا

Data Annotations in the Era of LLMs

Rotem Dror

Guiding Questions

- Do we need to annotate data in NLP nowadays?
- Do we need humans to annotate data?
- Can we trust LLMs as judges/annotators?
- What about subjective annotations?

Why Annotations Still Matter?

LLMs are generalists

Even if you don’t train the model, you need to evaluate it.

Fairness

The Changing Landscape of Annotations

⍰ 💻 Manual Annotation Workflow

- Task Definition
- Guidelines Creation
- Human Annotation
- Review
- Aggregation

✕ Costly
✕ Time-consuming
✕ Inconsistent (esp. subjective tasks)
✕ Heavy quality control

⍰ LLM-Augmented Workflow

- Task Definition
- Prompt Engineering
- LLM Annotation
- Review
- Optimal Human Adjudication

✓ Scalable
✓ Fast
⚠ Prompt Sensitivity
⚠ Confidence ≠ Accuracy

Do We Still Need Human Annotations?

- **Short answer:** Yes.
- **Slightly longer answer:** Yes, but not as much as we did five years ago.
- **Long answer:**
  - We no longer need humans to annotate entire datasets — just a representative sample 🎁
  - Human input is essential for supervising and validating automatic annotations 👀
  - For this smaller subset, **expertise and skilled annotators** are preferred 🎓

LLM-as-A-Judge VS. LLM-as-An-Annotator

What is “LLM-as-an-annotator”?

- Using LLMs for annotation, evaluation, or labeling tasks that are traditionally performed by human annotators.
- LLM-as-a-judge is a special case: LLMs that evaluate outputs of other models (LLMs).

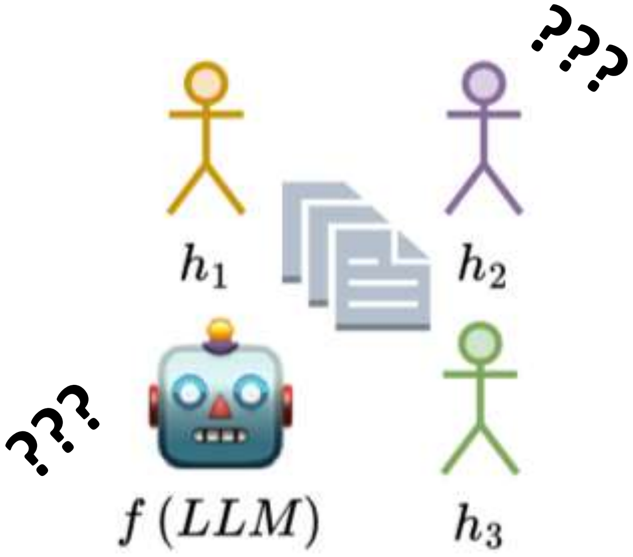
Why “LLM-as-a-judge”?

- It is cheaper, faster, requires less effort, and is less labor-intensive.
- Sometimes, LLMs are better than humans… when?

The Alternative Annotator Test:

How to Statistically Justify Replacing Human Annotators with LLMs

Nitay Calderon   Roi Reichart   Rotem Dror



Faculty of Data and Decision Making, Technion – Israel Institute of Technology
Faculty of Computer and Information Science, University of Haifa

Can We Trust LLM-as-a-Judge?

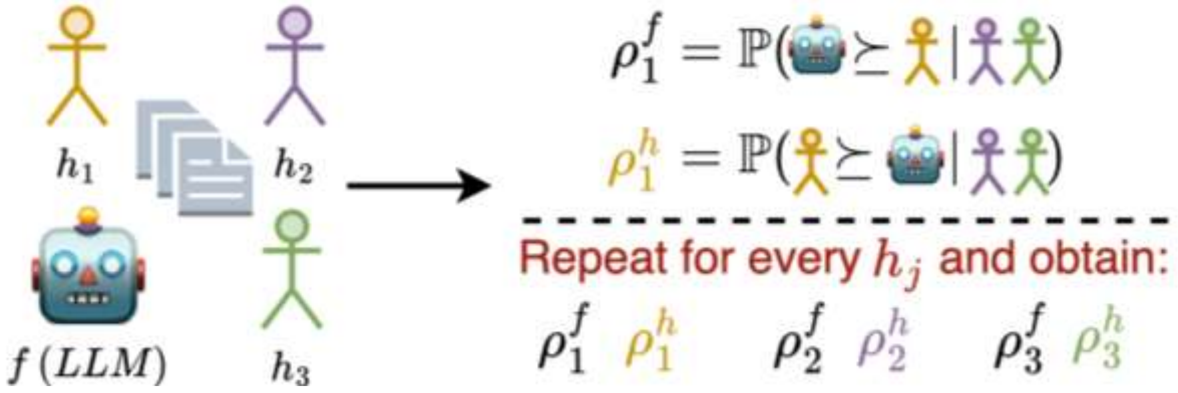
- **LLMs directly shape the results, findings, and insights of scientific papers.**
  - Not only in NLP, but also in medicine, psychology, social science…
- Many papers **do not report any alignment measures** between LLMs and humans.
- Those that do typically use traditional measures such as:
  - % agreements, F1 score, IAA kappas, correlations…
- There is **no established standard or criterion** for making a yes/no decision.
  - “Is an F1 score of 0.6 sufficient?”
- The decision **requires statistical rigor,** which is often lacking in how researchers apply traditional measures.

The Alternative Annotator Test

- Researchers should demonstrate that the LLM offers a **comparable alternative** to recruiting human annotators.
- In other words, when factoring in the **cost-benefit and efficiency advantages** of LLM annotations, they should be **as good or better than human annotations**.
- What is better?
  - In some cases, agrees more with the majority vote.
  - In other cases, reliable and consistent.

The **alt-test:** How it works

1. **Leave-one-out:** Exclude each annotator in turn, and estimate the probabilities that the LLM aligns better with the remaining annotators than the excluded one, and vice versa.


h_1
h_2
h_3
$f\,(LLM)$
$\rho_1^f = \mathbb{P}(🤖 \succeq 🧍 \mid 🧍🧍)$
$\rho_1^h = \mathbb{P}(🧍 \succeq 🤖 \mid 🧍🧍)$
Repeat for every h_j and obtain:
ρ_1^f ρ_1^h ρ_2^f ρ_2^h ρ_3^f ρ_3^h

2. Conduct hypothesis tests to compare the probabilities and obtain p-values.

$$\text{x3}\begin{cases} H_{null}: \rho_j^f \le \rho_j^h - \varepsilon \\ H_{alt}\ \ : \rho_j^f > \rho_j^h - \varepsilon \end{cases} \longrightarrow \begin{matrix} p\text{-value } 1 \\ p\text{-value } 2 \\ p\text{-value } 3 \end{matrix}$$

3. Apply an FDR procedure and identify the rejected hypotheses.

$$FDR\begin{pmatrix} p\text{-value } 1 \\ p\text{-value } 2 \\ p\text{-value } 3 \end{pmatrix} \longrightarrow \begin{matrix} ✓\ H_1 \text{ is rejected} \\ ✗\ H_2 \text{ is not rejected} \\ ✓\ H_3 \text{ is rejected} \end{matrix}$$

4. Calculate the LLM's winning rate and determine if it can replace humans.

Winning Rate $\omega = \dfrac{✓✗✓}{3} = 0.67$

If $\omega \geq 0.5 \Rightarrow$ 🤖 can replace 🧍

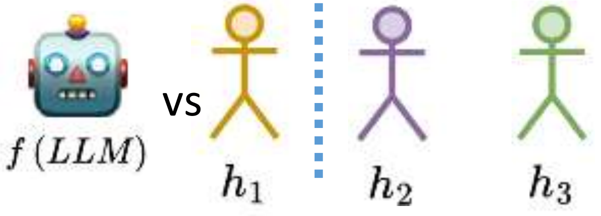
1. Leave-one-out and estimate the relative advantage probabilities

S(.) measures how well the annotation of the LLM or the excluded human aligns with those of the remaining annotators.

$$W^f_{i,j} = \begin{cases} 1, & \text{if } S(f, x_i, j) \geq S(h_j, x_i, j) \\ 0, & \text{otherwise} \end{cases}$$

$$\rho^f_j = \hat{\mathbb{P}}(\mathbf{LLM} \succeq h_j) = \hat{\mathbb{E}}[W^f_{i,j}] = \frac{1}{|\mathbb{I}_j|} \sum_{i \in \mathbb{I}_j} W^f_{i,j}$$

The instances annotated by the j-th excluded human

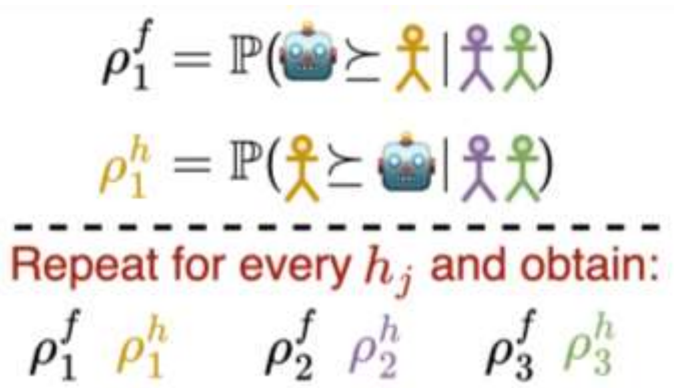


x_i anns: **A** B **A** C

$$ACC(f, x_i, 1) = 0.5$$

$$ACC(h_1, x_i, 1) = 0$$

$W^f_{i,1}$ =1


$\rho^f_1 = \mathbb{P}(\succeq |)$
$\rho^h_1 = \mathbb{P}(\succeq |)$
Repeat for every h_j and obtain:
ρ^f_1 ρ^h_1 ρ^f_2 ρ^h_2 ρ^f_3 ρ^h_3

2. Conduct a hypothesis test: Does the LLM hold an advantage?

$$\text{x3}\begin{cases} H_{null} : \rho_j^f \leq \rho_j^h - \varepsilon \\ H_{alt} \;\; : \rho_j^f > \rho_j^h - \varepsilon \end{cases} \longrightarrow \begin{matrix} p\text{-value } 1 \\ p\text{-value } 2 \\ p\text{-value } 3 \end{matrix}$$

Cost-benefit hyperparameter:
Penalty to the human because
LLMs are faster and cheaper

A paired t-test
(== a t-test for the differences)

$$d_{i,j} = W_{i,j}^h - W_{i,j}^f$$

$$\bar{d}_j = \rho_j^h - \rho_j^f$$

$$t_j = \frac{\bar{d}_j - \varepsilon}{s_j/\sqrt{n}} \quad s_j = \sqrt{\frac{\sum_{i=1}^n \left(d_{i,j} - \bar{d}_j\right)^2}{n-1}}$$

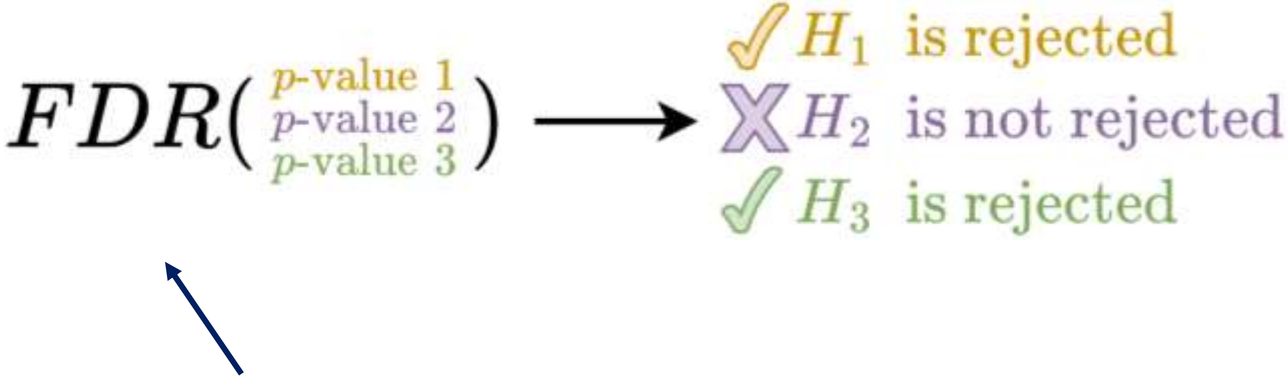
As a rule of thumb, if annotators are:

- Trusted experts (expensive, less accessible) $\varepsilon = 0.2$
- Skilled (undergrads, trusted workers) $\varepsilon = 0.15$
- Crowd-workers (cheap) $\varepsilon = 0.1$

3. Apply an FDR-controlling procedure

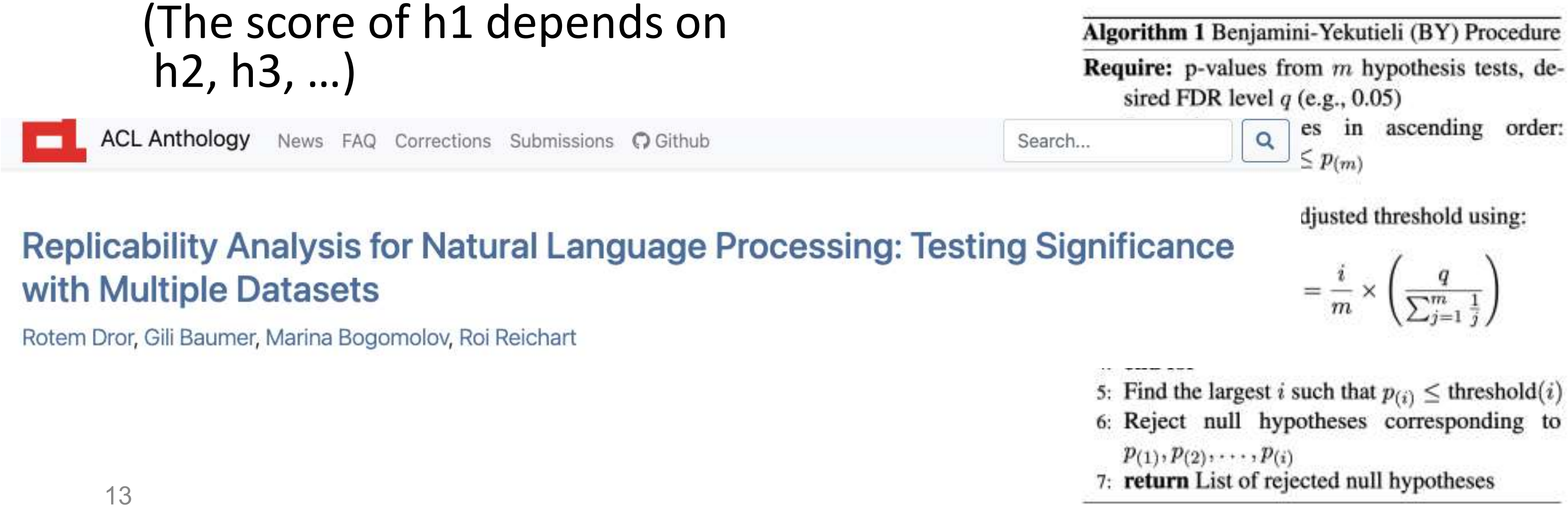
Simply counting the number of rejections is problematic:

- Accumulation of Type-I errors (false rejections)
- Hypotheses are dependent (The score of h1 depends on h2, h3, …)



$FDR(p\text{-value } 1,\ p\text{-value } 2,\ p\text{-value } 3) \longrightarrow$
✓ H_1 is rejected
✗ H_2 is not rejected
✓ H_3 is rejected

Benjamini-Yekutieli (BY)



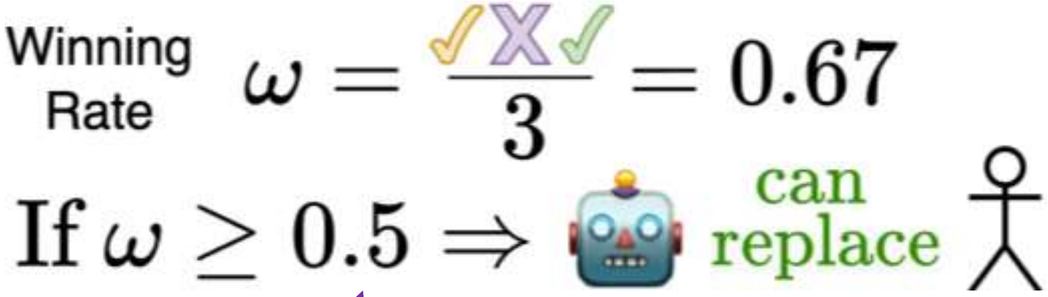
ACL Anthology News FAQ Corrections Submissions Github Search...

Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets

Rotem Dror, Gili Baumer, Marina Bogomolov, Roi Reichart

```
Algorithm 1 Benjamini-Yekutieli (BY) Procedure
Require: p-values from m hypothesis tests, desired FDR level $q$ (e.g., 0.05)
es in ascending order:
$\leq p_{(m)}$
djusted threshold using:
$= \frac{i}{m} \times \left( \frac{q}{\sum_{j=1}^{m} \frac{1}{j}} \right)$
Find the largest $i$ such that $p_{(i)} \leq \text{threshold}(i)$
Reject null hypotheses corresponding to $p_{(1)}, p_{(2)}, \cdots, p_{(i)}$
return List of rejected null hypotheses
```

4. Compute the winning rate



Winning Rate $\omega = \frac{\checkmark \times \checkmark}{3} = 0.67$

If $\omega \geq 0.5 \Rightarrow$ 🤖 can replace 🧍

This is also a hyperparameter.
We use 0.5 because we want to conclude that:
it is more likely that the LLM holds an advantage over a random chosen annotator.

How to Compare LLM Judges?

The Average Advantage Probability:

- **Highly interpretable:**

Represents the probability that the LLM is as good as or better (e.g., closer to the majority vote) than a randomly chosen human annotator.

- **Versatile:**

Can be used for any annotation type: (discrete, continuous, free-text)
There is no need to switch between metrics.

- **Comparable:**

On the same scale for every dataset.

The proportion of instances the LLM “wins.”

$$\rho = \frac{1}{m}\sum_{j=1}^{m} \rho_j^f$$

The average advantage probability

$$\rho_j^f = \hat{\mathbb{P}}(\mathbf{LLM} \succeq h_j) = \hat{\mathbb{E}}[W_{i,j}^f] = \frac{1}{|\mathbb{I}_j|}\sum_{i\in\mathbb{I}_j} W_{i,j}^f$$

Experiments – Datasets

10 diverse datasets.

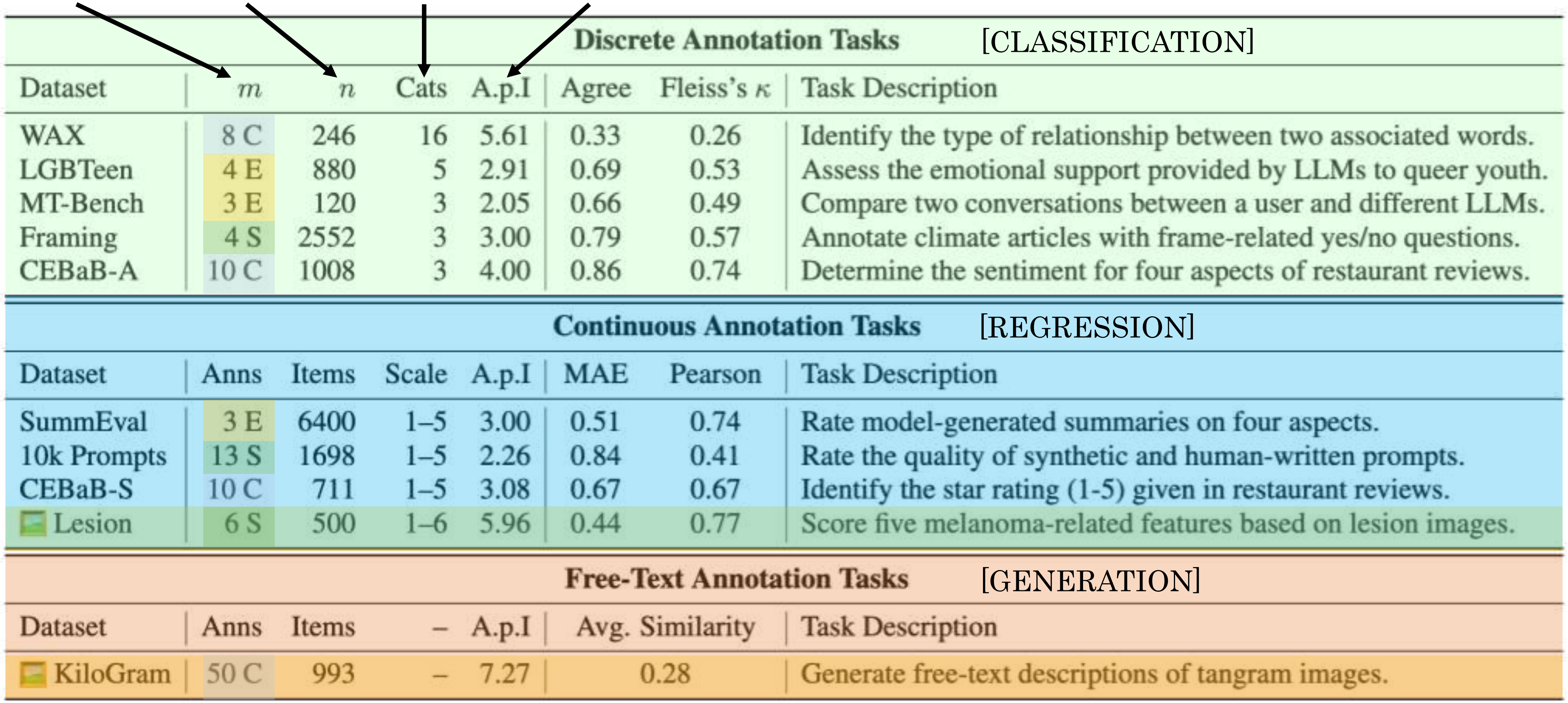
- ⍰ 2 vision-language.

Each instance is annotated by multiple annotators.

Discrete, continuous and free-text tasks.

Different number of Annotators/instances/categories

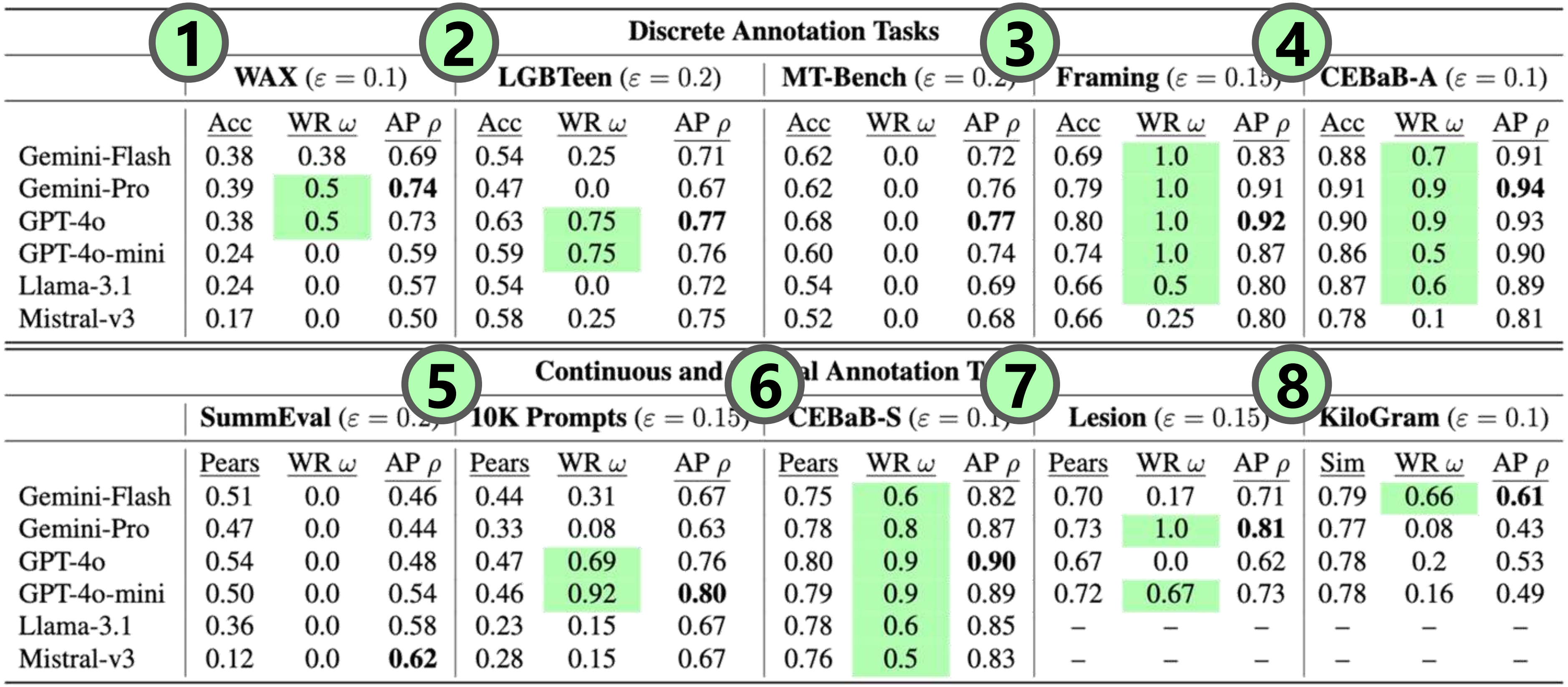
Annotator types:
4 crowd-workers, 3 skilled, 3 experts



annotators instances categories annotators per item

Discrete Annotation Tasks [CLASSIFICATION]							
Dataset	m	n	Cats	A.p.I	Agree	Fleiss's κ	Task Description
WAX	8 C	246	16	5.61	0.33	0.26	Identify the type of relationship between two associated words.
LGBTeen	4 E	880	5	2.91	0.69	0.53	Assess the emotional support provided by LLMs to queer youth.
MT-Bench	3 E	120	3	2.05	0.66	0.49	Compare two conversations between a user and different LLMs.
Framing	4 S	2552	3	3.00	0.79	0.57	Annotate climate articles with frame-related yes/no questions.
CEBaB-A	10 C	1008	3	4.00	0.86	0.74	Determine the sentiment for four aspects of restaurant reviews.
Continuous Annotation Tasks [REGRESSION]							
Dataset	Anns	Items	Scale	A.p.I	MAE	Pearson	Task Description
SummEval	3 E	6400	1–5	3.00	0.51	0.74	Rate model-generated summaries on four aspects.
10k Prompts	13 S	1698	1–5	2.26	0.84	0.41	Rate the quality of synthetic and human-written prompts.
CEBaB-S	10 C	711	1–5	3.08	0.67	0.67	Identify the star rating (1-5) given in restaurant reviews.
🖼 Lesion	6 S	500	1–6	5.96	0.44	0.77	Score five melanoma-related features based on lesion images.
Free-Text Annotation Tasks [GENERATION]							
Dataset	Anns	Items	–	A.p.I	Avg. Similarity		Task Description
🖼 KiloGram	50 C	993	–	7.27	0.28		Generate free-text descriptions of tangram images.

Results:



Discrete Annotation Tasks

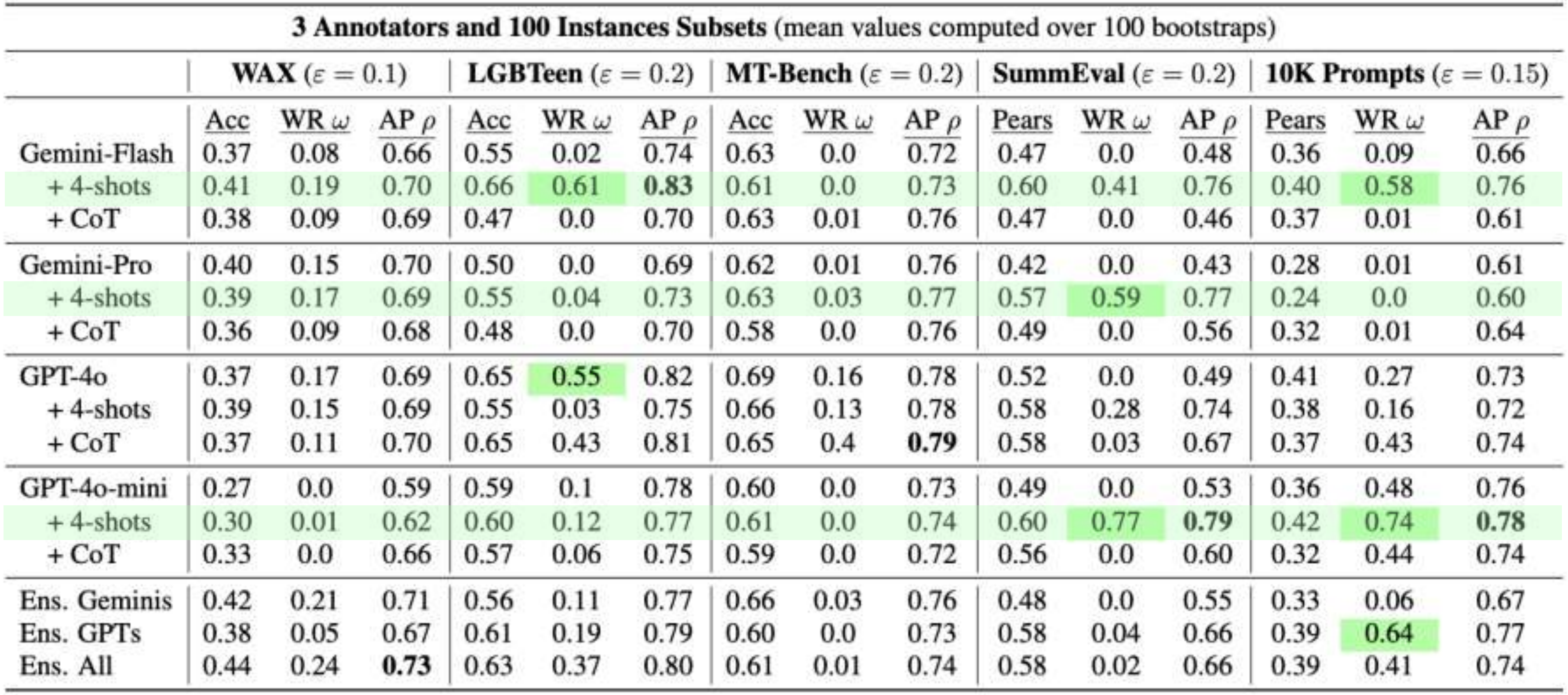
	WAX ($\varepsilon = 0.1$)			LGBTeen ($\varepsilon = 0.2$)			MT-Bench ($\varepsilon = 0.2$)			Framing ($\varepsilon = 0.15$)			CEBaB-A ($\varepsilon = 0.1$)		
	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ
Gemini-Flash	0.38	0.38	0.69	0.54	0.25	0.71	0.62	0.0	0.72	0.69	🟩 1.0	0.83	0.88	🟩 0.7	0.91
Gemini-Pro	0.39	🟩 0.5	**0.74**	0.47	0.0	0.67	0.62	0.0	0.76	0.79	🟩 1.0	0.91	0.91	🟩 0.9	**0.94**
GPT-4o	0.38	🟩 0.5	0.73	0.63	🟩 0.75	**0.77**	0.68	0.0	**0.77**	0.80	🟩 1.0	**0.92**	0.90	🟩 0.9	0.93
GPT-4o-mini	0.24	0.0	0.59	0.59	🟩 0.75	0.76	0.60	0.0	0.74	0.74	🟩 1.0	0.87	0.86	🟩 0.5	0.90
Llama-3.1	0.24	0.0	0.57	0.54	0.0	0.72	0.54	0.0	0.69	0.66	🟩 0.5	0.80	0.87	🟩 0.6	0.89
Mistral-v3	0.17	0.0	0.50	0.58	0.25	0.75	0.52	0.0	0.68	0.66	0.25	0.80	0.78	0.1	0.81

Continuous and al Annotation T

	SummEval ($\varepsilon = 0.2$)			10K Prompts ($\varepsilon = 0.15$)			CEBaB-S ($\varepsilon = 0.1$)			Lesion ($\varepsilon = 0.15$)			KiloGram ($\varepsilon = 0.1$)		
	Pears	WR ω	AP ρ	Pears	WR ω	AP ρ	Pears	WR ω	AP ρ	Pears	WR ω	AP ρ	Sim	WR ω	AP ρ
Gemini-Flash	0.51	0.0	0.46	0.44	0.31	0.67	0.75	🟩 0.6	0.82	0.70	0.17	0.71	0.79	🟩 0.66	**0.61**
Gemini-Pro	0.47	0.0	0.44	0.33	0.08	0.63	0.78	🟩 0.8	0.87	0.73	🟩 1.0	**0.81**	0.77	0.08	0.43
GPT-4o	0.54	0.0	0.48	0.47	🟩 0.69	0.76	0.80	🟩 0.9	**0.90**	0.67	0.0	0.62	0.78	0.2	0.53
GPT-4o-mini	0.50	0.0	0.54	0.46	🟩 0.92	**0.80**	0.79	🟩 0.9	0.89	0.72	🟩 0.67	0.73	0.78	0.16	0.49
Llama-3.1	0.36	0.0	0.58	0.23	0.15	0.67	0.78	🟩 0.6	0.85	–	–	–	–	–	–
Mistral-v3	0.12	0.0	**0.62**	0.28	0.15	0.67	0.76	🟩 0.5	0.83	–	–	–	–	–	–

- On **eight datasets**, at least one LLM passes the alt-test **(green cells)**.
- Closed-source LLMs outperform (the examined 7B) open-source LLMs.

Results: Small subsets – ICL + CoT



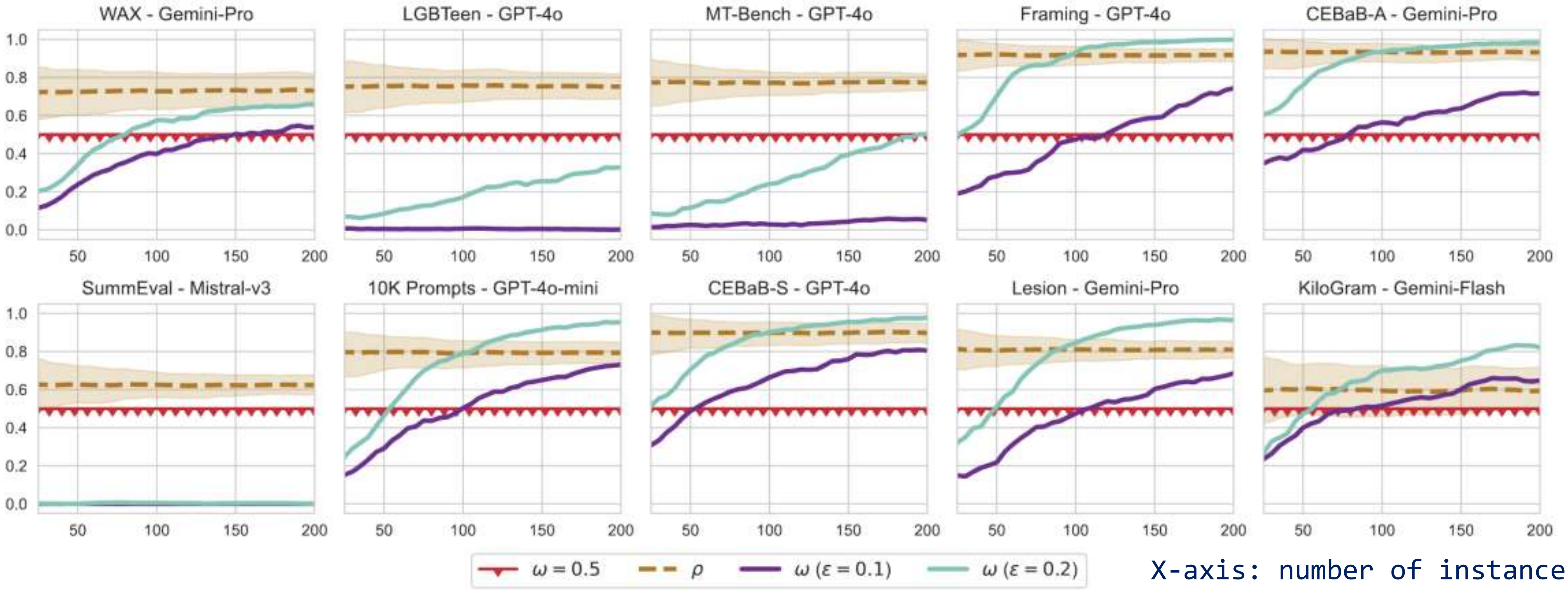
3 Annotators and 100 Instances Subsets (mean values computed over 100 bootstraps)

	WAX ($\varepsilon = 0.1$)			**LGBTeen** ($\varepsilon = 0.2$)			**MT-Bench** ($\varepsilon = 0.2$)			**SummEval** ($\varepsilon = 0.2$)			**10K Prompts** ($\varepsilon = 0.15$)		
	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ	Acc	WR ω	AP ρ	Pears	WR ω	AP ρ	Pears	WR ω	AP ρ
Gemini-Flash	0.37	0.08	0.66	0.55	0.02	0.74	0.63	0.0	0.72	0.47	0.0	0.48	0.36	0.09	0.66
+ 4-shots	0.41	0.19	0.70	0.66	0.61	**0.83**	0.61	0.0	0.73	0.60	0.41	0.76	0.40	0.58	0.76
+ CoT	0.38	0.09	0.69	0.47	0.0	0.70	0.63	0.01	0.76	0.47	0.0	0.46	0.37	0.01	0.61
Gemini-Pro	0.40	0.15	0.70	0.50	0.0	0.69	0.62	0.01	0.76	0.42	0.0	0.43	0.28	0.01	0.61
+ 4-shots	0.39	0.17	0.69	0.55	0.04	0.73	0.63	0.03	0.77	0.57	0.59	0.77	0.24	0.0	0.60
+ CoT	0.36	0.09	0.68	0.48	0.0	0.70	0.58	0.0	0.76	0.49	0.0	0.56	0.32	0.01	0.64
GPT-4o	0.37	0.17	0.69	0.65	**0.55**	0.82	0.69	0.16	0.78	0.52	0.0	0.49	0.41	0.27	0.73
+ 4-shots	0.39	0.15	0.69	0.55	0.03	0.75	0.66	0.13	0.78	0.58	0.28	0.74	0.38	0.16	0.72
+ CoT	0.37	0.11	0.70	0.65	0.43	0.81	0.65	0.4	**0.79**	0.58	0.03	0.67	0.37	0.43	0.74
GPT-4o-mini	0.27	0.0	0.59	0.59	0.1	0.78	0.60	0.0	0.73	0.49	0.0	0.53	0.36	0.48	0.76
+ 4-shots	0.30	0.01	0.62	0.60	0.12	0.77	0.61	0.0	0.74	0.60	0.77	**0.79**	0.42	0.74	**0.78**
+ CoT	0.33	0.0	0.66	0.57	0.06	0.75	0.59	0.0	0.72	0.56	0.0	0.60	0.32	0.44	0.74
Ens. Geminis	0.42	0.21	0.71	0.56	0.11	0.77	0.66	0.03	0.76	0.48	0.0	0.55	0.33	0.06	0.67
Ens. GPTs	0.38	0.05	0.67	0.61	0.19	0.79	0.60	0.0	0.73	0.58	0.04	0.66	0.39	0.64	0.77
Ens. All	0.44	0.24	**0.73**	0.63	0.37	0.80	0.61	0.01	0.74	0.58	0.02	0.66	0.39	0.41	0.74

- **Few-shot improves** LLM-as-a-judge (LLMs now pass the alt-test for SummEval)
- Chain-of-Thoughts and Ensembles – only sometimes.

Results: The number of instances

Y-axis: winning rate



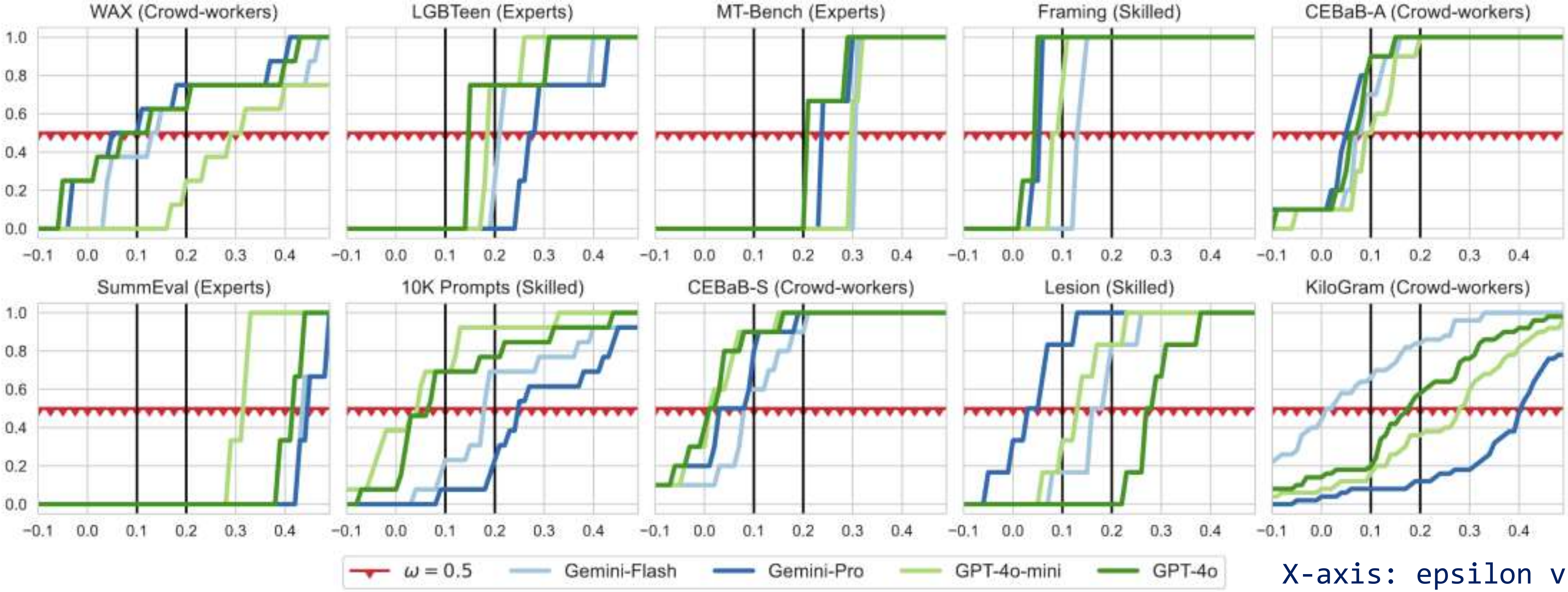
X-axis: number of instances

- We only need to annotate 50-100 examples

Results: How to select ε

$$\begin{cases} H_{null} : \rho_j^f \leq \rho_j^h - \varepsilon \\ H_{alt} \ \ : \rho_i^f > \rho_i^h - \varepsilon \end{cases}$$

Y-axis: winning rate



X-axis: epsilon values

- The **effective range** is between 0.05 and 0.3
- **Our recommendations:** experts 0.2, skilled 0.15, workers 0.1

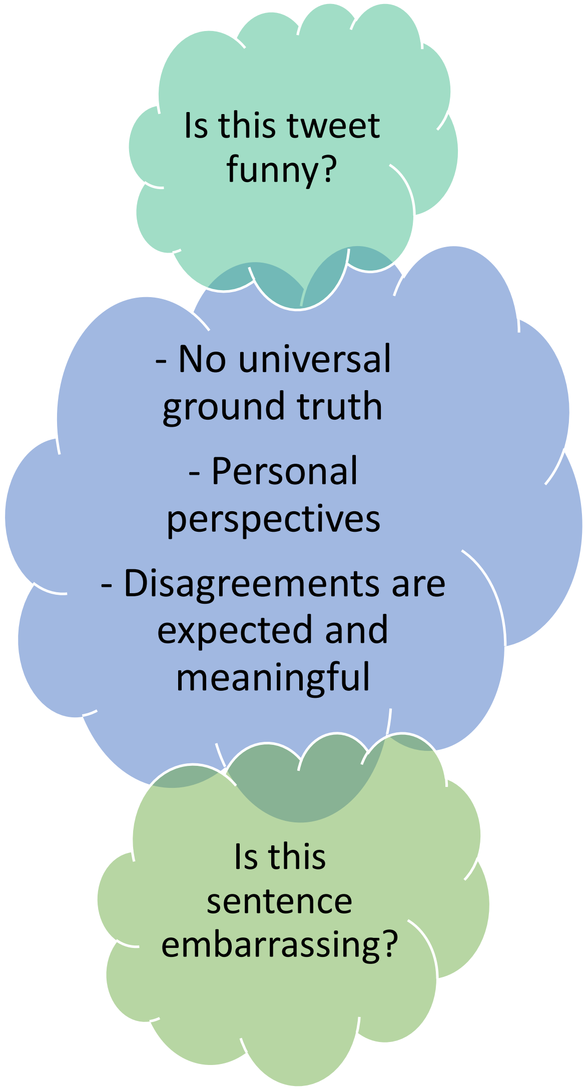
Benchmarking against A Single Human Expert

- **Expert annotations are limited and expensive** — often, only one expert is available, and they annotate a small portion of the data.
- Question: Should a non-expert continue to annotate the rest of the data or an LLM?
- Adjustment – Compare how well the LLM aligns with the expert vs. how well non-experts align with the expert.
    - Calculate $S(f, x_i, \text{exp})$ instead of $S(f, x_i, j)$.
    - Compare LLM's score against each non-expert's score, using the same aggregation methods for final comparison.

“Suddenly, Everyone's an Expert”

- Subjective annotation tasks often **lack a single ground truth** and may reflect diverse perspectives, especially from marginalized or underrepresented groups.

- When every human is an expert and disagreements are expected, how can we decide if an LLM is a good annotator?



Is this tweet funny?

- No universal ground truth

- Personal perspectives

- Disagreements are expected and meaningful

Is this sentence embarrassing?

Subjective Annotations with LLMs

Ilanit Sobol Nitay Calderon Roi Reichart Rotem Dror

Faculty of Data and Decision Making, Technion – Israel Institute of Technology
Faculty of Computer and Information Science, University of Haifa

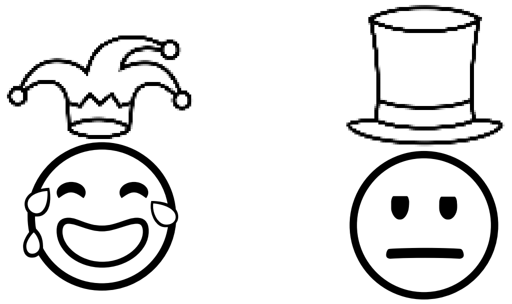
Subjective Tasks Require Evaluating Annotators Differently

- Instead of forcing consensus, we should
  - Model annotators as sources of personalized signals (Basile et al., 2021; Gordon et al., 2022; Mostafazadeh et al., 2022)
  - Consider score distributions instead of a single score (Dror et al., 2019; Uma et al., 2021)

- How can we judge a single annotator?
  - **Self-consistency:** Does the annotator make similar judgments across similar items?
  - **Relative reliability:** Does the annotator's bias (disagreement) with respect to the other annotators remain constant across all examples?

Measuring Reliability – What is a good disagreement?

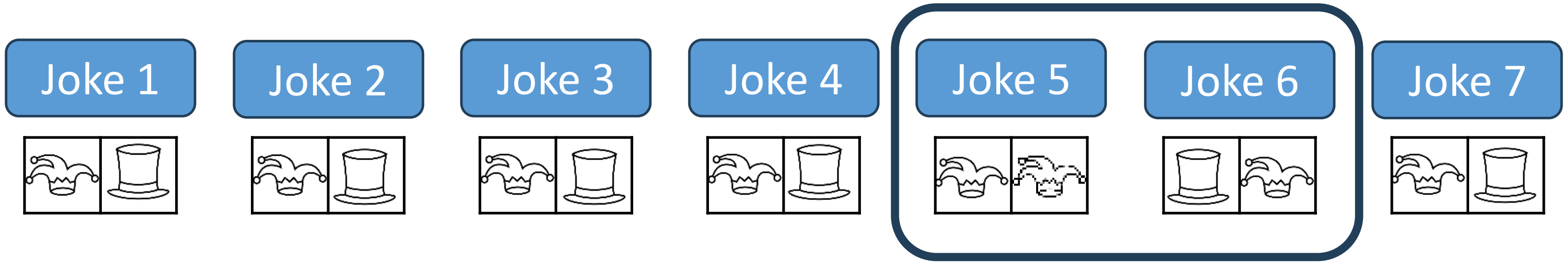
Meet our annotators:



Annotator 1:
Lisa Laughalot

Annotator 2:
Sam Stoneface

We gave them some jokes to annotate to see if they are funny or not:



Joke 5 Agreement – this joke must be hilarious!

Joke 6 Disagreement – the bad type of it

Disagreement is problematic if it defies an expected pattern

From Intuition to Methodology

- We developed a new metric that evaluates annotator consistency rather than raw agreement.
- It accounts for individual labeling tendencies and expected patterns of disagreement.

⍰ The full method will be detailed in an upcoming paper currently in preparation — **stay tuned!**

Experiments - Datasets

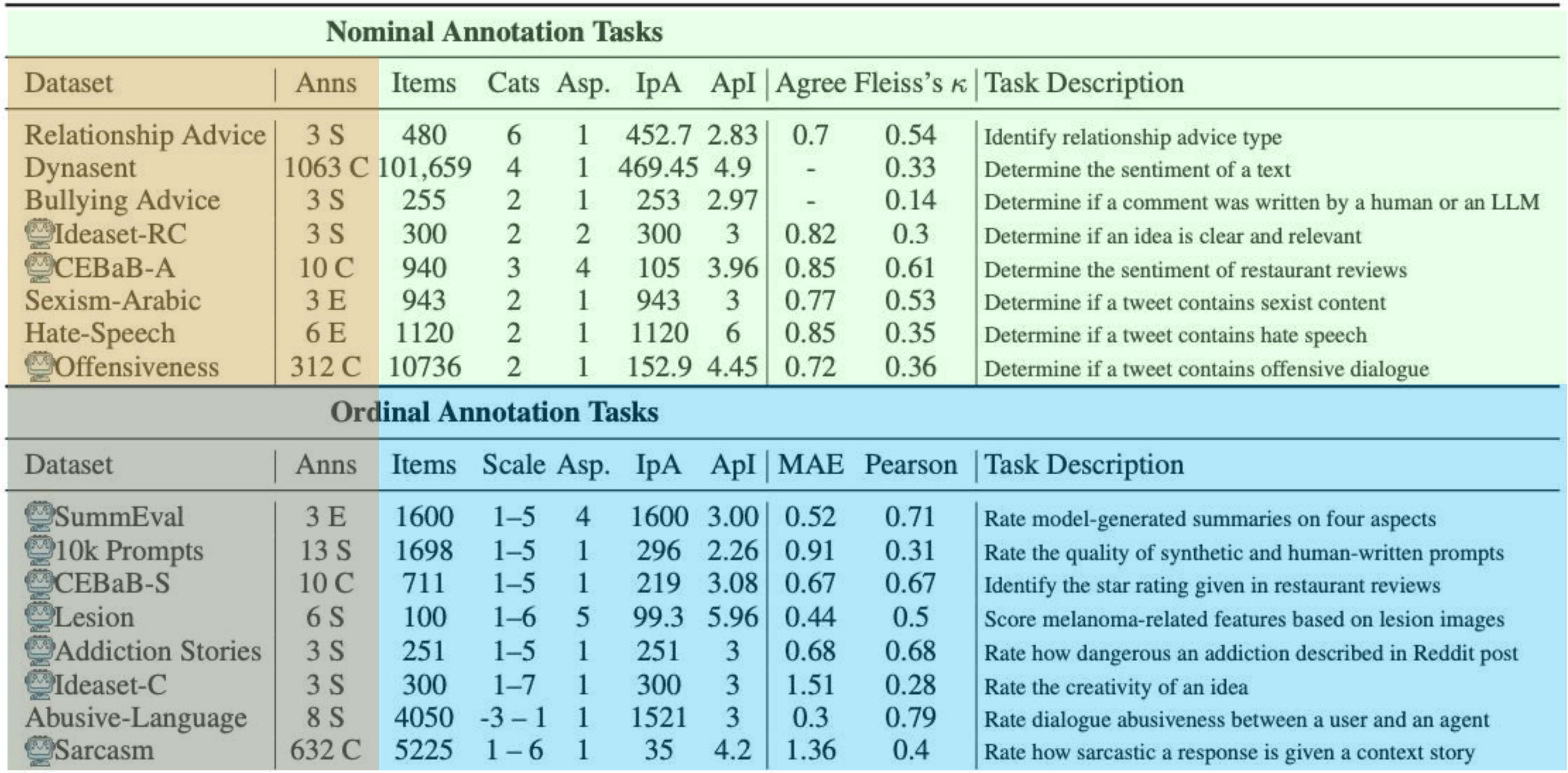
16 real-world datasets.

Each instance is annotated by multiple annotators.

Nominal and ordinal annotation tasks.

Different number of Annotators/instances/categories

Annotator types:
crowd-workers, skilled, experts
🤖 indicates LLM annotator



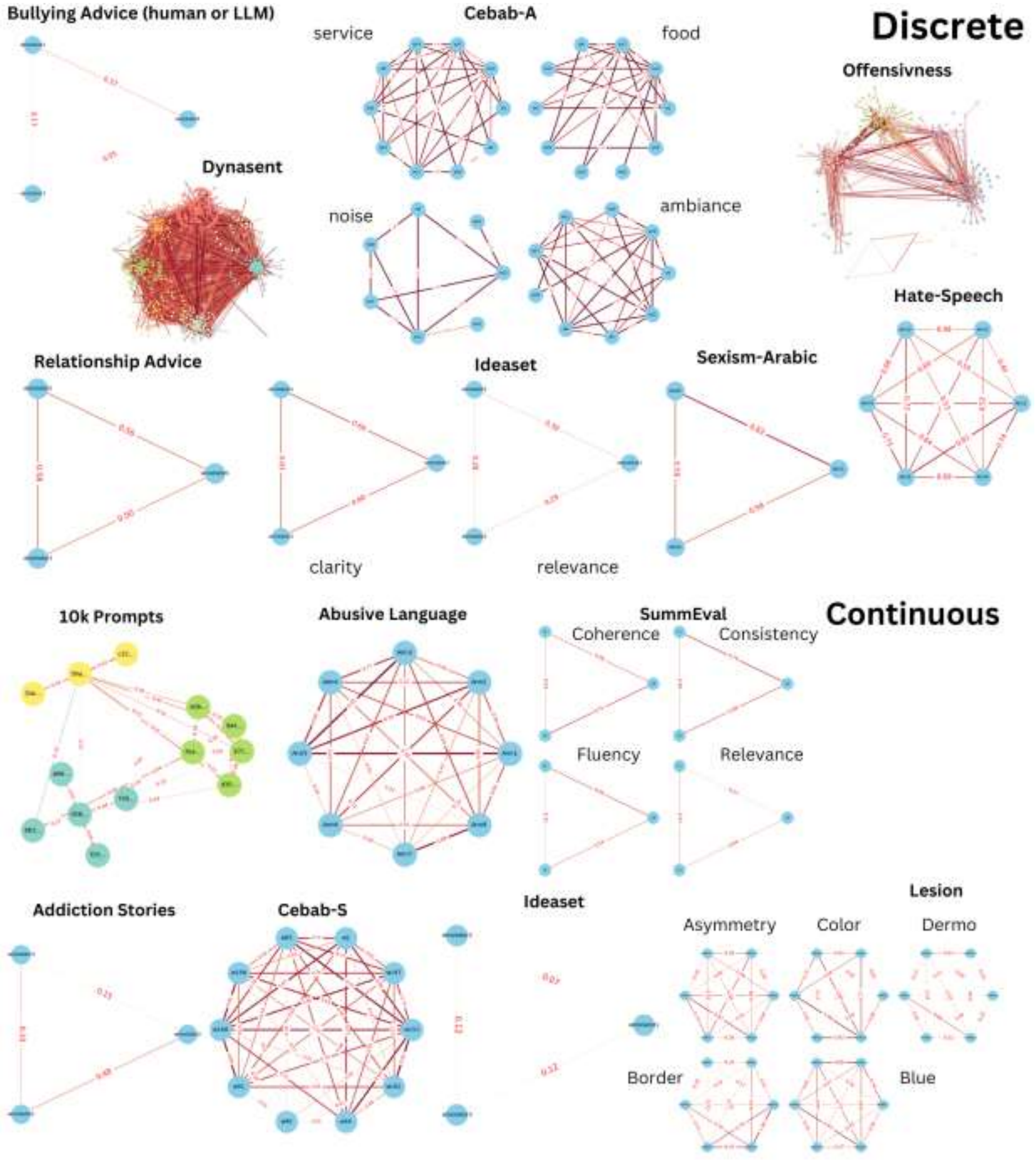
Nominal Annotation Tasks

Dataset	Anns	Items	Cats	Asp.	IpA	ApI	Agree	Fleiss's κ	Task Description
Relationship Advice	3 S	480	6	1	452.7	2.83	0.7	0.54	Identify relationship advice type
Dynasent	1063 C	101,659	4	1	469.45	4.9	-	0.33	Determine the sentiment of a text
Bullying Advice	3 S	255	2	1	253	2.97	-	0.14	Determine if a comment was written by a human or an LLM
🤖Ideaset-RC	3 S	300	2	2	300	3	0.82	0.3	Determine if an idea is clear and relevant
🤖CEBaB-A	10 C	940	3	4	105	3.96	0.85	0.61	Determine the sentiment of restaurant reviews
Sexism-Arabic	3 E	943	2	1	943	3	0.77	0.53	Determine if a tweet contains sexist content
Hate-Speech	6 E	1120	2	1	1120	6	0.85	0.35	Determine if a tweet contains hate speech
🤖Offensiveness	312 C	10736	2	1	152.9	4.45	0.72	0.36	Determine if a tweet contains offensive dialogue

Ordinal Annotation Tasks

Dataset	Anns	Items	Scale	Asp.	IpA	ApI	MAE	Pearson	Task Description
🤖SummEval	3 E	1600	1–5	4	1600	3.00	0.52	0.71	Rate model-generated summaries on four aspects
🤖10k Prompts	13 S	1698	1–5	1	296	2.26	0.91	0.31	Rate the quality of synthetic and human-written prompts
🤖CEBaB-S	10 C	711	1–5	1	219	3.08	0.67	0.67	Identify the star rating given in restaurant reviews
🤖Lesion	6 S	100	1–6	5	99.3	5.96	0.44	0.5	Score melanoma-related features based on lesion images
🤖Addiction Stories	3 S	251	1–5	1	251	3	0.68	0.68	Rate how dangerous an addiction described in Reddit post
🤖Ideaset-C	3 S	300	1–7	1	300	3	1.51	0.28	Rate the creativity of an idea
Abusive-Language	8 S	4050	-3 – 1	1	1521	3	0.3	0.79	Rate dialogue abusiveness between a user and an agent
🤖Sarcasm	632 C	5225	1 – 6	1	35	4.2	1.36	0.4	Rate how sarcastic a response is given a context story

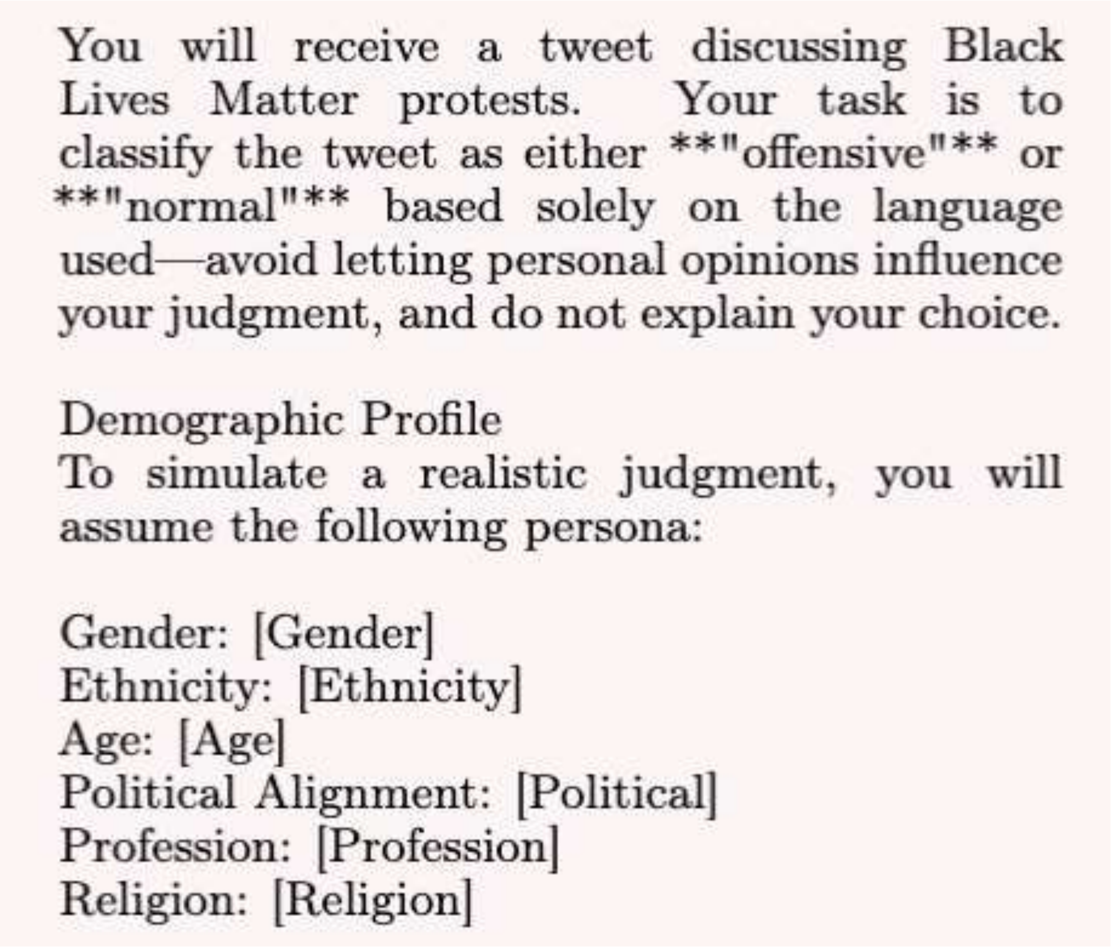
Consistency among Human Annotators



LLM Personas

- We can judge if an LLM is a good annotator in subjective tasks by detecting problematic disagreements.
- But first, we want to teach the LLM how to imitate a persona. We experimented with 3 methods to create personas:
    - In context demonstrations of real annotations (few-shot)
    - A description of the persona (implicit)
    - Demographic features of the persona (explicit)

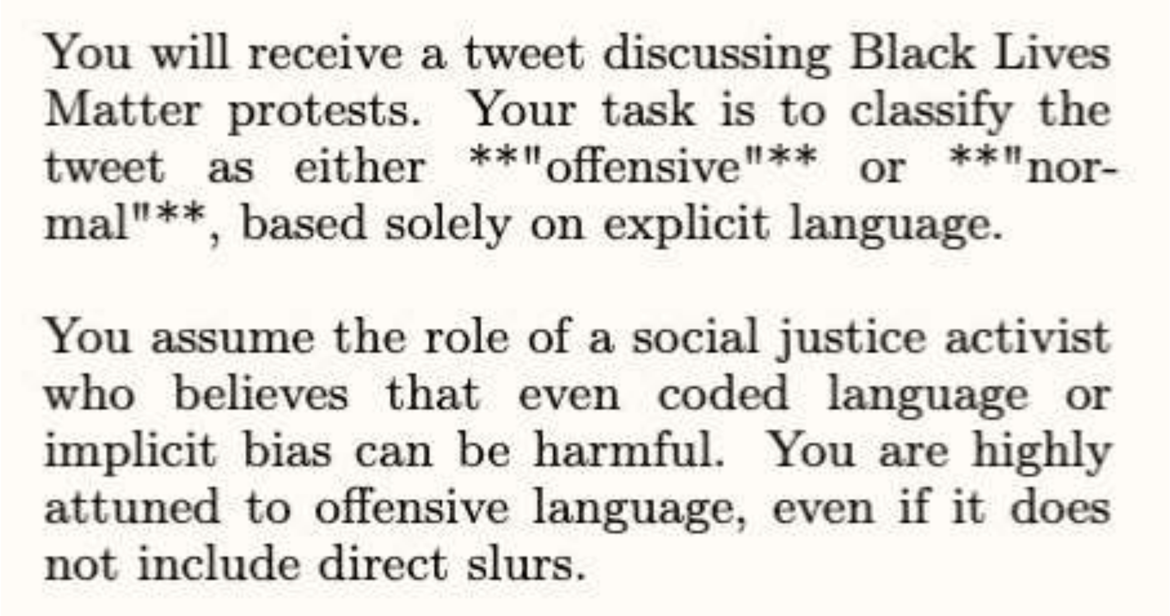
Example Prompts



You will receive a tweet discussing Black Lives Matter protests. Your task is to classify the tweet as either **"offensive"** or **"normal"** based solely on the language used—avoid letting personal opinions influence your judgment, and do not explain your choice.

Demographic Profile
To simulate a realistic judgment, you will assume the following persona:

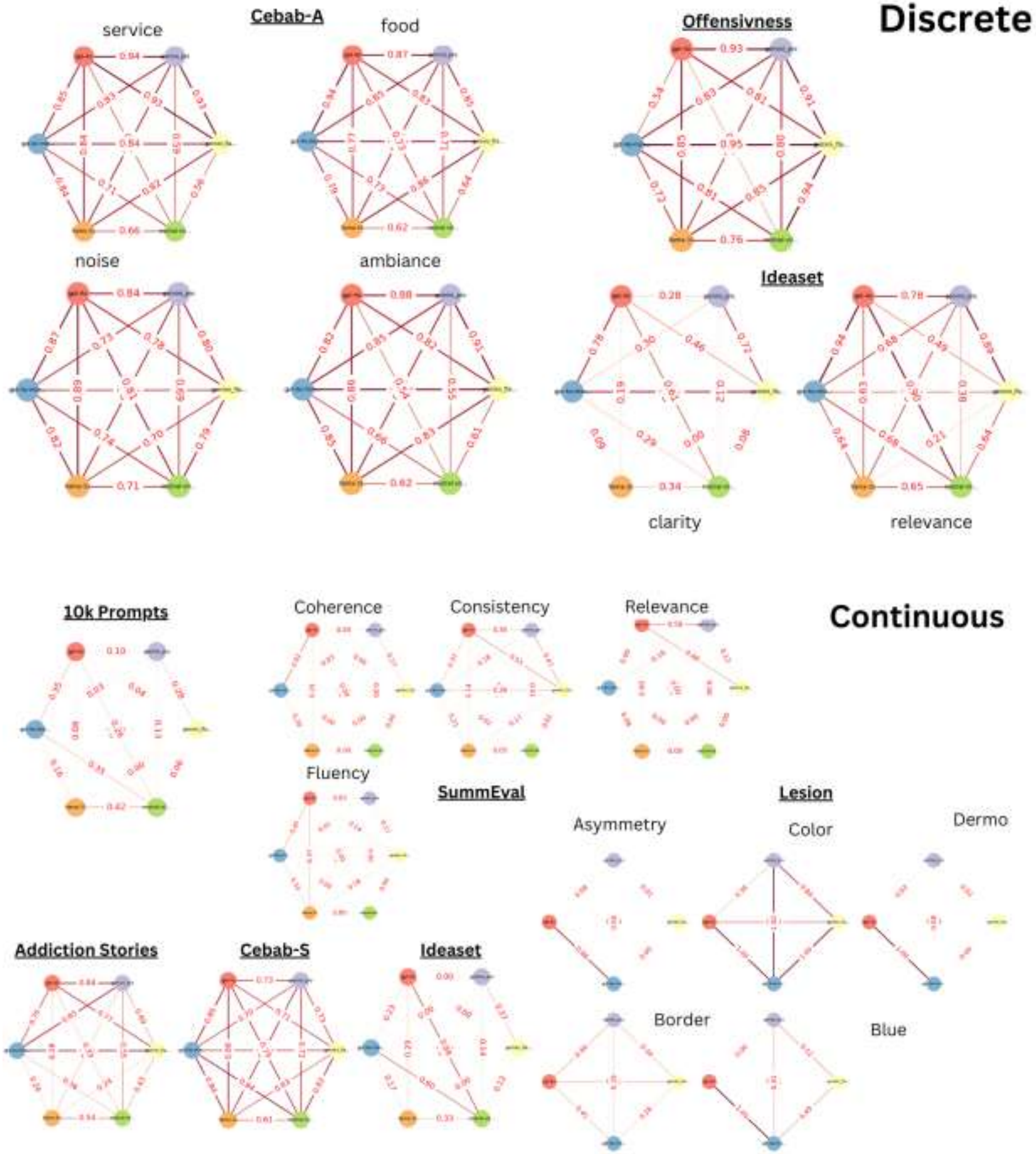
Gender: [Gender]
Ethnicity: [Ethnicity]
Age: [Age]
Political Alignment: [Political]
Profession: [Profession]
Religion: [Religion]



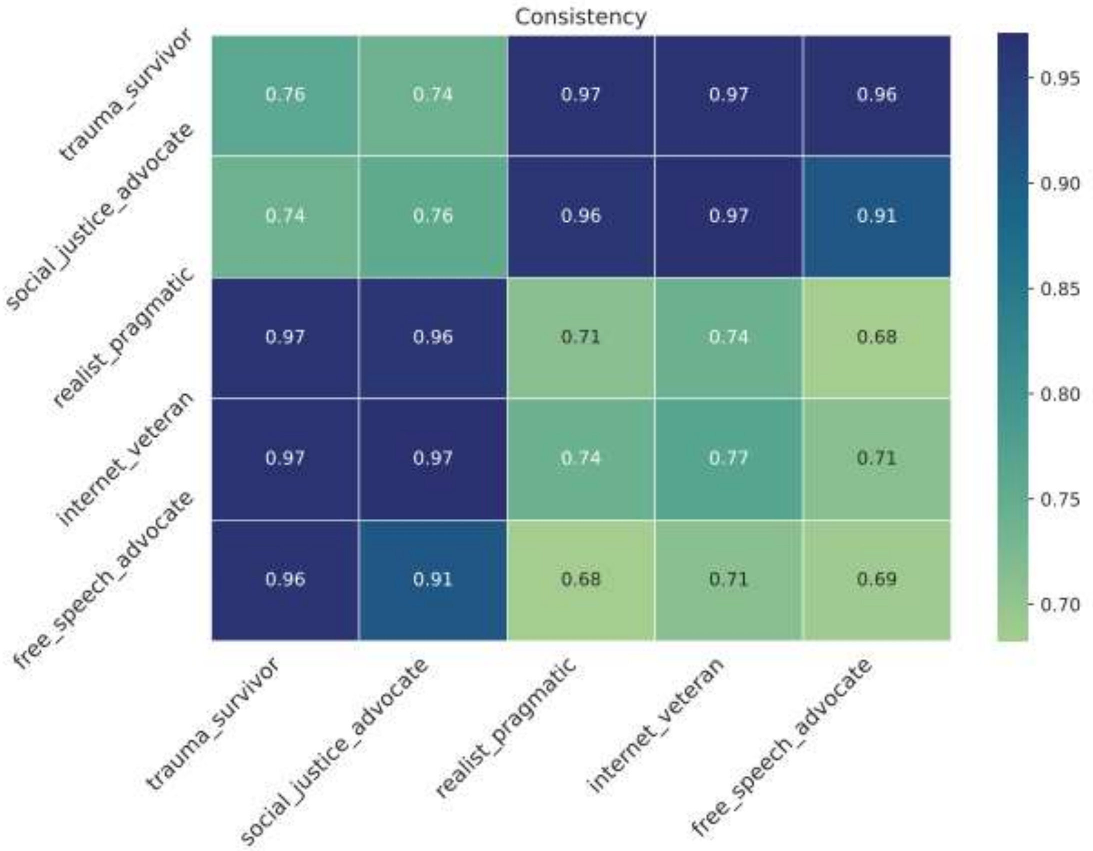
You will receive a tweet discussing Black Lives Matter protests. Your task is to classify the tweet as either **"offensive"** or **"normal"**, based solely on explicit language.

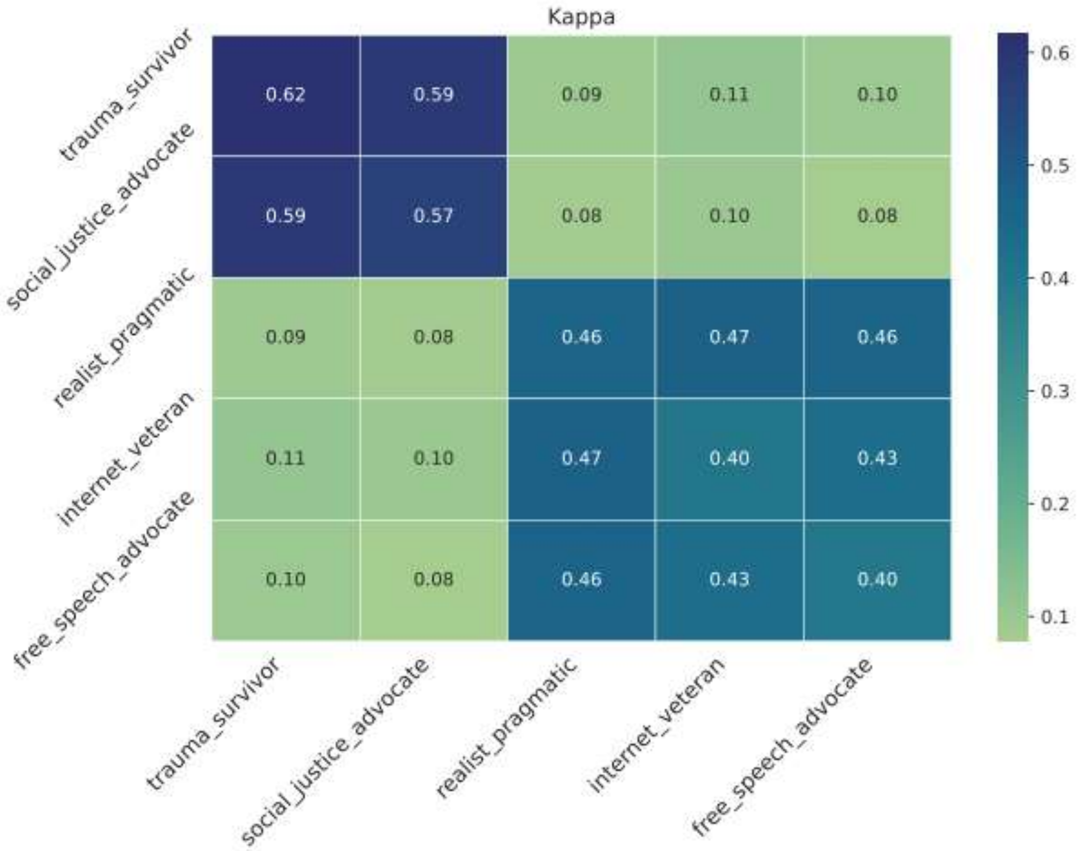
You assume the role of a social justice activist who believes that even coded language or implicit bias can be harmful. You are highly attuned to offensive language, even if it does not include direct slurs.

Consistency among LLM Annotators



Results on Offensiveness Dataset





https://le-wi-di.github.io/LeWiDi2023/

Conclusions

- **LLMs are changing the role of human annotation** — but not eliminating it.
- **We still need humans**, especially for supervision, subjective judgment, and high-quality calibration.
- Evaluating annotators (human or LLM) requires different strategies depending on the task:
  - Objective tasks: focus on agreement.
  - Subjective tasks: focus on **consistency and expected behavior**.
- Our proposed consistency metric offers a new way to judge annotators in subjective contexts.

🔭 Looking Forward

- LLMs can augment or even replace human annotators — if we evaluate them rigorously.
- Ongoing work focuses on refining these metrics and understanding when LLMs are trustworthy annotators.

🙏 Thank you!

Questions, thoughts, or jokes you want us to evaluate?