GESIS Leibniz Institute for the Social Sciences

Who and What?

Using Linguistic Features and Annotator Characteristics to Analyze Annotation Variation

Leibniz Association

Maximilian Maurer

GESIS – Leibniz Institute for the Social Sciences
Cologne & Heinrich Heine University Düsseldorf

03.06.2026

A couple of words about me

- 3rd year doctoral researcher at GESIS & HHU

A couple of words about me

- 3rd year doctoral researcher at GESIS & HHU
- Pet peeve for interpretability and efficient tools

elfen:
A Python Package for Efficient Linguistic Feature Extraction for Natural Language Datasets

A couple of words about me

- 3rd year doctoral researcher at GESIS & HHU
- Pet peeve for interpretability and efficient tools
- Interested in variation within *subjective* NLP
    - Stylistic, semantic, pragmatic, socio-linguistic

A couple of words about me

- 3rd year doctoral researcher at GESIS & HHU
- Pet peeve for interpretability and efficient tools
- Interested in variation within *subjective* NLP
  - Stylistic, semantic, pragmatic, socio-linguistic
  - **Production (human-written and model-generated)**, perception (annotation disagreement!)

A Systematic Analysis of Linguistic Features in AI-Generated Text Detection Across Domains and Models

Anonymous ACL submission

Promptology: A Large-Scale Systematic Analysis of Prompt Elements and Contextual Factors in Argument Generation

Anonymous ACL submission

Comparing Human and AI Deception in Online Reviews

Anonymous ACL submission

AI Argues Differently: Distinct Argumentative and Linguistic Patterns of LLMs in Persuasive Contexts

Esra Dönmez[♦1,2], Maximilian Maurer[★3,4], Gabriella Lapesa[3,4], Agnieszka Falenska[1,2]

A couple of words about me

- 3rd year doctoral researcher at GESIS & HHU
- Pet peeve for interpretability and efficient tools
- Interested in variation within *subjective* NLP
  - Stylistic, semantic, pragmatic, socio-linguistic
  - Production (human-written and model-generated),
    perception
    (annotation disagreement!)

GESIS-DSM at PerpectiveArg2024:
A Matter of Style? Socio-Cultural Differences in Argumentation

Maximilian Martin Maurer[1], Julia Romberg[1], Myrthe Reuver[3♥],
Negash Desalegn Weldekiros[1,4], and Gabriella Lapesa[1,2]

Towards a Perspectivist Turn in Argument Quality Assessment

Julia Romberg[1], Maximilian Maurer[1,3], Henning Wachsmuth[2] and Gabriella Lapesa[1,3]

Who and What?
Using Linguistic Features and Annotator Characteristics to Analyze Annotation Variation

Maximilian Maurer[1,2], Maximilian Linde[1] and Gabriella Lapesa[1,2]

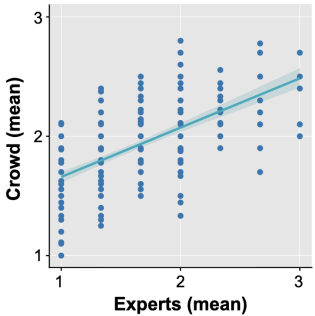
Who and What?
Using Linguistic Features and Annotator Characteristics to Analyze Annotation Variation

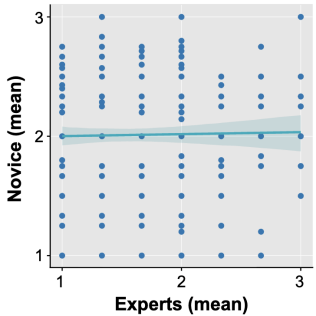
Maximilian Maurer[1,2], Maximilian Linde[1] and Gabriella Lapesa[1,2]

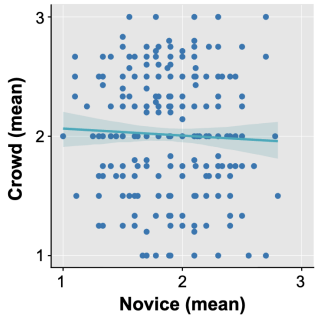
Why (re-)analyze annotation behavior?

Why (re-)analyze annotation behavior?

1. Substantial differences between annotators/groups



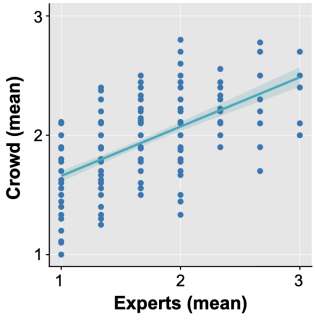


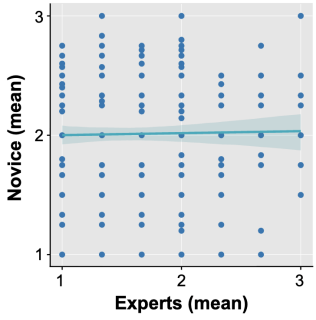


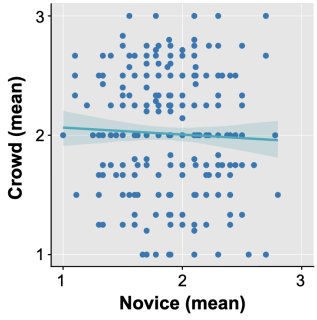
Romberg et al. (2025)

Why (re-)analyze annotation behavior?

1. Substantial differences between annotators/groups





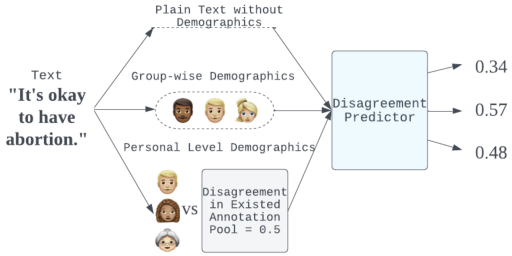


Romberg et al. (2025)

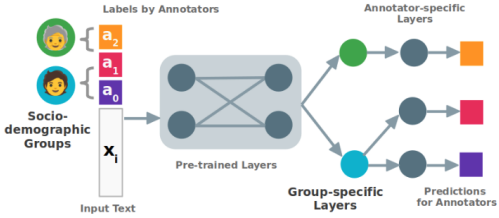
→ Where does this variation come from?

Why (re-)analyze annotation behavior?

2. Unclear whether demographics help



Wan et al. (2023): Yes



Orlikowski et al. (2023): No

Why (re-)analyze annotation behavior?

2. Unclear whether demographics help

Data, Task, Annotator population?

What about intersectional identities and lived experiences?

Do (L)LMs even capture identities interacting with text perception?

Why (re-)analyze annotation behavior?

2. Unclear whether demographics help

Data, Task, Annotator population?

What about intersectional identities and lived experiences?

Do (L)LMs even capture identities interacting with text perception?

Why (re-)analyze annotation behavior?

2. Unclear whether demographics help

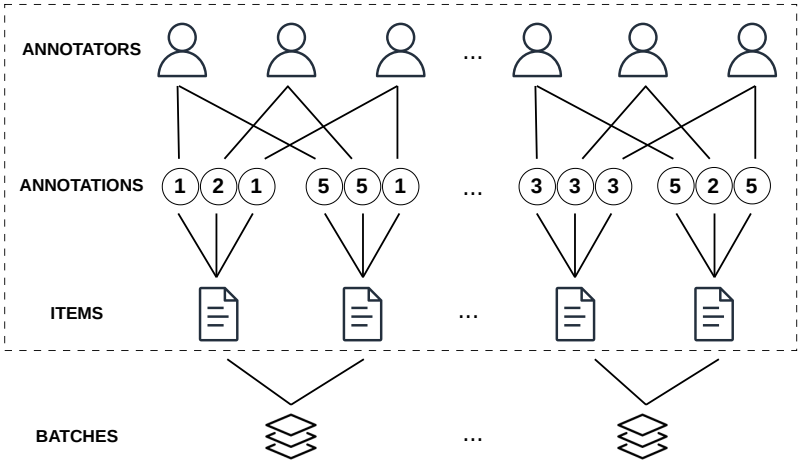
Data, Task, Annotator population?

What about intersectional identities and lived experiences?

Do (L)LMs even capture identities interacting with text perception?

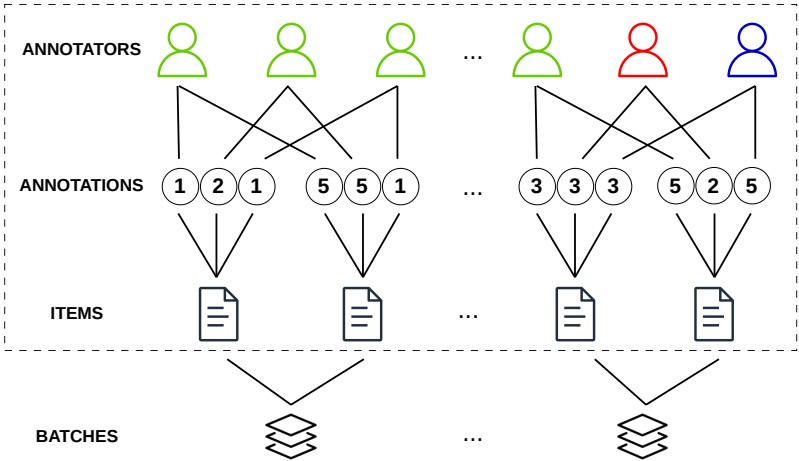
Why (re-)analyze annotation behavior?

3. Annotation setups have a special structure



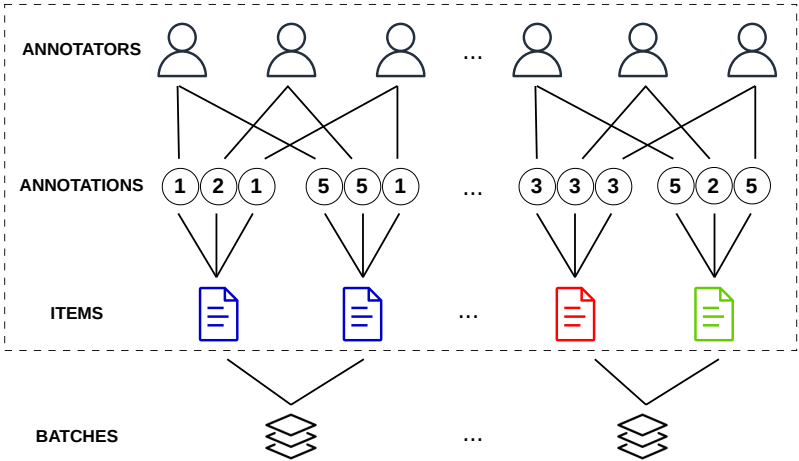
Why (re-)analyze annotation behavior?

3. Annotation setups have a special structure



Why (re-)analyze annotation behavior?

3. Annotation setups have a special structure



Goal

Model annotation as a function of the who and the what, and their interaction in different scenarios:

Goal

Model annotation as a function of the who and the what, and their interaction in different scenarios:

1. Comparing related tasks

Goal

Model annotation as a function of the who and the what, and their interaction in different scenarios:

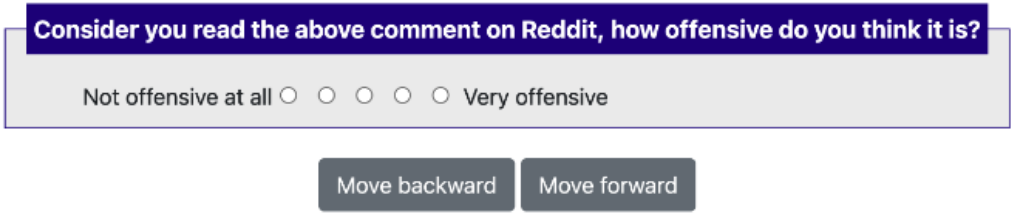
1. Comparing related tasks
2. Comparing different annotator groups on the same items

Goal

Model annotation as a function of the who and the what, and their interaction in different scenarios:

1. Comparing related tasks
2. Comparing different annotator groups on the same items
3. Comparing batches for the same annotation guidelines

Task


Consider you read the above comment on Reddit, how offensive do you think it is?
Not offensive at all
Very offensive
Move backward
Move forward

Setup

Which category below includes your age? [Radio]
o Under 18
o 18 - 24
o 25 - 34
o 35 - 44
o 45 - 54
o 55 - 64
o 65 or older
o Prefer not to say

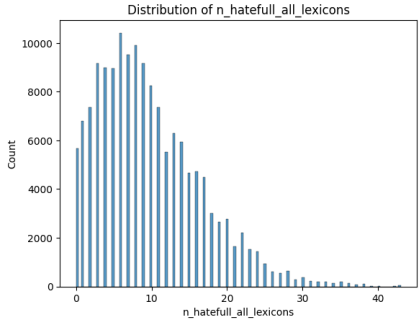
Race [Checkbox]
o White
o Hispanic or Latino
o Black or African American
o Native American or American Indian
o Asian / Pacific Islander
o Other [open ended]
o Prefer not to say

- Look at attributes of **who**, what, and interactions
  - Annotator characteristics from existing datasets

Setup

Which category below includes your age? [Radio]
o Under 18
o 18 - 24
o 25 - 34
o 35 - 44
o 45 - 54
o 55 - 64
o 65 or older
o Prefer not to say

Race [Checkbox]
o White
o Hispanic or Latino
o Black or African American
o Native American or American Indian
o Asian / Pacific Islander
o Other [open ended]
o Prefer not to say

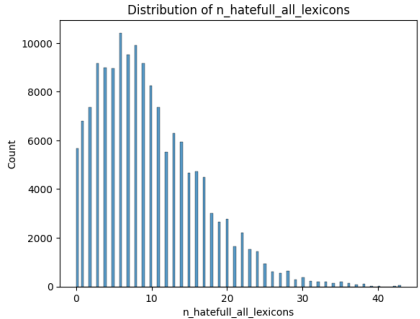


- Look at attributes of who, **what**, and interactions
  - Annotator characteristics from existing datasets
  - ($>$ 300) linguistic features from `elfen` + domain lexicons
    $\rightarrow$ interpretable, grounded

Setup

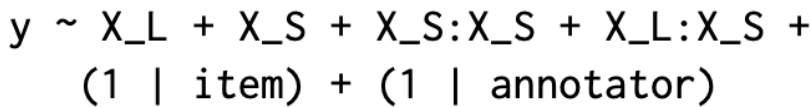
Which category below includes your age? [Radio]
o Under 18
o 18 - 24
o 25 - 34
o 35 - 44
o 45 - 54
o 55 - 64
o 65 or older
o Prefer not to say

Race [Checkbox]
o White
o Hispanic or Latino
o Black or African American
o Native American or American Indian
o Asian / Pacific Islander
o Other [open ended]
o Prefer not to say



- Look at attributes of who, what, and **interactions**
  - Annotator characteristics from existing datasets
  - ($>$ 300) linguistic features from `elfen` + domain lexicons
    $\rightarrow$ interpretable, grounded

Setup

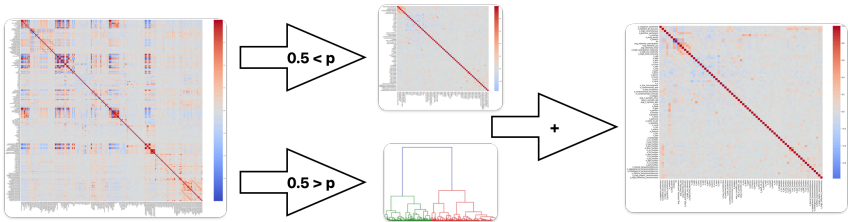


- Take into account the structure
  - Bayesian multilevel regression
  - Random intercepts for annotators and items

Setup

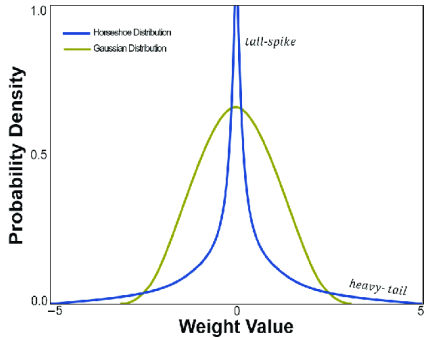
- Keep it manageable (evidence, interpretability, computationally)

Setup



- Keep it manageable (evidence, interpretability, computationally)
  - **Linguistic feature preselection**

Setup



- Keep it manageable (evidence, interpretability, computationally)
  - Linguistic feature preselection
  - **Strong regularization (Horseshoe prior)**

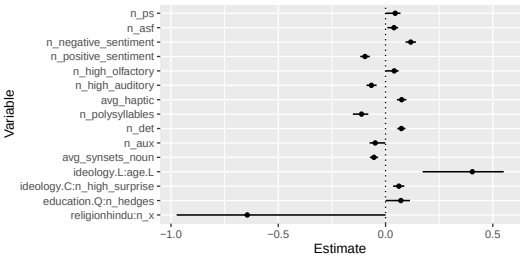
Comparing Related Tasks

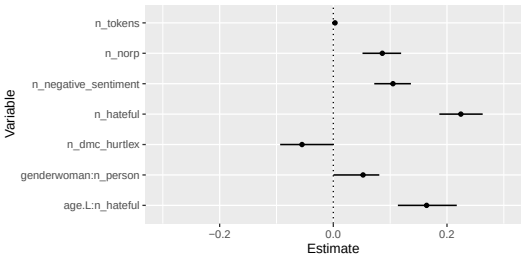
Results: Comparing Related Tasks

Dataset	Task	Items	Annotations	Annotators	Ann. per Item
MHS	Hatefulness	3,556	17,693	1,385	4.0±0.2
POPQUORN	Offensiveness	1,500	13,036	262	8.7±1.0

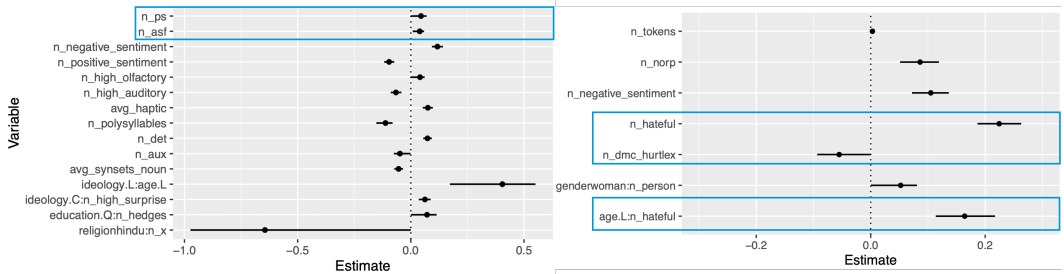
Can we expect similar effects for related tasks?

Results: Comparing Related Tasks



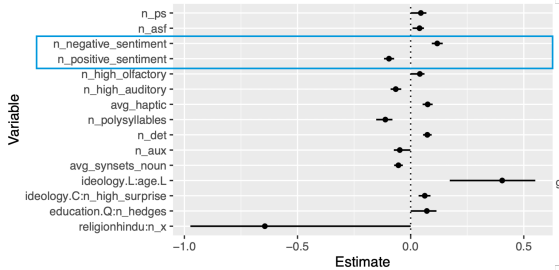


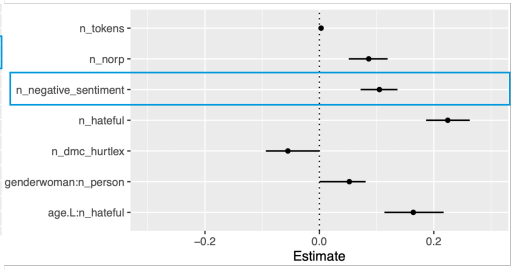
Results: Comparing Related Tasks



→ **Slurs and vulgar terms generally increase the offensiveness/hatefulness ratings**

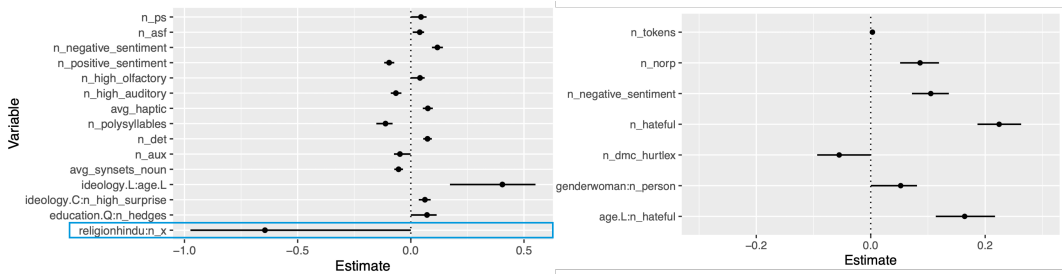
Results: Comparing Related Tasks





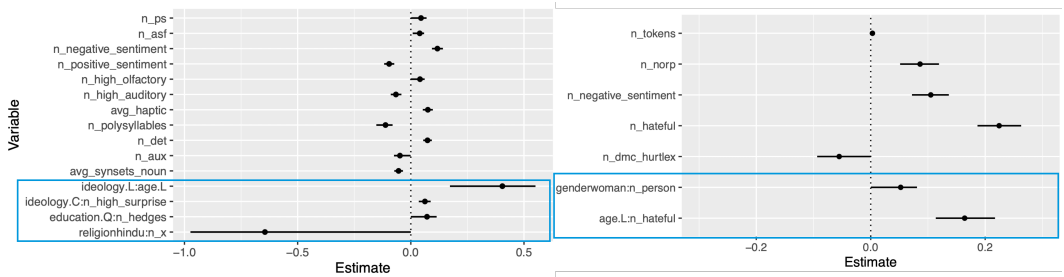
→ **Sentiment points in expected directions**

Results: Comparing Related Tasks



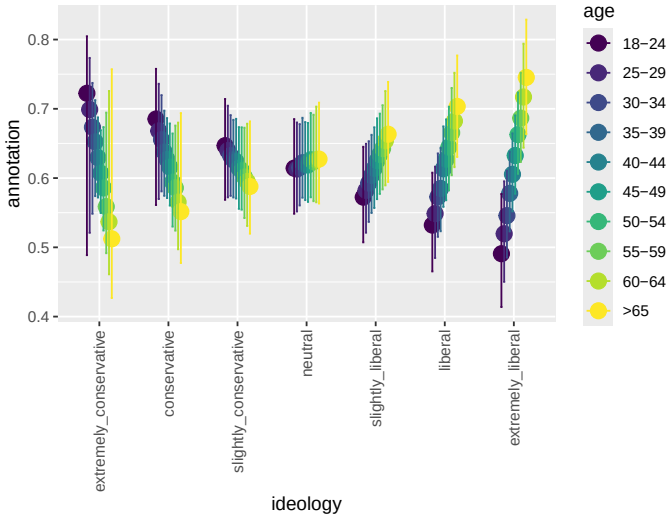
→ **We find spurious effects**

Results: Comparing Related Tasks

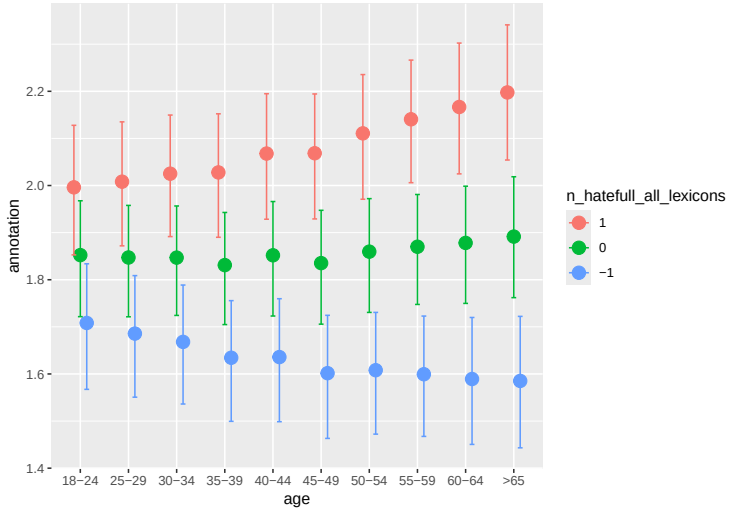


→ **Interactions and intersectional identities play a role!**

Results: Comparing Related Tasks



Results: Comparing Related Tasks



Bigger Picture: Comparing Related Tasks

- Similar tendencies on expected axes (domain-specific, sentiment)
- Considerable variation on more general features

Bigger Picture: Comparing Related Tasks

- Similar tendencies on expected axes (domain-specific, sentiment)
- Considerable variation on more general features
- Individual demographics proxies play less of a role

Bigger Picture: Comparing Related Tasks

- Similar tendencies on expected axes (domain-specific, sentiment)
- Considerable variation on more general features
- Individual demographics proxies play less of a role
- interactions and intersectional identities may to different extents

Simulating New Annotators

Simulating New Annotators

Do effects found for an annotator population map to another demographically similar one on the same items?

Simulating New Annotators

Dataset	Task	Items	Annotations	Annotators	Ann. per Item
D3CODE	Offensiveness	4,402	139,379	4,309	31.7±16.6

- Diverse annotator pool (countries, SES, demographics, moral foundations)

Simulating New Annotators

Dataset	Task	Items	Annotations	Annotators	Ann. per Item
D3CODE	Offensiveness	4,402	139,379	4,309	31.7±16.6

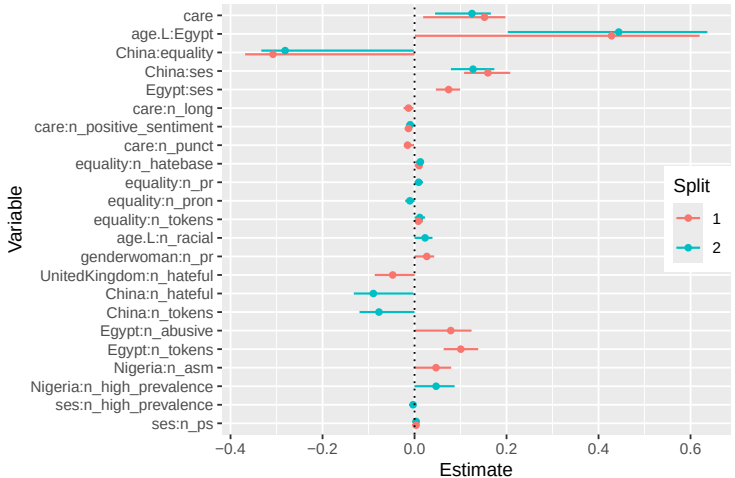
- Diverse annotator pool (countries, SES, demographics, moral foundations)
- Split annotator set per item into halves
- Ensure overall annotator characteristics distributions

Simulating New Annotators

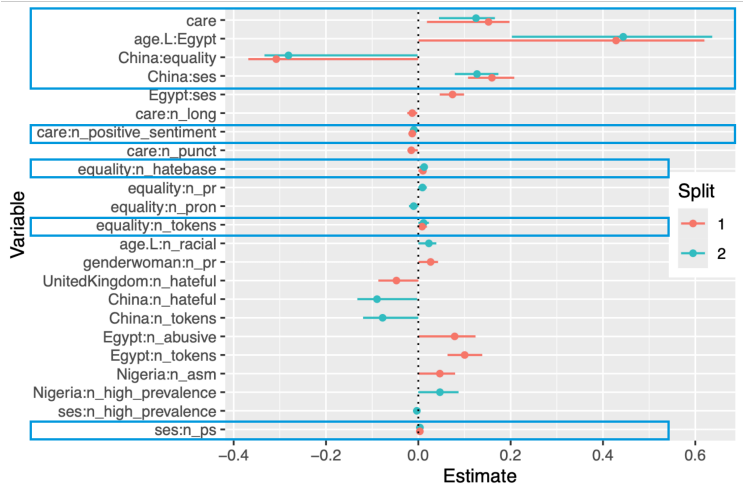
Dataset	Task	Items	Annotations	Annotators	Ann. per Item
D3CODE	Offensiveness	4,402	139,379	4,309	31.7±16.6

- Diverse annotator pool (countries, SES, demographics, moral foundations)
- Split annotator set per item into halves
- Ensure overall annotator characteristics distributions
- Fit a model per split $\rightarrow$ compare

Results: Simulating New Annotators

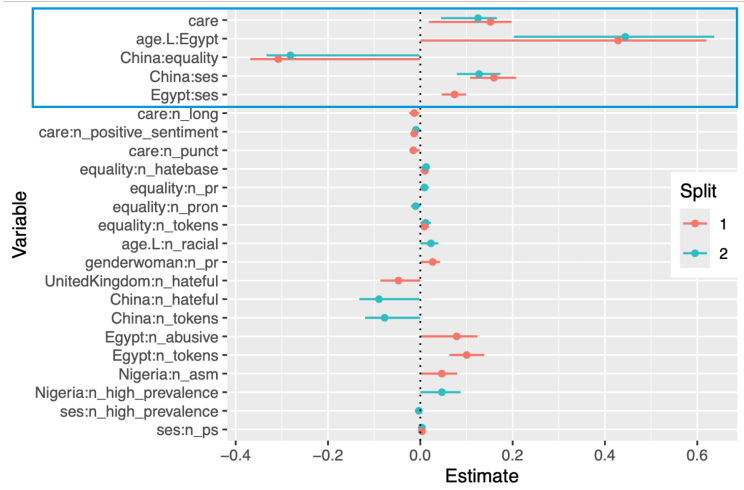


Results: Simulating New Annotators



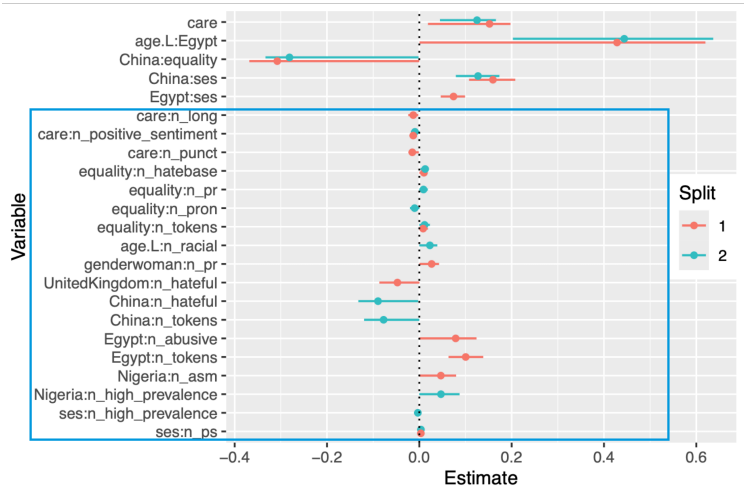
$\rightarrow$ **Only $\frac{8}{23}$ surviving effects are shared across splits**

Results: Simulating New Annotators



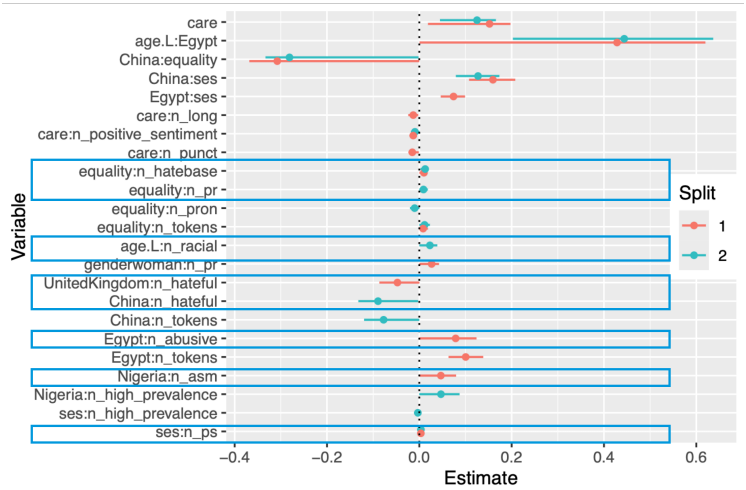
→ **(Intersectional) characteristics play a role**

Results: Simulating New Annotators



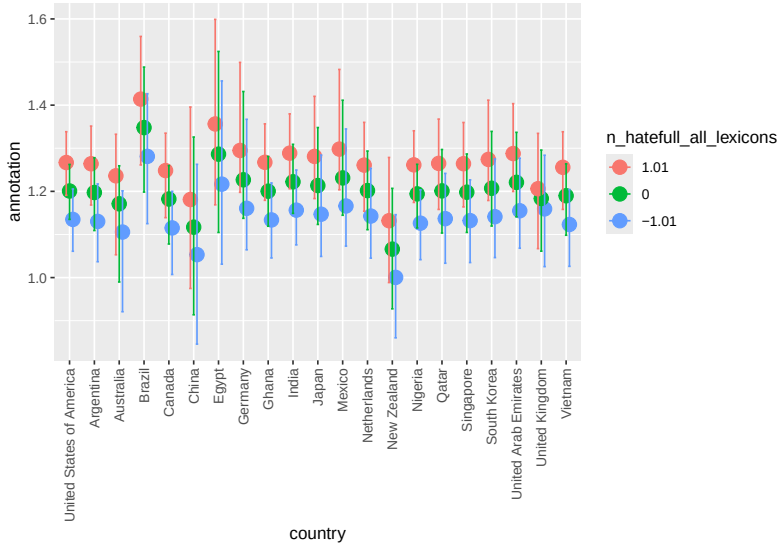
→ **Interactions play a role** (but some only for half of the annotators)

Results: Simulating New Annotators



→ **Interactions of annotator characteristics with slurs and vulgar terms**

Results: Simulating New Annotators



Bigger Picture: New Annotators

- We can expect some aligned labeling behavior
- But: considerable level of variation between similar annotator sets
  $\rightarrow$ individual differences? item-specific attitudes?

Simulating new batches

Simulating New Batches

Dataset	Task	Items	Annotations	Annotators	Ann. per Item
CTDP	Toxicity	97,489	221,087	10,958	4.7 ± 0.6

- Lived experiences, item-specific attitudes, task attitudes, demographics

Simulating New Batches

Dataset	Task	Items	Annotations	Annotators	Ann. per Item
CTDP	Toxicity	97,489	221,087	10,958	4.7 ± 0.6

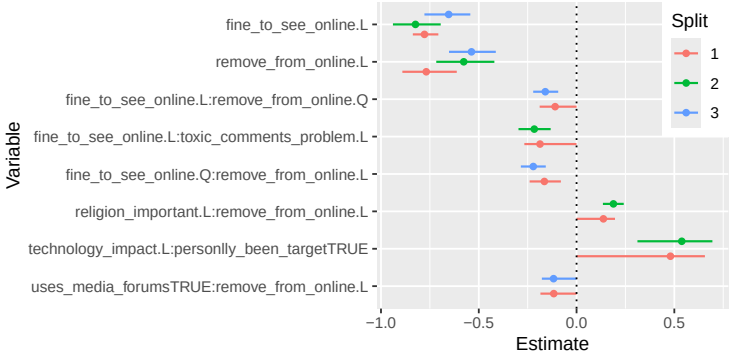
- Lived experiences, item-specific attitudes, task attitudes, demographics
- Reconstruct batches
- Subsets of 300 batches each (20 items $\times$ 5 annotators $\times$ 300 batches)

Simulating New Batches

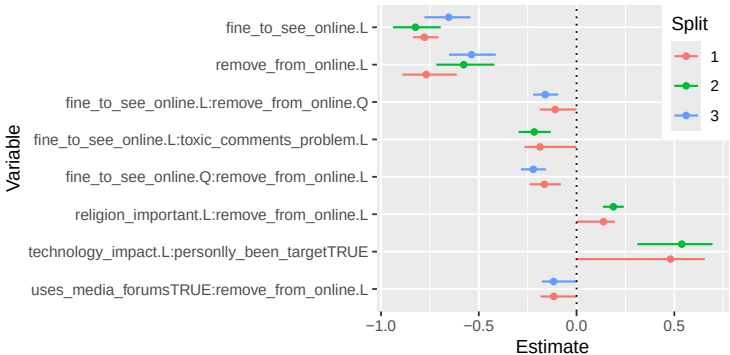
Dataset	Task	Items	Annotations	Annotators	Ann. per Item
CTDP	Toxicity	97,489	221,087	10,958	4.7 ± 0.6

- Lived experiences, item-specific attitudes, task attitudes, demographics
- Reconstruct batches
- Subsets of 300 batches each (20 items $\times$ 5 annotators $\times$ 300 batches)
- Fit models for three subsets

Results: Simulating New Batches

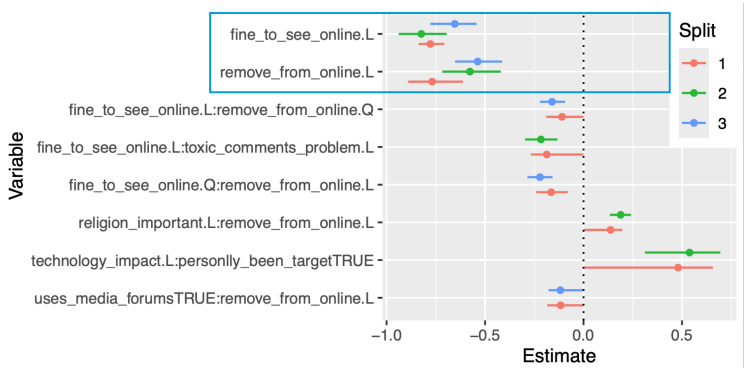


Results: Simulating New Batches



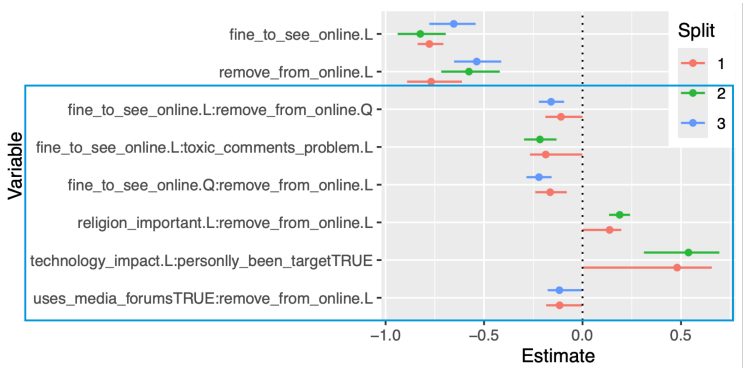
$\rightarrow \frac{8}{36}$ **effects surviving in** ≥ 2 **splits**

Results: Simulating New Batches



→ **Only item-specific attitudes survive in all three splits**

Results: Simulating New Batches



→ **Item-specific attitudes interact with lived experience and identity**

Conclusions

Conclusions

- Considerable variation even within a bundle of tasks; annotators, items
  $\rightarrow$ **Should we expect generalization?**

Conclusions

- Considerable variation even within a bundle of tasks; annotators, items
  $\rightarrow$ Should we expect generalization?
- To understand variation, we may want to take into account item-annotator interactions or collect item-specific attitudes

Conclusions

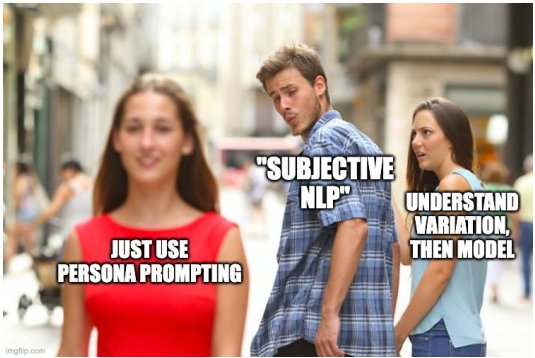
- Considerable variation even within a bundle of tasks; annotators, items
  $\rightarrow$ Should we expect generalization?
- To understand variation, we may want to take into account item-annotator interactions or collect item-specific attitudes
- Let's document our collected annotations better: batches, annotation setting, etc.

Conclusions

- Considerable variation even within a bundle of tasks; annotators, items
  $\rightarrow$ Should we expect generalization?
- To understand variation, we may want to take into account item-annotator interactions or collect item-specific attitudes
- Let's document our collected annotations better: batches, annotation setting, etc.
- **Reflect**
  - Which items do we want or need more annotations on?
  - Which annotator characteristics do we need to collect?

Thank you! Questions?

Preprint



Contact me

Appendix

Annotator Characteristics

Feature	Type	Reference	Dataset
Gender	nominal	*male*	all
Age	ordinal		all
Education	ordinal		CTDP, POPQUORN, MHS
Race	nominal	*white*	CTDP, POPQUORN, MHS
Political ideology	ordinal		MHS
Political affiliation	nominal	*Liberal*	CTDP
Socio-economic status	ordinal		D3CODE, POPQUORN
Moral foundations	interval		D3CODE
Country	nominal	*USA*	D3CODE
Media usage	nominal	*no*	CTDP
Task-specific questionnaire	ordinal		CTDP
	nominal	*no*	CTDP
Occupation	nominal	*employed full-time*	POPQUORN
LGBTQ status	nominal	*heterosexual*	CTDP
Trans status	nominal	*no*	MHS

Domain-specific Lexicons

Source	Feature	Explanation
Hatebase	n_hatebase	Number of tokens found on Hatebase
Abusive Words	n_abusive	Number of tokens found in the Abusive Words lexicon
Hurtlex	n_ps	Number of negative stereotype/ethnic slur tokens
	n_rci	Number of location/demonym tokens
	n_pa	Number of profession/occupation tokens
	n_ddf	Number of tokens related to physical disabilities and diversity
	n_ddp	Number of tokens related to cognitive disabilities and diversity
	n_dmc	Number of tokens related to moral and behavioral defects
	n_rci	Number of tokens related to physical disabilities and diversity
	n_is	Number of tokens related to social and economic disadvantage
	n_or	Number of tokens related to plants
	n_an	Number of tokens related to animals
	n_asm	Number of tokens related to male genitalia
	n_asf	Number of tokens related to female genitalia
	n_ps	Number of tokens related to prostitution
	n_om	Number of tokens related to homosexuality
	n_qas	Number of tokens with potentially negative connotations
	n_cds	Number of derogatory tokens
	n_re	Number of tokens related to felonies, crime, and immoral behavior
	n_svp	Number of tokens related to the seven deadly sins of the Christian tradition
Harassment Lexicon	n_generic	Number of tokens related to harassment
	n_sexual	Number of tokens related to sexual harassment
	n_appearance	Number of tokens related to appearance-related harassment
	n_racial	Number of tokens related to racial harassment
	n_intelligence	Number of tokens related to intellectual harassment
	n_politics	Number of tokens related to political harassment
	n_hateful	Number of tokens in the union of all lexicons

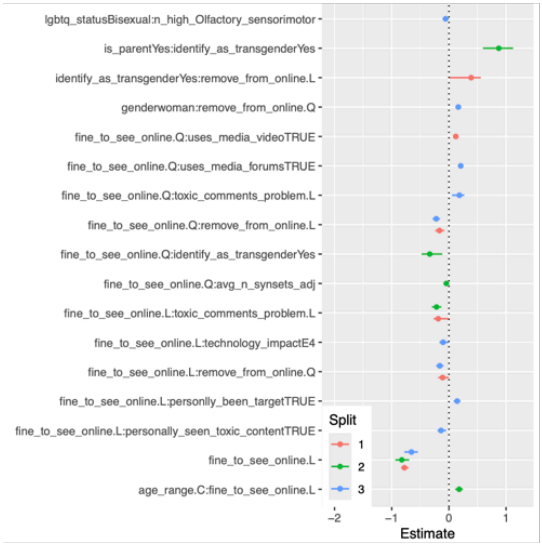
Results

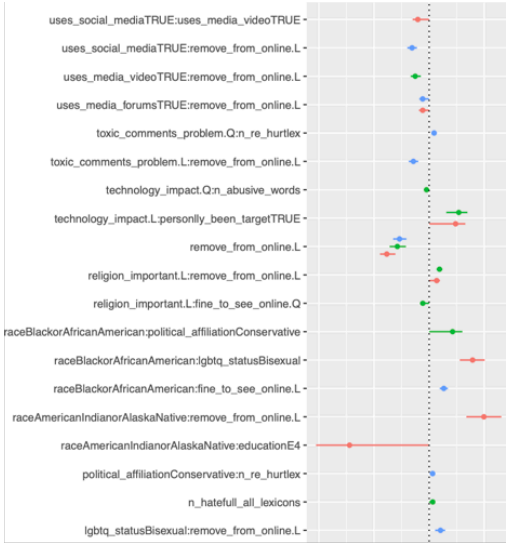
Do Random Intercepts Both for Annotators and Items Make a Difference?

	`(1 | item)`			`(1 | item) + (1| annotator)`		
	Coef.	Std.Err.	$P > \|t\|$	Coef.	Std.Err.	$P > \|t\|$
(Intercept)	2.10	0.05	< 2e-16	2.09	0.18	<2e-16
gender: Non-binary	-0.23	0.06	0.00	-0.23	0.21	0.29
gender: Woman	-0.02	0.02	0.30	-0.01	0.07	0.84
race: Black or African American	0.18	0.04	0.04	0.19	0.16	0.24
race: Hispanic or Latino	-0.04	0.08	0.02	-0.40	0.28	0.16
race: White	-0.11	0.04	0.01	-0.10	0.13	0.46
age: 18-24	-0.11	0.04	0.01	-0.12	0.15	0.42
age: 25-29	-0.30	0.04	0.00	-0.30	0.16	0.06
age: 30-34	-0.28	0.04	0.00	-0.28	0.15	0.07
age: 35-39	-0.26	0.04	0.00	-0.26	0.15	0.08
age: 40-44	-0.15	0.04	0.00	-0.15	0.15	0.34
age: 45-49	-0.20	0.04	0.00	-0.20	0.16	0.21
age: 50-54	-0.26	0.05	0.00	-0.27	0.17	0.11
age: 54-59	-0.11	0.04	0.00	-0.11	0.14	0.43
age: 60-64	0.20	0.05	0.00	0.19	0.18	0.29
education: Graduate degree	0.07	0.03	0.01	0.07	0.09	0.48
education: High school diploma or equivalent	0.02	0.02	0.51	0.02	0.08	0.77

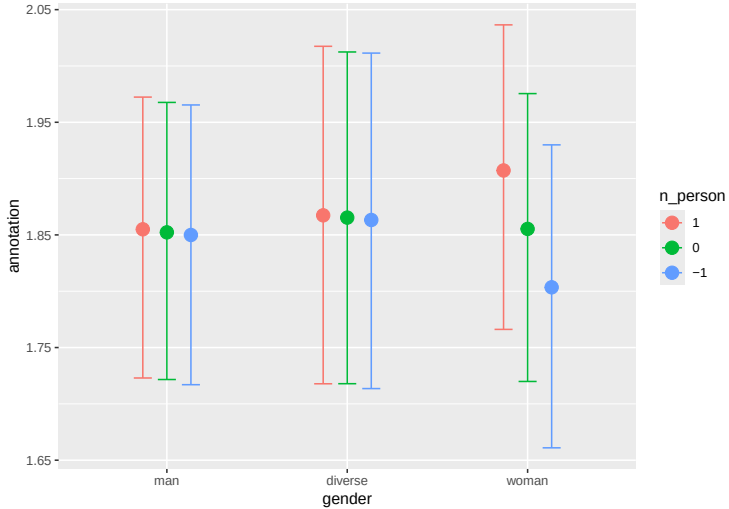
$\rightarrow$ **Yes**

Results: Simulating New Batches

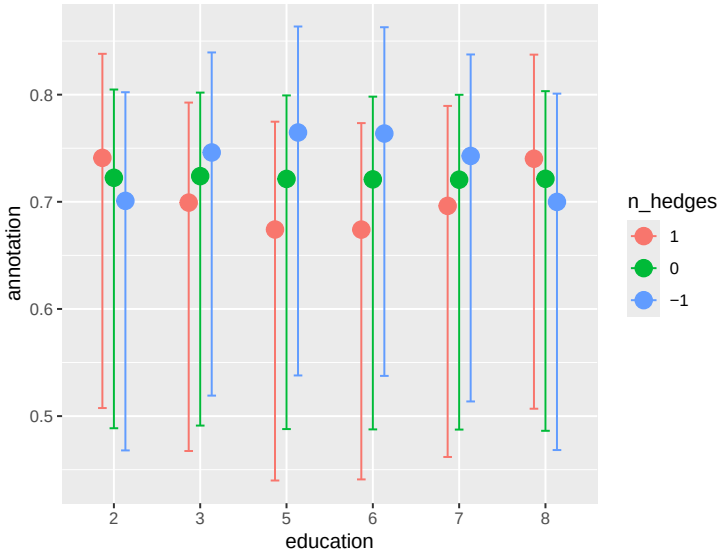




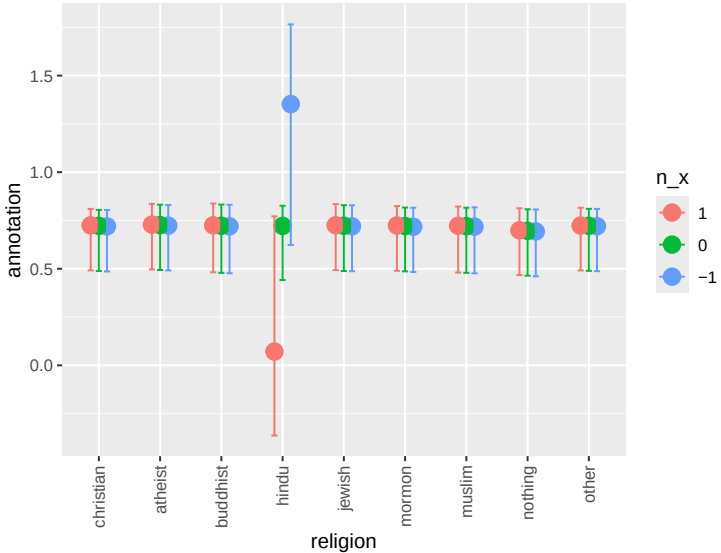
Interactions: POPQUORN



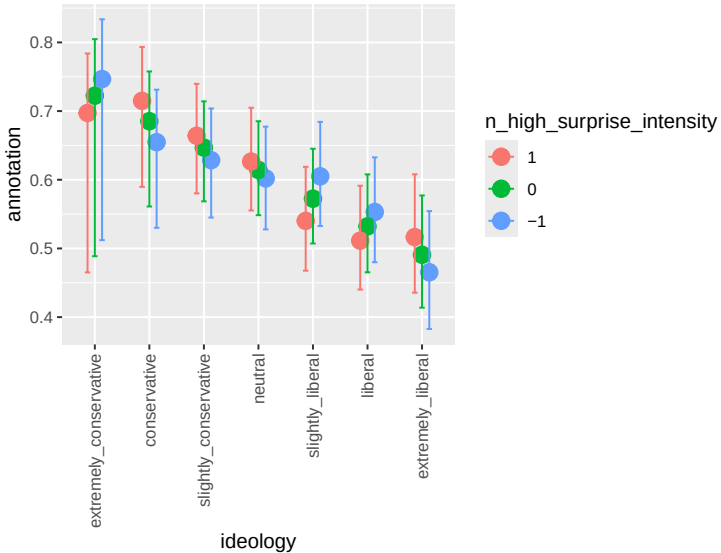
Interactions: MHS



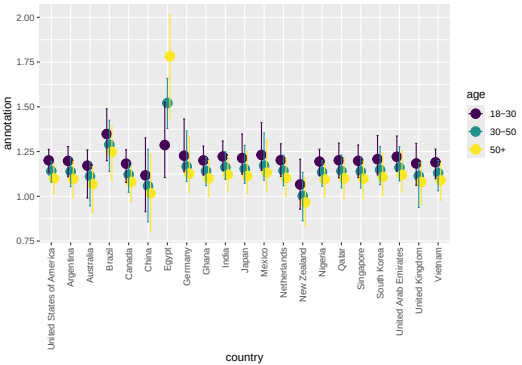
Interactions: MHS

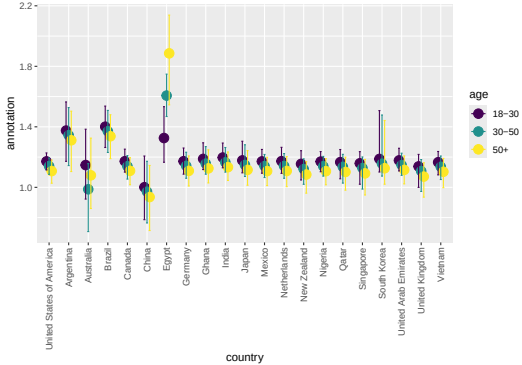


Interactions: MHS

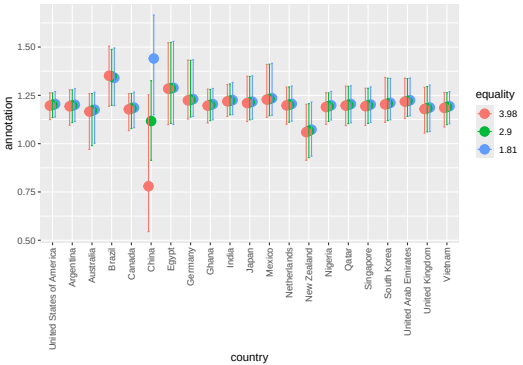


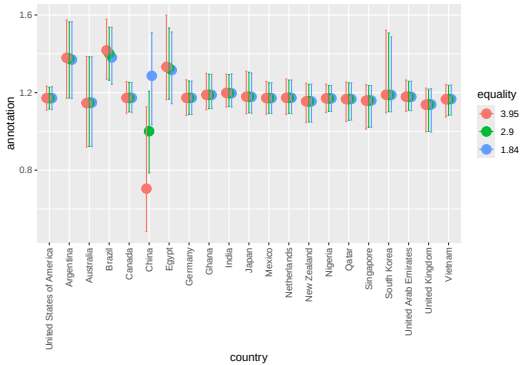
Interactions: D3CODE



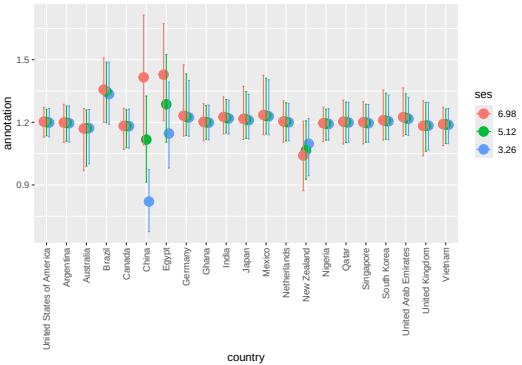


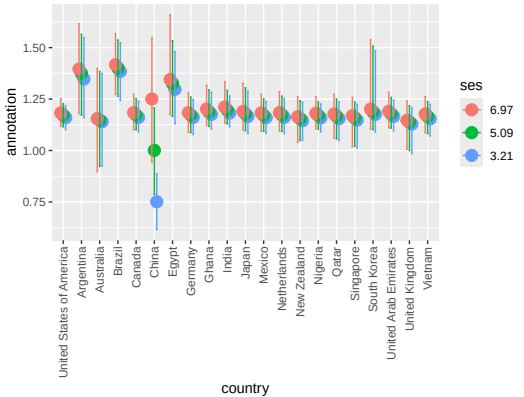
Interactions: D3CODE



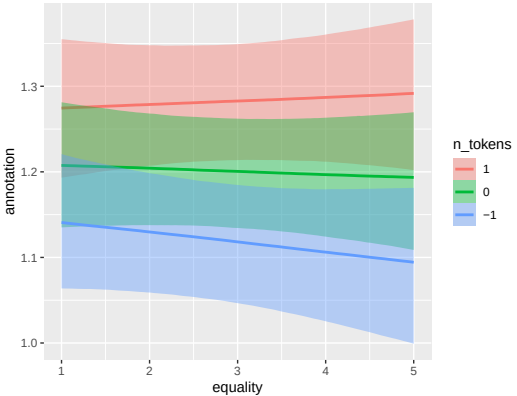


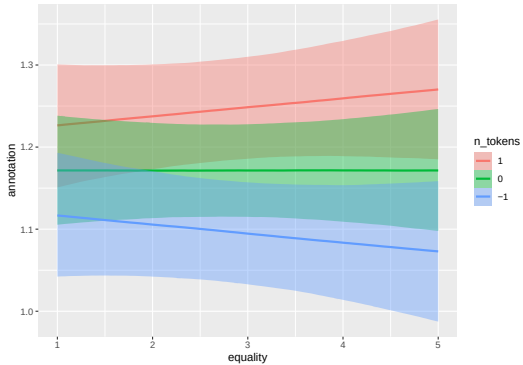
Interactions: D3CODE



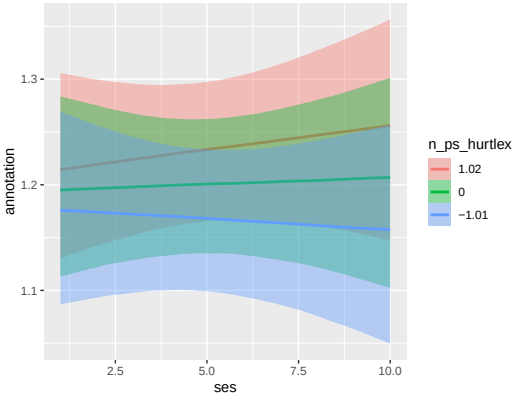


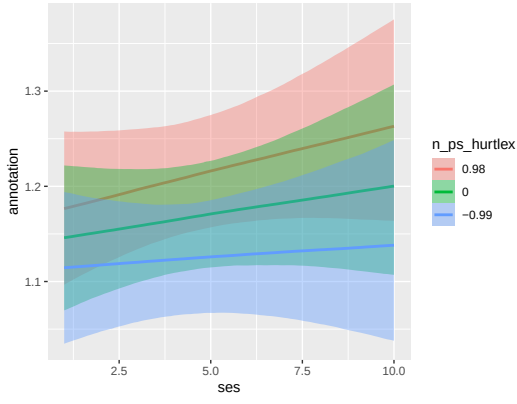
Interactions: D3CODE



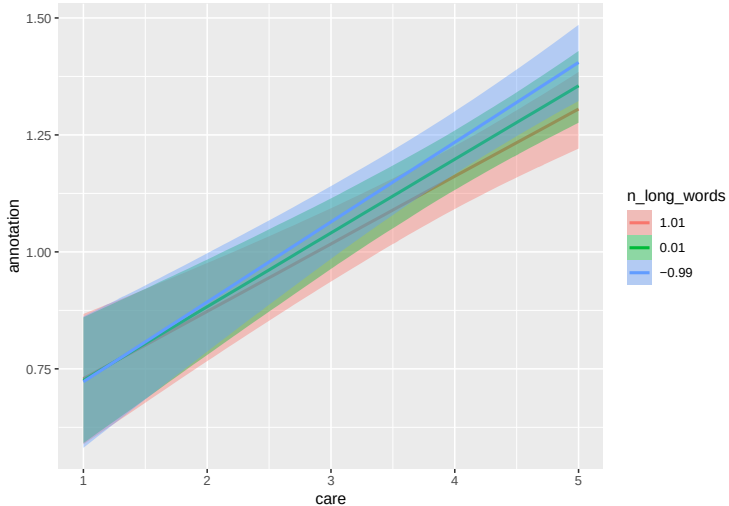


Interactions: D3CODE

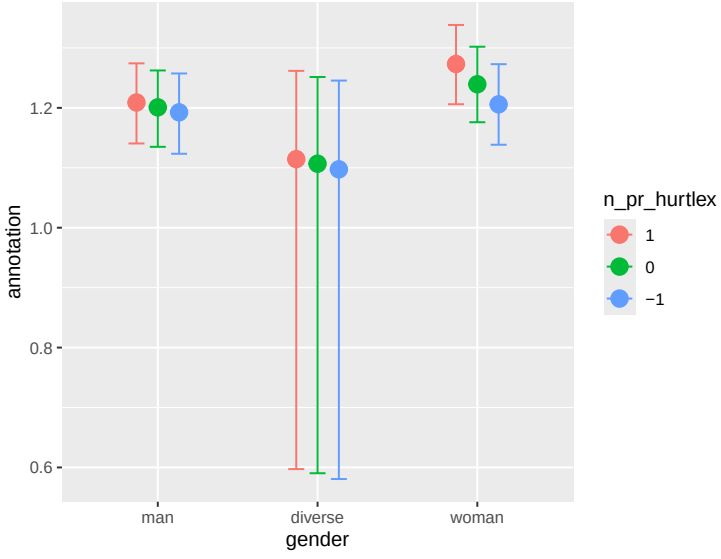




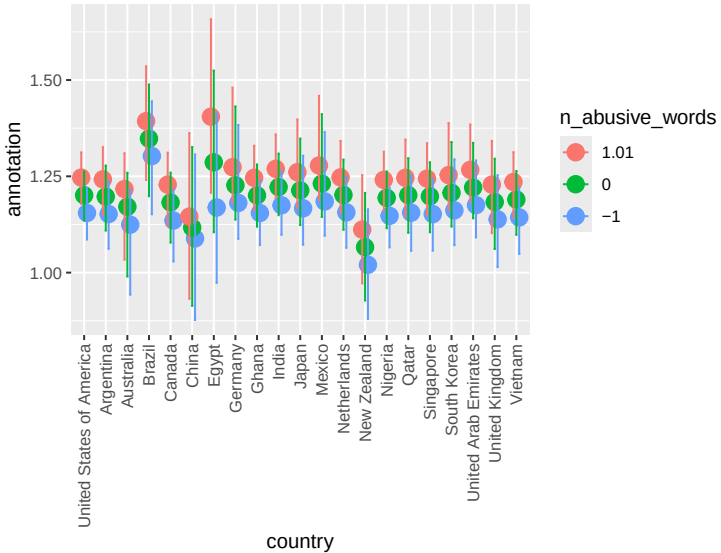
Interactions: D3CODE



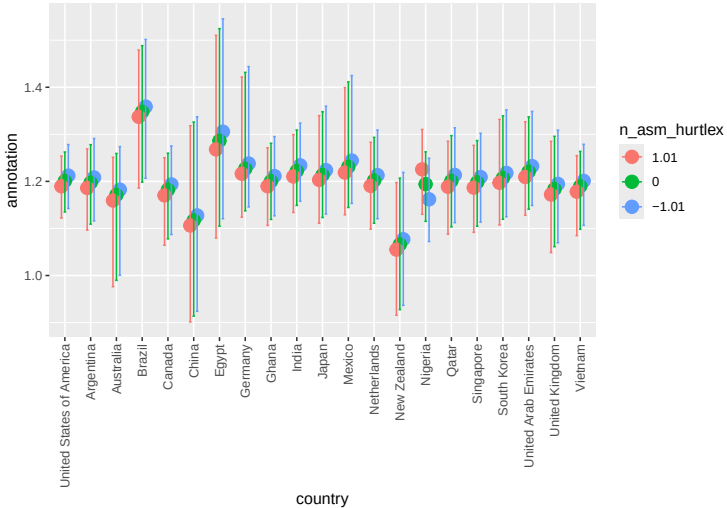
Interactions: D3CODE



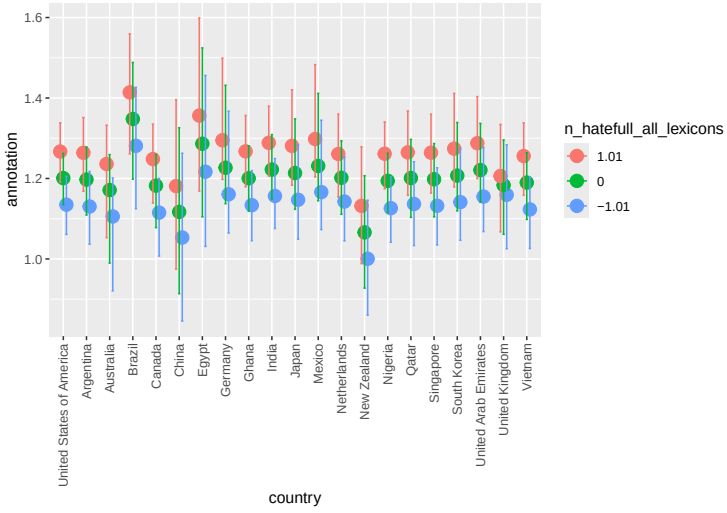
Interactions: D3CODE



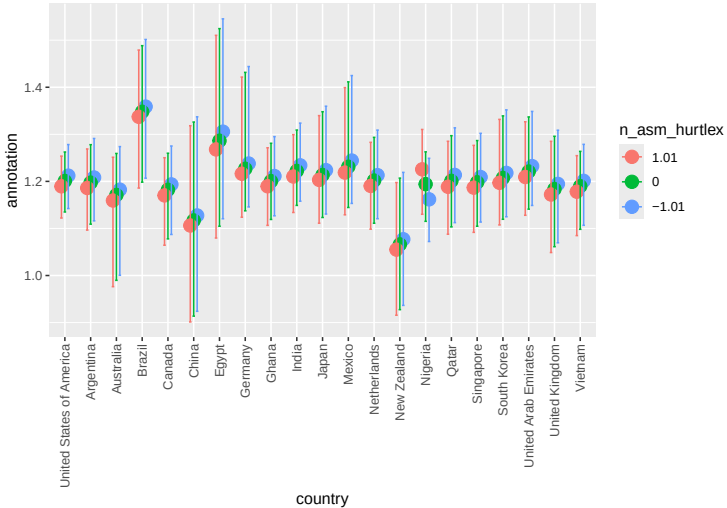
Interactions: D3CODE



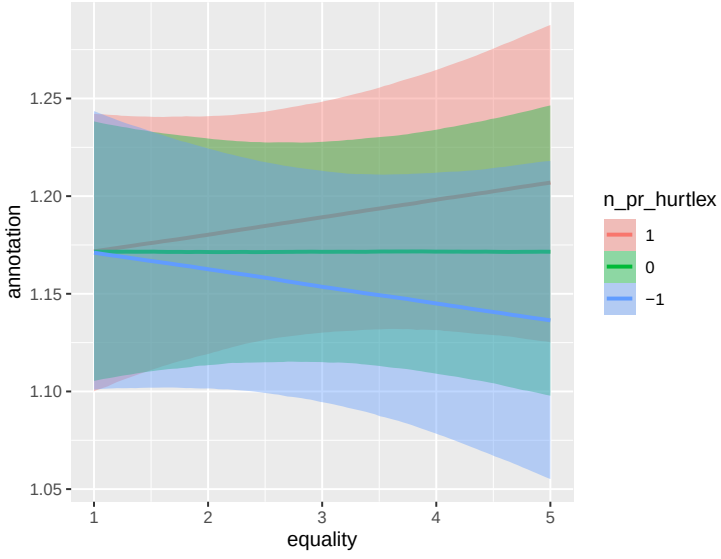
Interactions: D3CODE



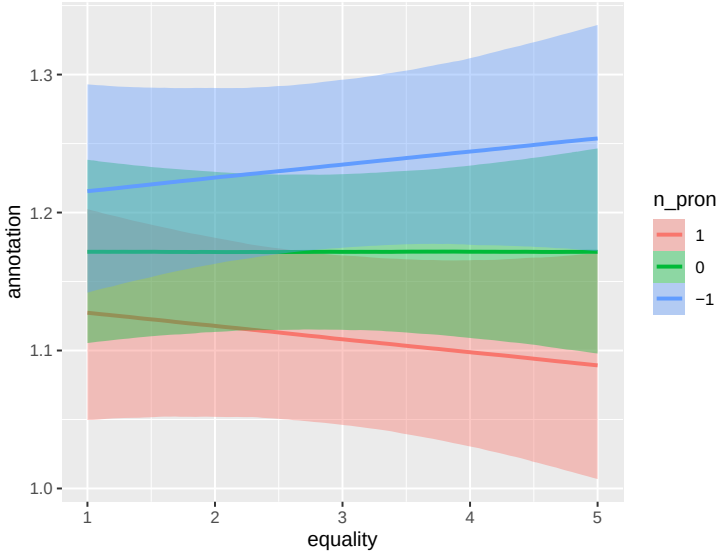
Interactions: D3CODE



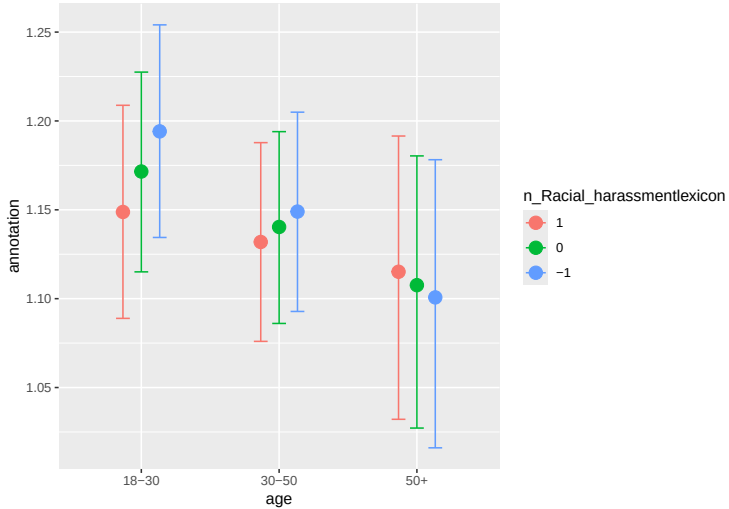
Interactions: D3CODE



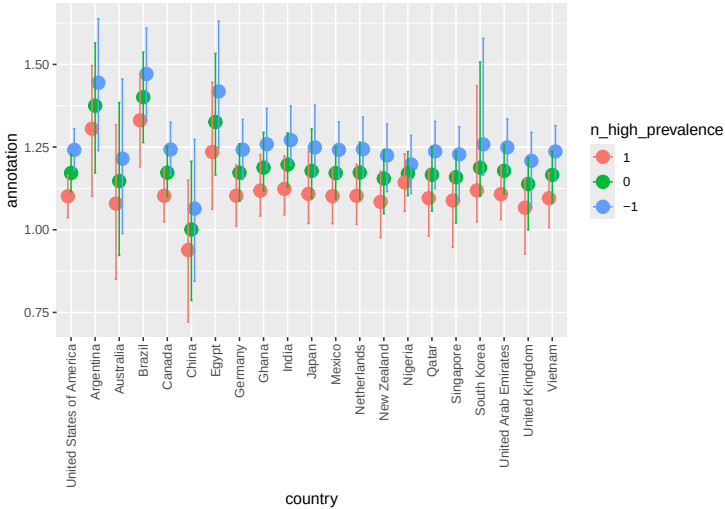
Interactions: D3CODE



Interactions: D3CODE



Interactions: D3CODE



References

References

Julia Romberg, Maximilian Maurer, Henning Wachsmuth, and Gabriella Lapesa. 2025.

Towards a Perspectivist Turn in Argument Quality Assessment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7458–7485, Albuquerque, New Mexico. Association for Computational Linguistics.

Wan, R., Kim, J., & Kang, D. 2023.

Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information. in *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12), 14523–14530.

References

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023.

The Ecological Fallacy in Annotation: Modeling Human Label Variation goes beyond Sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.