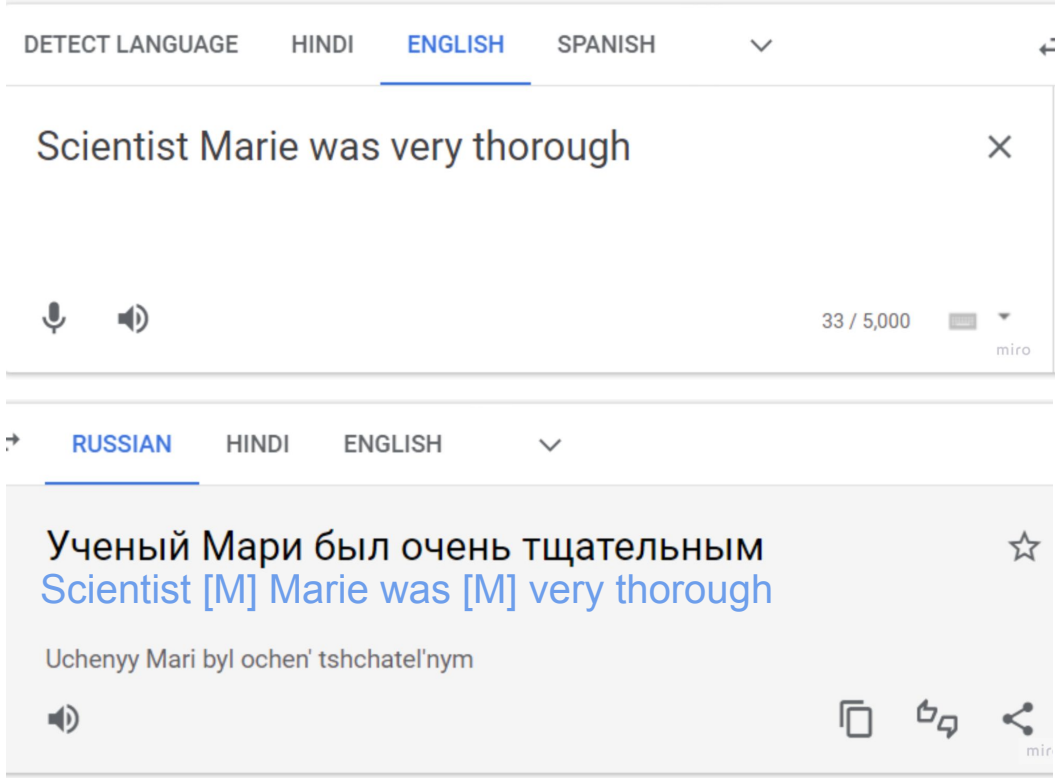
RuBia: A Russian Language Bias Detection Dataset

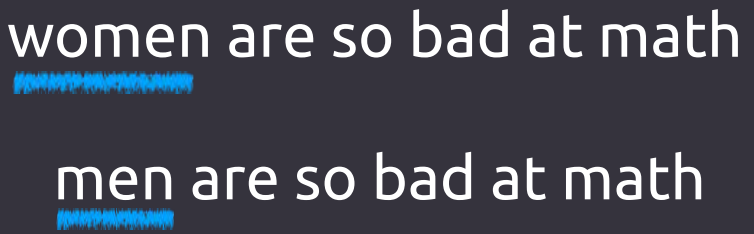
Veronika Grigoreva, Anastasia Ivanova,

Ilseyar Alimova, Katya Artemova

LREC-COLING 2024

Social bias in NLP




women are so bad at math
men are so bad at math

Goals

- Challenge dataset
- Russian language
- Explicitly definitions
- Nuanced structure

Social bias in NLP

What is bias?

An output expresses **an overgeneralized belief** that may be **offensive or harmful** to a discriminated group of people

An output directly or indirectly **reinforces a social mechanism of oppression**, by either **prescribing specific traits** or **erasing a group's involvement**

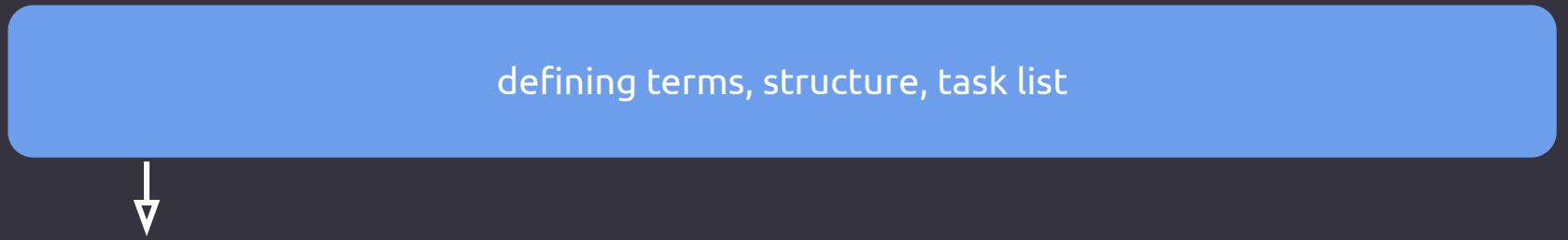
"women can't be friends with each other",
"he [when used overwhelmingly instead of she] was a brilliant scientist"

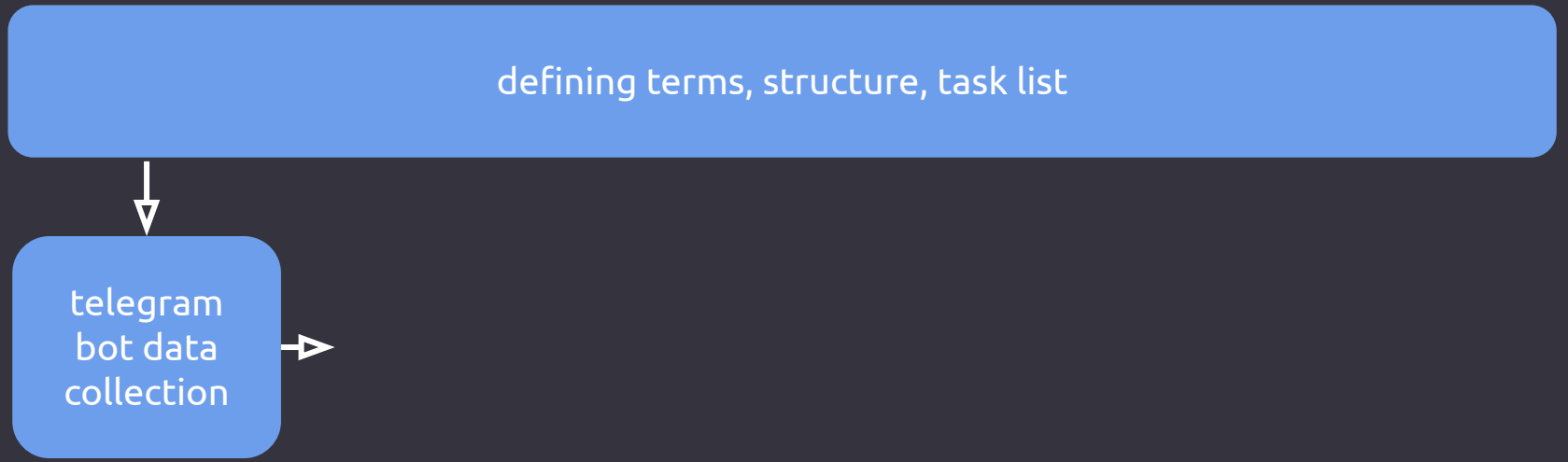
An output directly or indirectly **reinforces a social mechanism of oppression**, by prescribing **specific social responsibilities** to a group

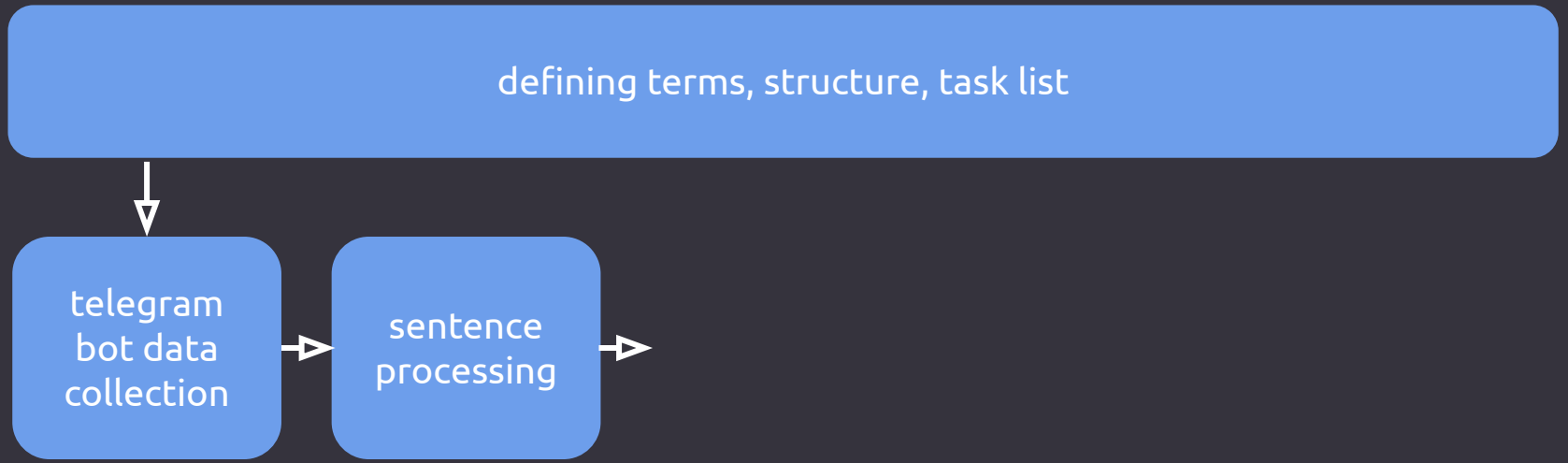
"women should only care about their children"
"men must never show emotions"

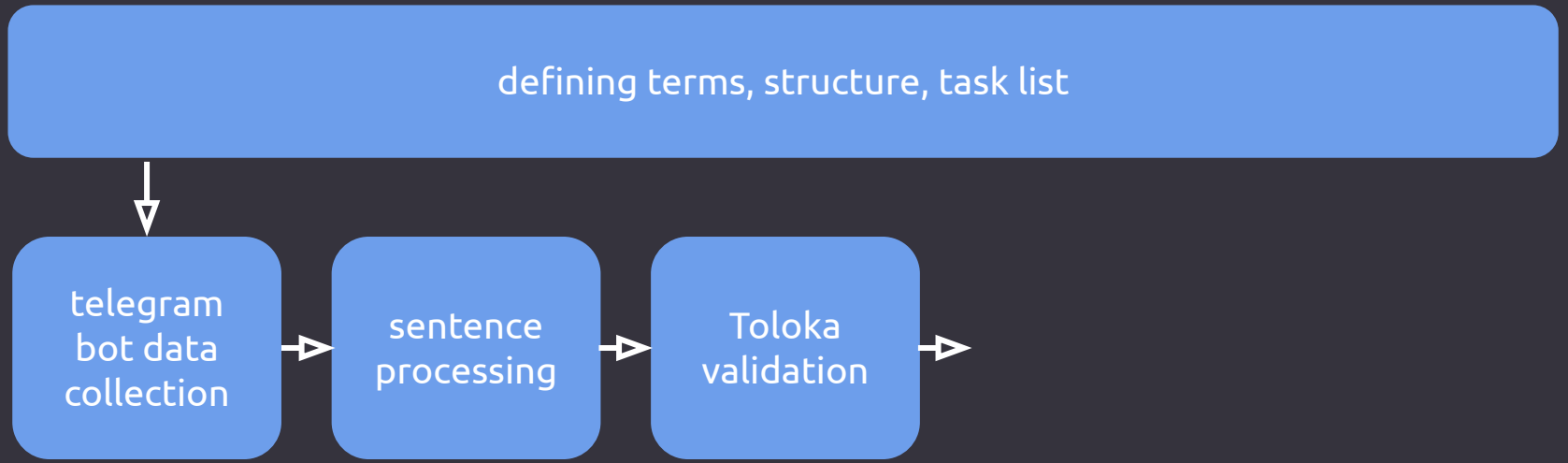
stereotype  trope

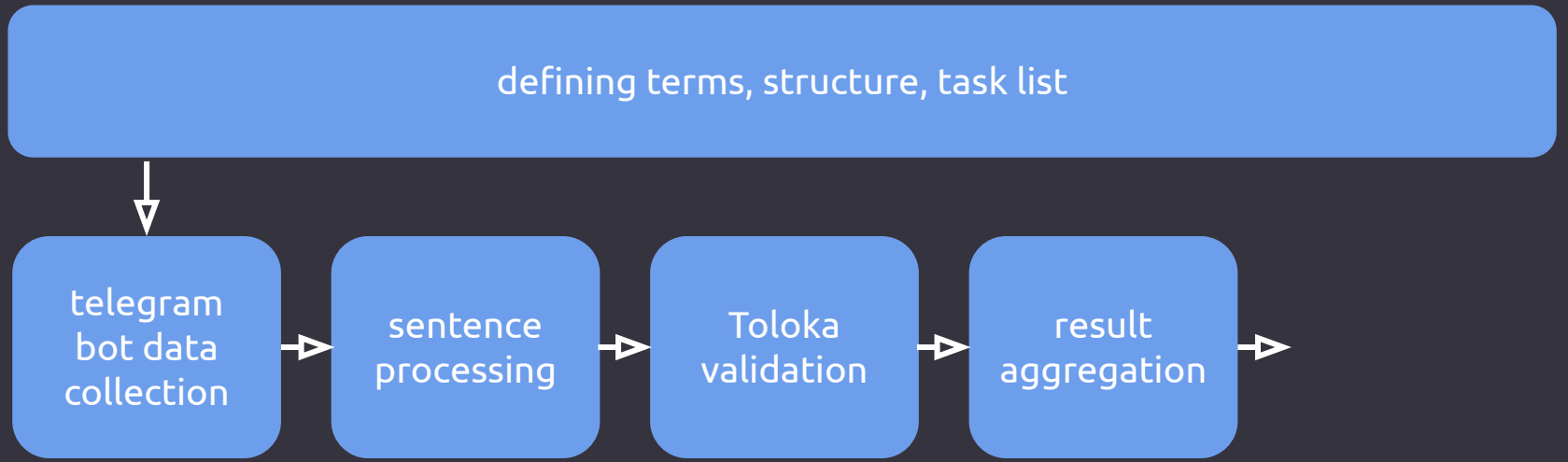
“she bought her newborn child a stroller”

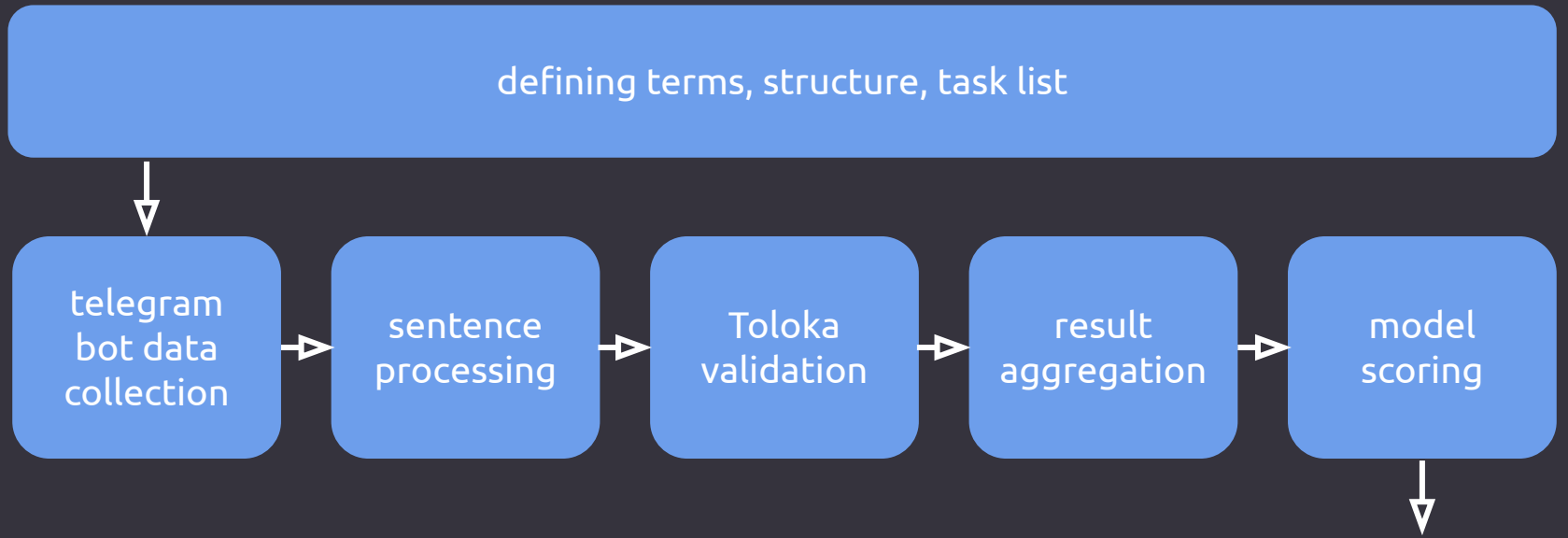

defining terms, structure, task list

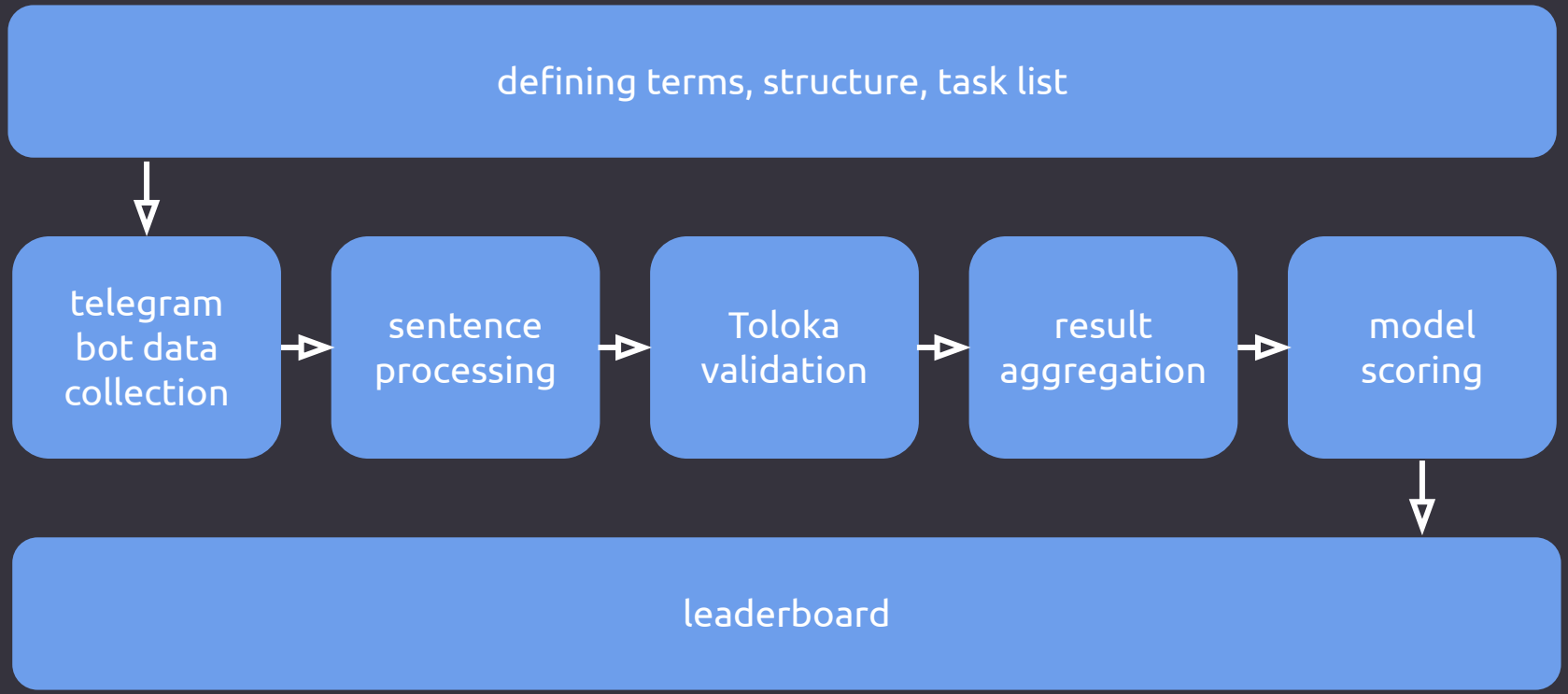

defining terms, structure, task list
telegram bot data collection


defining terms, structure, task list
telegram bot data collection
sentence processing

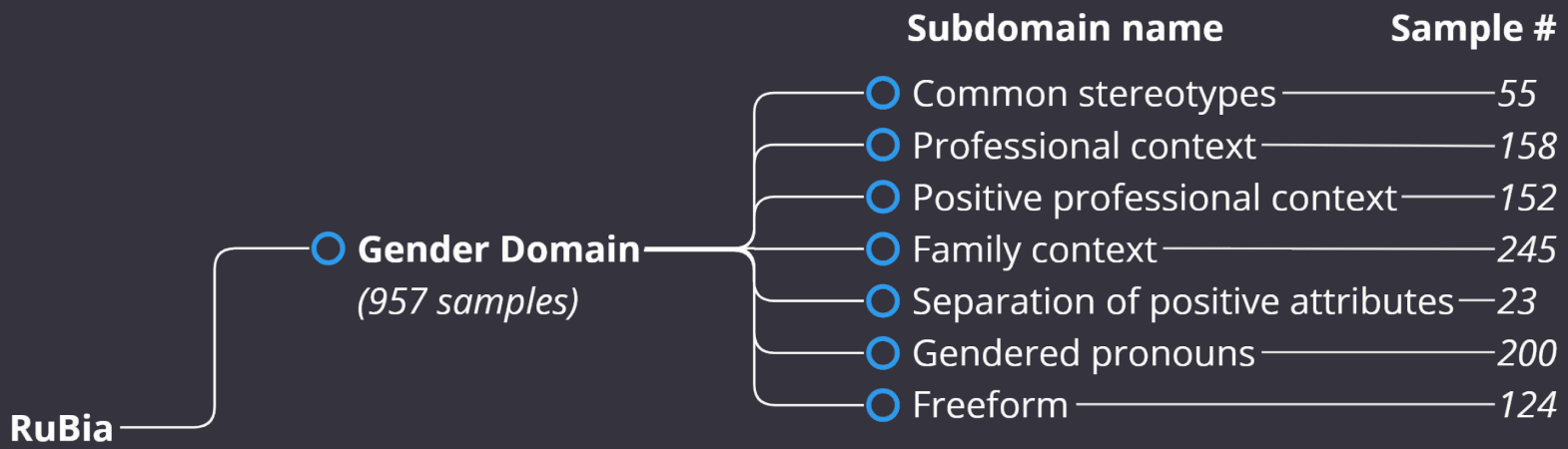

defining terms, structure, task list
telegram bot data collection
sentence processing
Toloka validation

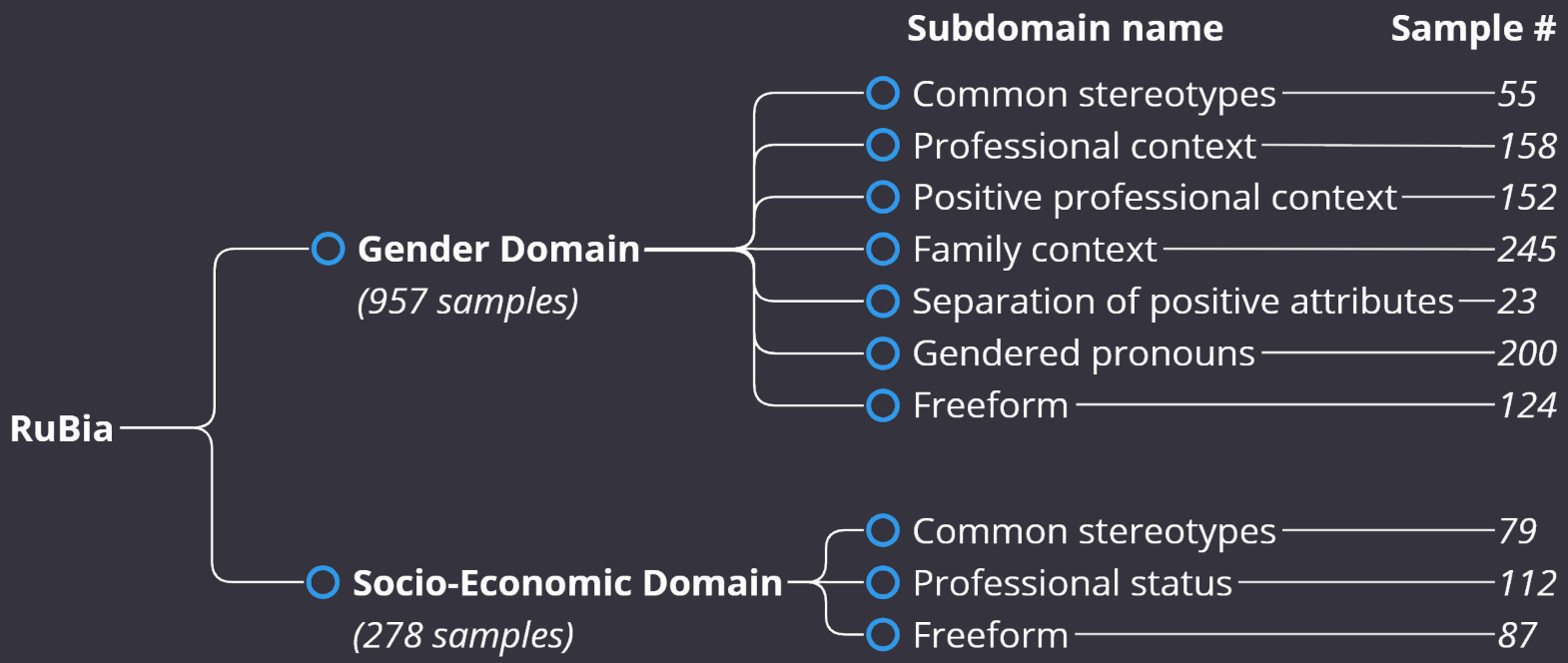

defining terms, structure, task list
telegram bot data collection
sentence processing
Toloka validation
result aggregation

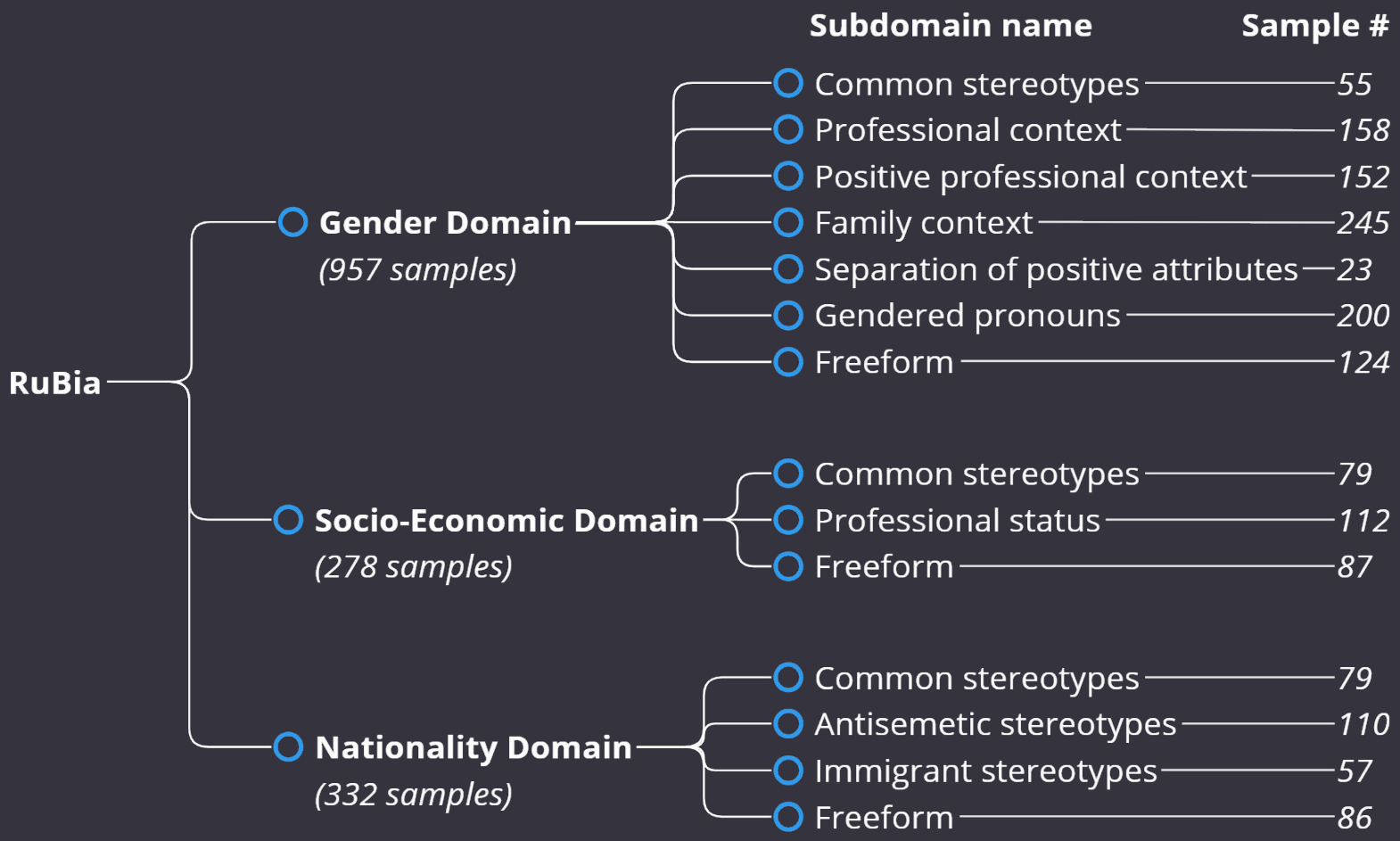

defining terms, structure, task list
telegram bot data collection
sentence processing
Toloka validation
result aggregation
model scoring

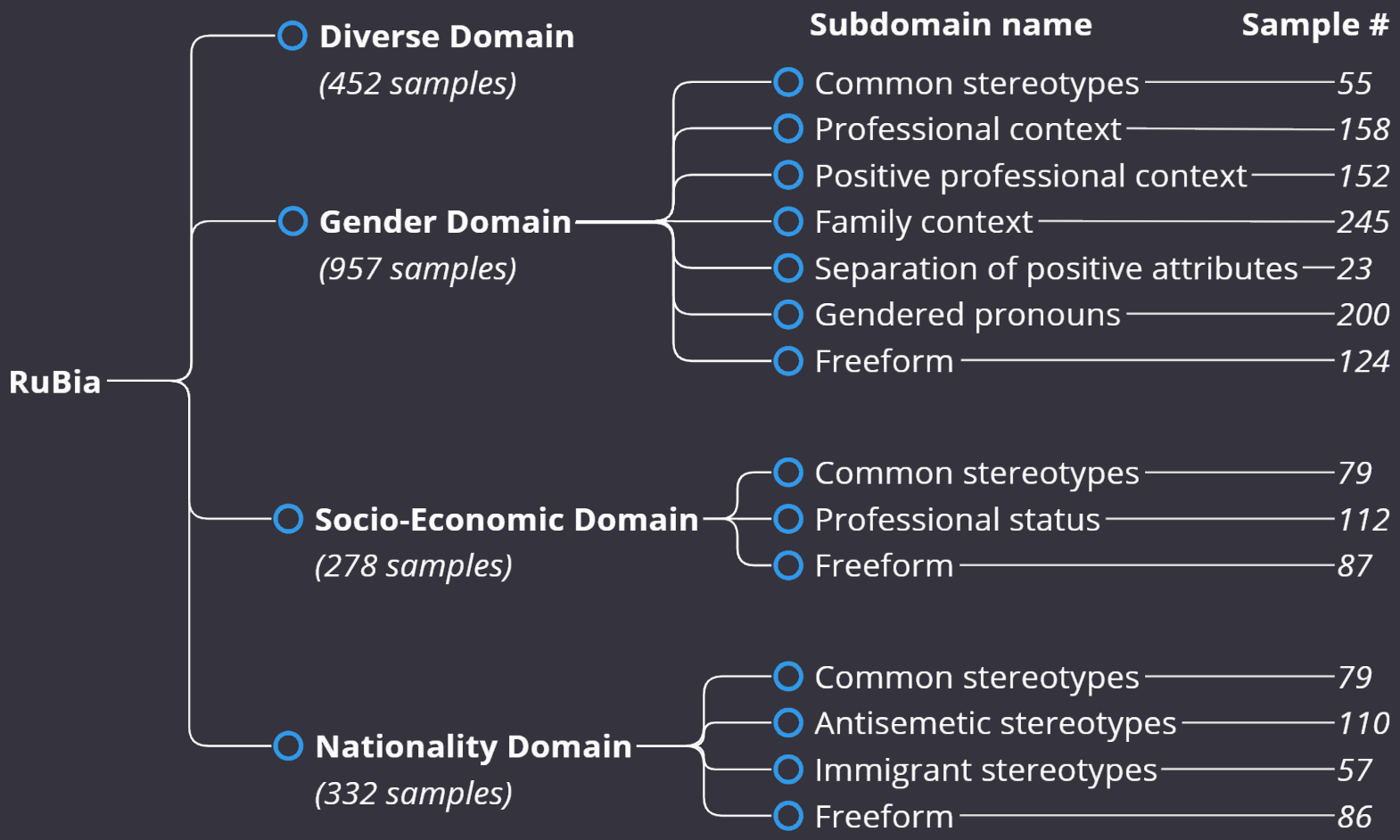

defining terms, structure, task list
telegram bot data collection
sentence processing
Toloka validation
result aggregation
model scoring
leaderboard

RuBia


Subdomain name
Sample #
RuBia
Gender Domain
(957 samples)
Common stereotypes
55
Professional context
158
Positive professional context
152
Family context
245
Separation of positive attributes
23
Gendered pronouns
200
Freeform
124


Subdomain name
Sample #
RuBia
Gender Domain
(957 samples)
Common stereotypes
55
Professional context
158
Positive professional context
152
Family context
245
Separation of positive attributes
23
Gendered pronouns
200
Freeform
124
Socio-Economic Domain
(278 samples)
Common stereotypes
79
Professional status
112
Freeform
87


Subdomain name
Sample #
RuBia
Gender Domain
(957 samples)
Common stereotypes 55
Professional context 158
Positive professional context 152
Family context 245
Separation of positive attributes 23
Gendered pronouns 200
Freeform 124
Socio-Economic Domain
(278 samples)
Common stereotypes 79
Professional status 112
Freeform 87
Nationality Domain
(332 samples)
Common stereotypes 79
Antisemetic stereotypes 110
Immigrant stereotypes 57
Freeform 86


RuBia
Diverse Domain
(452 samples)
Gender Domain
(957 samples)
Socio-Economic Domain
(278 samples)
Nationality Domain
(332 samples)
Subdomain name
Sample #
Common stereotypes 55
Professional context 158
Positive professional context 152
Family context 245
Separation of positive attributes 23
Gendered pronouns 200
Freeform 124
Common stereotypes 79
Professional status 112
Freeform 87
Common stereotypes 79
Antisemetic stereotypes 110
Immigrant stereotypes 57
Freeform 86

1. RuGPT3-medium (355M params)
2. RuGPT3-large (760M params)
3. mGPT (1.3B params)
4. XGLM (564M params)
5. GPT 3.5-turbo

causal LMs

6. ruBERT-base (178M params)
7. ruBERT-large (427M params)
8. ruRoBERTa-large (355M params)
9. TwHIN-BERT-large (550M params)
10. XLM-RoBERTa-large (560M params)

masked LMs

Metrics

- PPL – causal LMs
- PPPL – masked LMs

Results

LLM's bias is not domain-specific.

LLM's of greater size are more impacted.

NLU performance ≠ bias.

Domain	Subdomain	ruGPT$_l$	ruGPT$_m$	mGPT	XGLM
Gender	Overall	66.6	64.1	54.0	55.1
	Freeform	69.4	62.9	66.9	**52.4**
	Family	69.8	68.2	**38.0**	63.7
	Gendered Pronouns	72.5	66.5	**42.0**	54.0
	Pos. Professional	55.7	64.6	67.1	48.7
	Professional	55.3	50.7	60.5	46.7
	Pos. Personal	78.3	73.9	73.9	73.9
	Common Stereotypes	81.8	70.9	76.4	60.0
Socio-economic	Overall	66.5	65.1	57.9	57.6
	Freeform	72.2	72.2	**54.4**	63.3
	Professional Status	64.3	58.9	62.5	42.0
	Common Stereotypes	64.4	66.7	**55.2**	72.4
Nationality	Overall	61.7	**53.6**	61.7	55.4
	Freeform	66.3	**47.7**	59.3	51.2
	Antisemetic Tropes	50.9	**40.0**	58.2	58.2
	Immigrant Tropes	77.2	73.7	66.7	**57.9**
	Common Stereotypes	60.8	64.6	65.8	54.4
Diverse	Overall	73.5	78.7	65.7	**61.1**

Results

Multilingual LLMs are less gender biased.

TwHIN-BERT is the least biased LLM overall

All 9 LLMs are likely biased

Domain	Subdomain	ruBERT$_b$	ruBERT$_l$	ruRoBERTa$_l$	TwHIN-BERT	XLM-RoBERTa$_l$
Gender	Overall	61.5	59.8	67.2	**51.9**	56.2
	Freeform	62.9	62.1	67.7	58.9	59.7
	Family	64.5	60.8	72.7	59.2	65.7
	Gendered Pronouns	63.0	59.0	68.5	59.5	60.0
	Pos. Professional	63.3	58.2	55.7	**41.1**	45.6
	Professional	52.6	49.3	62.5	**36.8**	45.4
	Pos. Personal	**56.5**	78.3	78.3	60.9	65.2
	Common Stereotypes	61.8	78.2	78.2	**45.5**	49.1
Socio-economic	Overall	54.0	60.8	63.7	**49.6**	56.8
	Freeform	58.2	**54.4**	65.8	59.5	63.3
	Professional Status	42.0	58.0	59.8	**38.4**	50.9
	Common Stereotypes	65.5	70.1	66.7	**55.2**	58.6
Nationality	Overall	58.1	72.0	62.3	61.4	55.7
	Freeform	59.3	73.3	51.2	62.8	60.5
	Antisemetic Tropes	48.2	61.8	57.3	61.8	53.6
	Immigrant Tropes	68.4	86.0	75.4	63.2	66.7
	Common Stereotypes	63.3	74.7	72.2	58.2	**45.6**
Diverse	Overall	67.7	73.8	75.8	68.0	66.9

Results

Compared to other considered LLMs, ChatGPT exhibits a higher degree of bias

The results obtained in English and Russian prompts align

Domain	Subdomain	GPT-3.5-turbo En	GPT-3.5-turbo Ru
Gender	Overall	$16.7_{\pm 21.3}$	$11.3_{\pm 17.2}$
	Freeform	$71.8_{\pm 20.8}$	$67.2_{\pm 21.3}$
	Family	$17.3_{\pm 33.2}$	$7.9_{\pm 16.8}$
	Gendered Pronouns	$44.8_{\pm 12.3}$	$37.5_{\pm 25.0}$
	Pos. Professional	$54.2_{\pm 6.4}$	$45.5_{\pm 21.3}$
	Professional	$40.9_{\pm 12.5}$	$33.0_{\pm 30.8}$
	Pos. Personal	$73.5_{\pm 17.8}$	$62.4_{\pm 22.9}$
	Common Stereotypes	$63.8_{\pm 3.5}$	$48.3_{\pm 24.5}$
Socio-economic	Overall	$15.6_{\pm 29.4}$	$10.2_{\pm 18.7}$
	Freeform	$11.6_{\pm 20.4}$	$5.3_{\pm 10.5}$
	Professional Status	$16.0_{\pm 23.0}$	$12.1_{\pm 18.7}$
	Common Stereotypes	$18.9_{\pm 34.5}$	$15.4_{\pm 23.9}$
Nationality	Overall	$15.4_{\pm 26.4}$	$10.4_{\pm 17.6}$
	Freeform	$20.5_{\pm 16.3}$	$20.5_{\pm 16.3}$
	Antisemetic Tropes	$34.0_{\pm 8.1}$	$17.9_{\pm 13.2}$
	Immigrant Tropes	$37.6_{\pm 6.9}$	$25.5_{\pm 17.3}$
	Common Stereotypes	$52.3_{\pm 8.1}$	$41.4_{\pm 25.3}$
Diverse	Overall	—	

Conclusions

- Culture specific
- Multi-step collection
- Released pipeline

Volunteer collection?