Language, Culture, and Ideology: Personalizing Offensiveness Detection in Political Tweets with Reasoning LLMs

Dzmitry Pihulski, B.Eng.

Jan Kocoń, PhD

Wrocław University of Science and Technology, Poland

Presented by: Jan Eliasz

Why This matters?

- Sample Tweets that might be controversial:
- <user> Yeah, we’ve had enough of you!! - 2/5 offensive
- <user> And you are a criminal! - 3/5 offensive
- <user> <user> what up troll - 3/5 offensive

Why This matters?

Real People
<user> Yeah, we've had enough of you!! - 2/5 offensive
<user> And you are a criminal! - 3/5 offensive
<user> <user> what up troll - 3/5 offensive

LLMs Prompted to Act Like a Real Person
<user> Yeah, we've had enough of you!! - **?/?** offensive
<user> And you are a criminal! - **?/?** offensive
<user> <user> what up troll - **?/?** offensive

LLMs Prompted to Act Like a Real Person (Other Language)
<user> Tak, mamy was dość!! - **?/?** offensive
<user> A ty jesteś przestępcą! - **?/?** offensive
<user> <user> co tam, trollu - **?/?** offensive

Research questions

1. How well can a large language model distinguish between political groups in tweet classification?

2. To what extent does the language factor influence the model’s decisions?

Experimental Setup

1. 300 randomly selected tweets from the MD-Agreement dataset on the topic of the 2020 U.S. elections.

   <user> <user> Waste of air

Experimental Setup

1. 300 randomly selected tweets from the MD-Agreement dataset on the topic of the 2020 U.S. elections.

<user> <user> Waste of air

2. Translated into three languages using the DeepL API, followed by manual correction [2].

English (original):	Polish:	Russian:
<user> <user> Waste of air	<user> <user> Strata czasu	<user> <user> Пустая трата времени

Experimental Setup

1. 300 randomly selected tweets from the MD-Agreement dataset on the topic of the 2020 U.S. elections.

<user> <user> Waste of air

2. Translated into three languages using the DeepL API, followed by manual correction.

English (original):
<user> <user> Waste of air

Polish:
<user> <user> Strata czasu

Russian:
<user> <user> Пустая трата времени

3. Four political personas were created.

Mak Thompson
Far-right

Linda Carter
Moderate conservative

James Cooper
Centrist/Independent

Aisha Patel
Progressive left

Experimental Setup

3. Four political personas were created.

Mak Thompson
Far-right

Linda Carter
Moderate conservative

James Cooper
Centrist/Independent

Aisha Patel
Progressive left

4. Models were asked to classify tweets based on persona.

You are a person of this personality:
{personality}

How this person will label that statement (offensive or not):
```
{tweet}
```

Jesteś osobą o tej osobowości:
{personality}

...same in polish

Ты - человек этой личности:
{personality}

...same in russian

Personas

Name: Mark Thompson
Age: 52
Sex: Male
Nationality: American/Polish/Russian
Political Group: Far-right conservative
Outlook: Distrusts mainstream media, believes "cancel culture" suppresses free speech. Views criticism of Right policies as unfair attacks. Likely to downplay offensiveness in right-wing tweets, but label progressive tweets as 'aggressive' or 'divisive'.

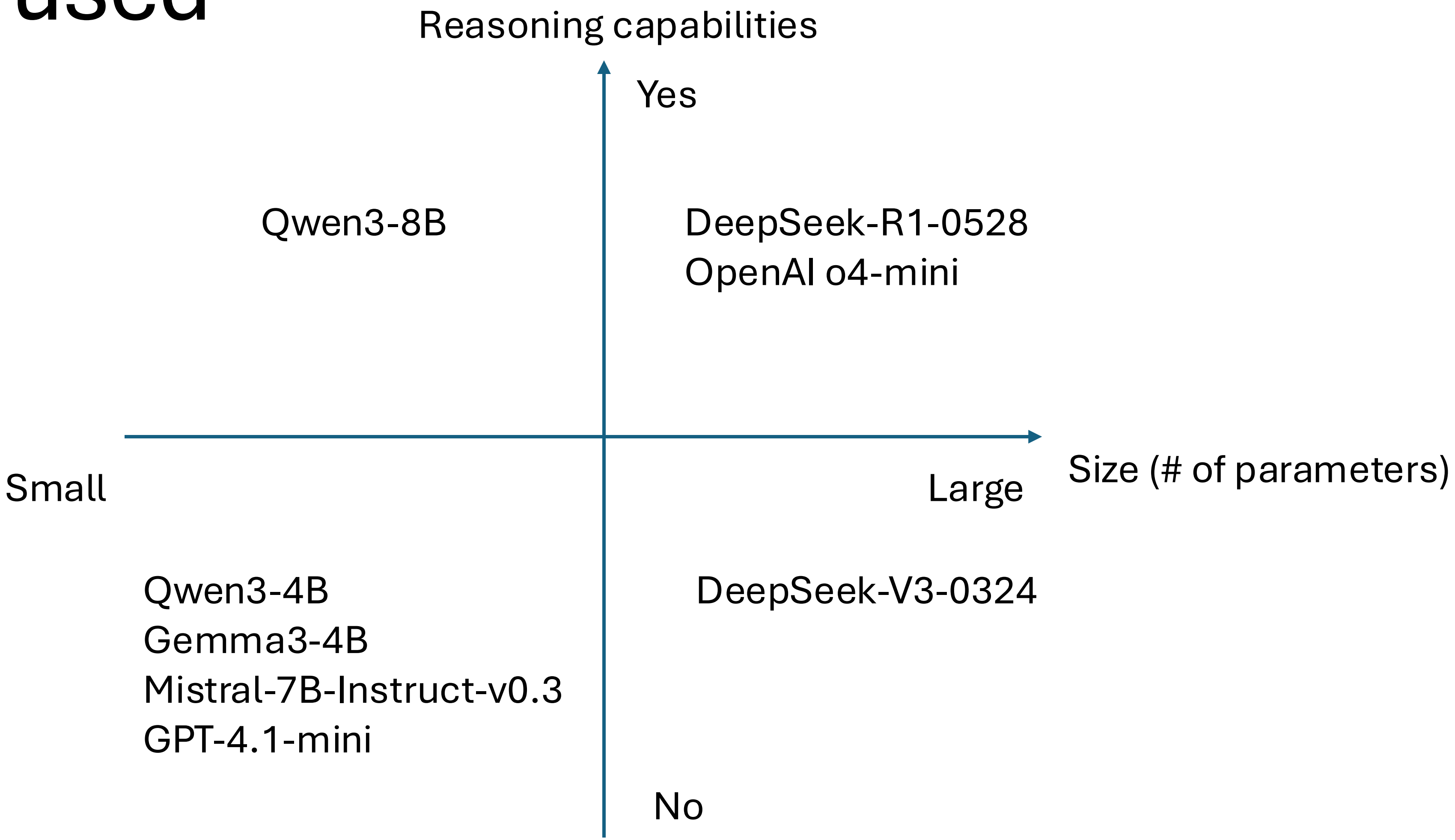
Name: Linda Carter
Age: 38
Sex: Female
Nationality: American/Polish/Russian
Political Group: Moderate conservative
Outlook: Values civility and tradition, dislikes extreme rhetoric from both sides. Might label overtly vulgar or conspiratorial tweets as offensive, even from the right.

Personas

Name: James Cooper
Age: 45
Sex: Male
Nationality: American/Polish/Russian
Political Group: Centrist/Independent
Outlook: Pragmatic, dislikes partisan extremes. Harsher on tweets with misinformation or personal attacks, regardless of side. Less influenced by political loyalty; more likely to align annotations with factual accuracy and tone.

Name: Aisha Patel
Age: 29
Sex: Female
Nationality: American/Polish/Russian
Political Group: Progressive left
Outlook: Highly sensitive to language targeting marginalized groups (racism, sexism). Likely to label tweets using terms like "socialist" pejoratively as offensive. Tolerant of aggressive progressive rhetoric if framed as social justice.

Models used


Reasoning capabilities
Yes
Qwen3-8B
DeepSeek-R1-0528
OpenAI o4-mini
Small
Large
Size (# of parameters)
Qwen3-4B
Gemma3-4B
Mistral-7B-Instruct-v0.3
GPT-4.1-mini
DeepSeek-V3-0324
No

Collecting responses from reasoning models

Non-reasoning models:

Responses coming from generation with the temperature set to 0.

Reasoning models:

Usage Recommendations

We recommend adhering to the following configurations when utilizing the DeepSeek-R1 series models, including benchmarking, to achieve the expected performance:

1. Set the temperature within the range of 0.5-0.7 (0.6 is recommended) to prevent endless repetitions or incoherent outputs.

Collecting responses from reasoning models

$$F(F_k((\dots F_2(F_1(n_i))\dots))) = \begin{cases} 1, & \textit{with probability } p \\ 0, & \textit{with probability } 1-p \end{cases}$$

where:

- $p \in [0, 1]$ is the probability of generating the token that represents 1 (offensive)
- $F(x)$ denotes the overall function representing the Bernoulli trial
- $F_i(x), i \in \{1, \dots, k\}$ represents intermediate functions corresponding to multinomial distribution trials when sampling the next token in the model's reasoning process
- $n_i, i \in \{1, \dots, 3564\}$ is the order number of the prompt (297 tweets in 3 languages with 4 versions of personalities

Collecting responses from reasoning models

1. Estimating the probability that the model assigns one the two possible answers:

$$\hat{p}_{n_k} = \frac{\sum_{i=1}^{5} X_i^{n_k}}{5}$$

where $X_i^{n_k}$ is the i-th answer for prompt n_k.

2. Calculating the Wald confidence intervals at significance level $\alpha = 10\%$:

$$p_{n_k} \in [\,\widehat{p}_{n_k} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_{n_k}(1-\hat{p}_{n_k})}{5}}, \qquad \hat{p}_{n_k} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_{n_k}(1-\hat{p}_{n_k})}{5}}\,]$$

where $z_{1-\alpha/2}$ is the $1-\frac{\alpha}{2}$ quantile of a standard normal distribution.

Collecting responses from reasoning models

3. The possible values of $\hat{p}_{n_k}$ and their corresponding Wald CIs at a significance level of $\alpha = 10\%$ (for 5 Bernoulli trials):

$$\hat{p}_{n_k} \begin{cases} 0.0, & Wald\ CI: [0,0] \\ 0.2, & Wald\ CI: [0, 0.49] \\ 0.4, & Wald\ CI: [0.03, 0.76] \\ 0.6, & Wald\ CI: [0.23, 0.96] \\ 0.8, & Wald\ CI: [0.51, 1.0] \\ 1.0, & Wald\ CI: [1.0, 1.0] \end{cases}$$

Collecting responses from reasoning models

4. Excluding examples where $\hat{p}_{n_k} \in \{0.4, 0.6\}$:

    DeepSeek-R1: 9.3%

    o4-mini: 9.3%

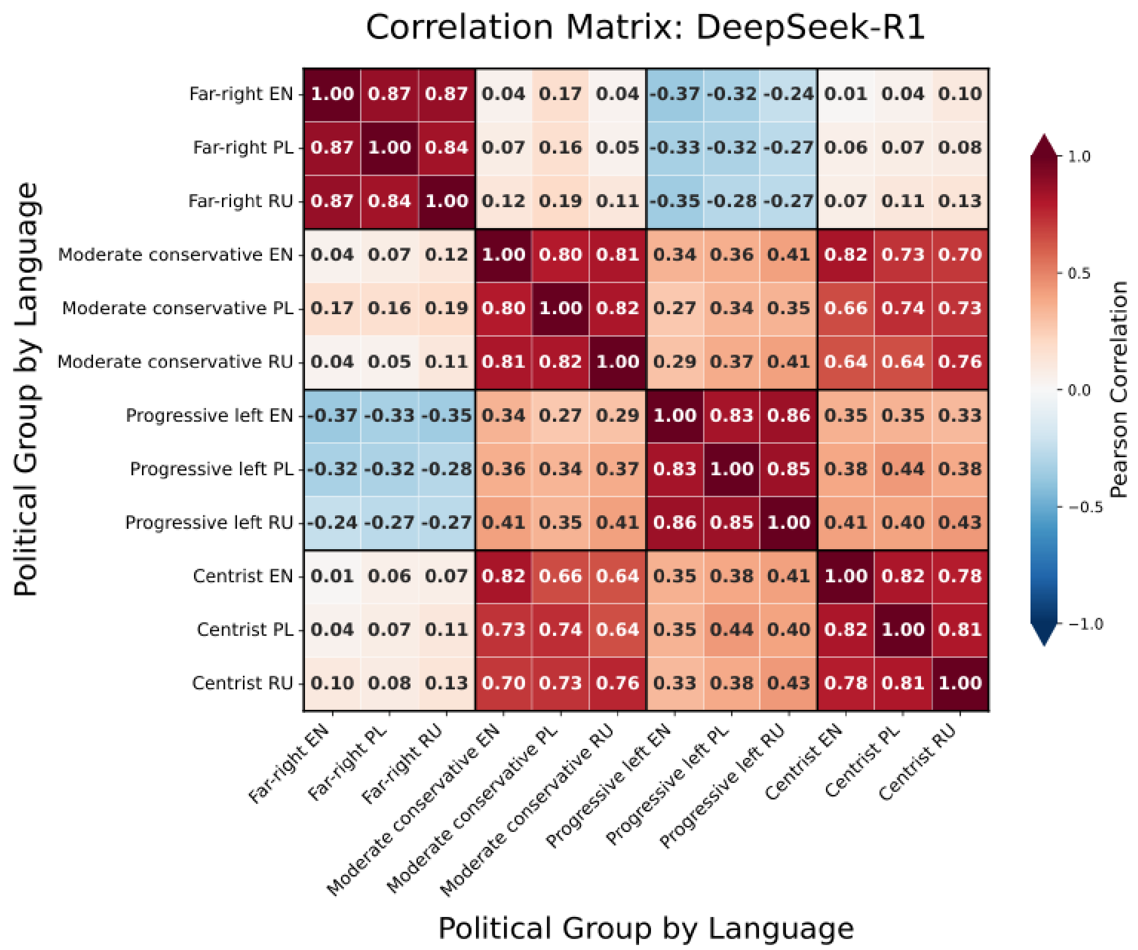
    Qwen3-8B: 22%

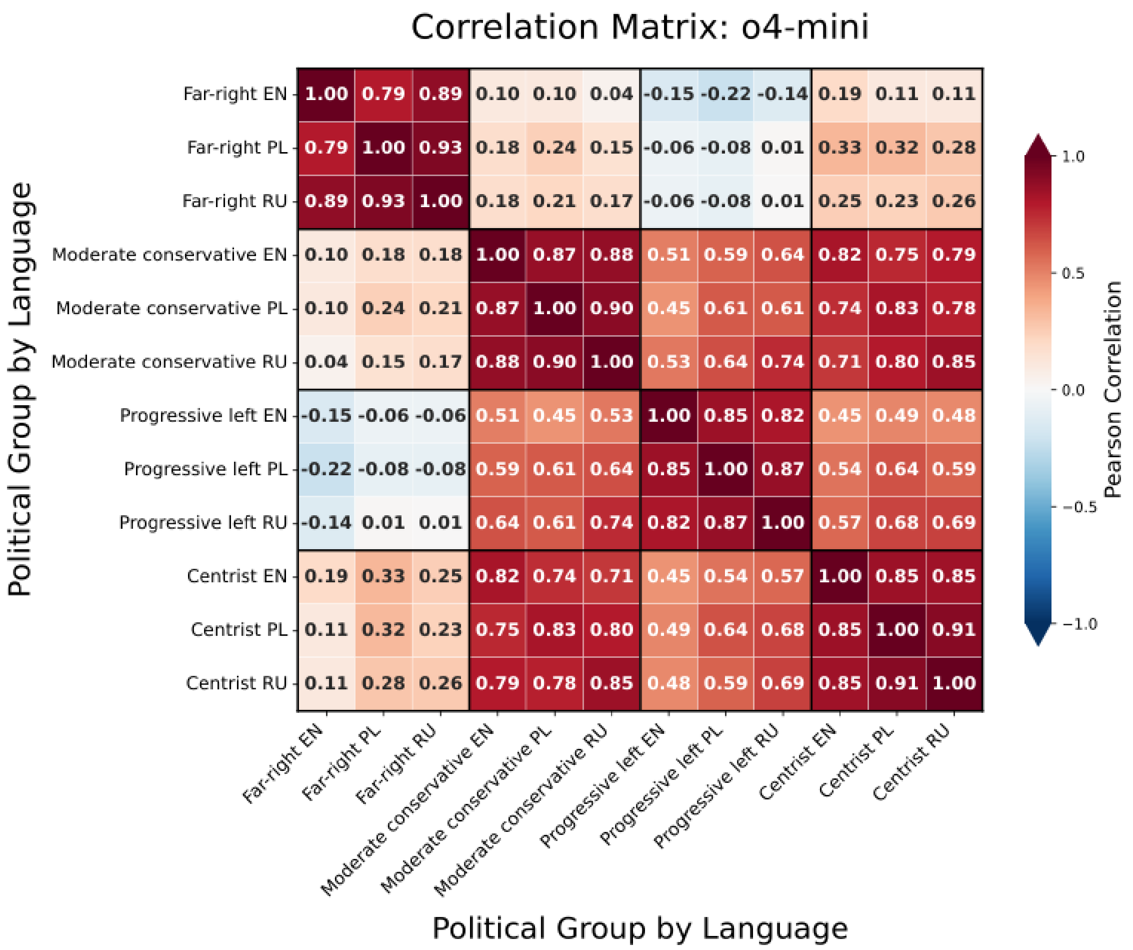
5. Aggregating statistically sufficient data:

    $\hat{p}_{n_k} \in \{0.0, 0.2\}$ — label 0

    $\hat{p}_{n_k} \in \{0.8, 1.0\}$ — label 1

Key results (large reasoning models)





Metrics

- Cross-Language Consistency (CLC):

  high values mean more variation across languages

$$CLC = \frac{\sum_{i=1}^{4} \sum_{j=i}^{4} Var(C_{ij})}{10^{-3} \sum_{i=1}^{4} \sum_{j=i}^{4} 1}$$

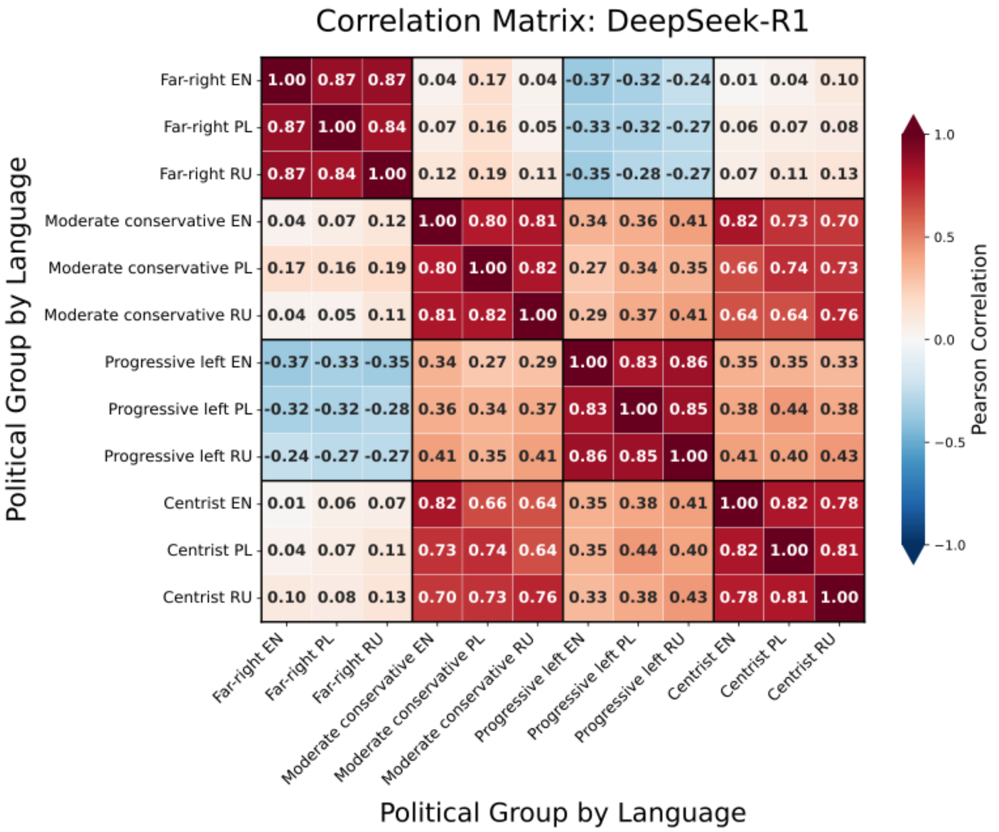
- Inter-Group Differentiation (IGD):

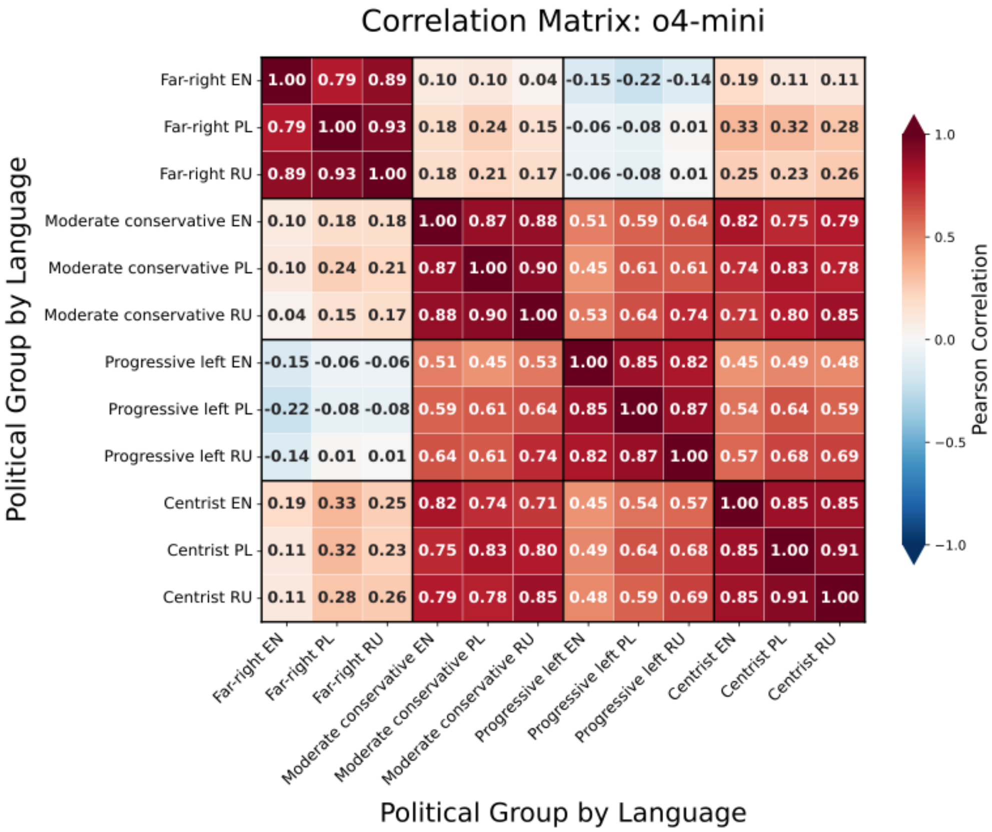
  high values mean better ideological separation

$$IGD = \frac{Var(\{Mean(C_{ij}) \mid 1 \leq i < j \leq 4\})}{10^{-3}}$$

where C_{ij} is the 3x3 block of correlation values between groups i and j.

Key results (large reasoning models)



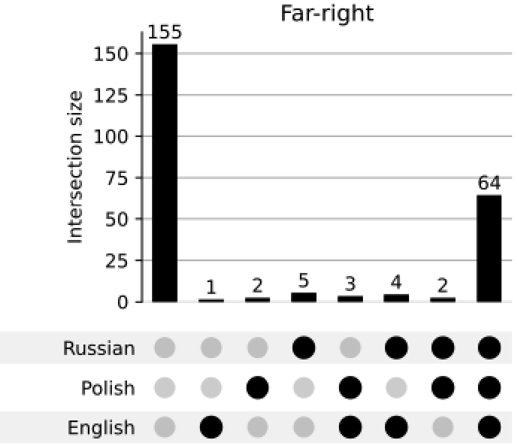


METRICS:

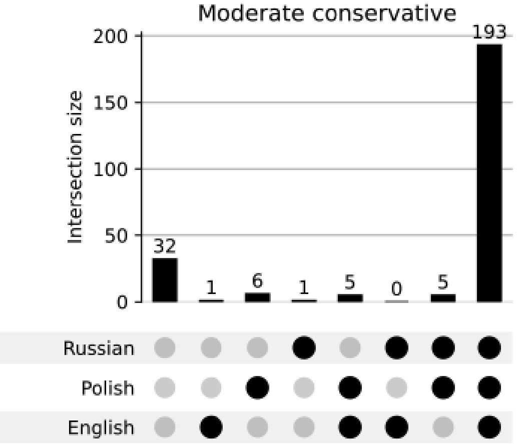
CLC (Cross-Language Consistency) - measures the variability of correlations within and between political groups across different languages.

IGD (Inter-Group Differentiation) - measures how distinct the model's responses are between different political groups, based on the average correlation values.

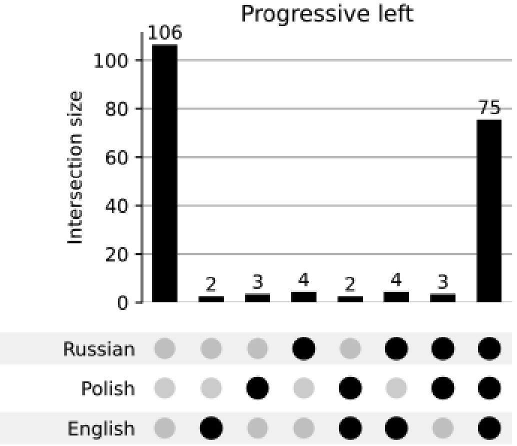
DeepSeek-R1 — languages disagreement



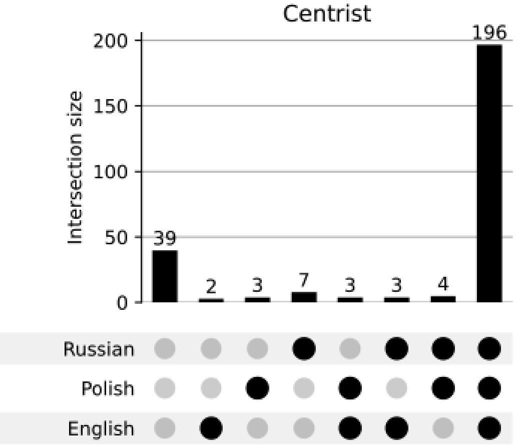
(a) Far-right group.



(b) Moderate conservative group.

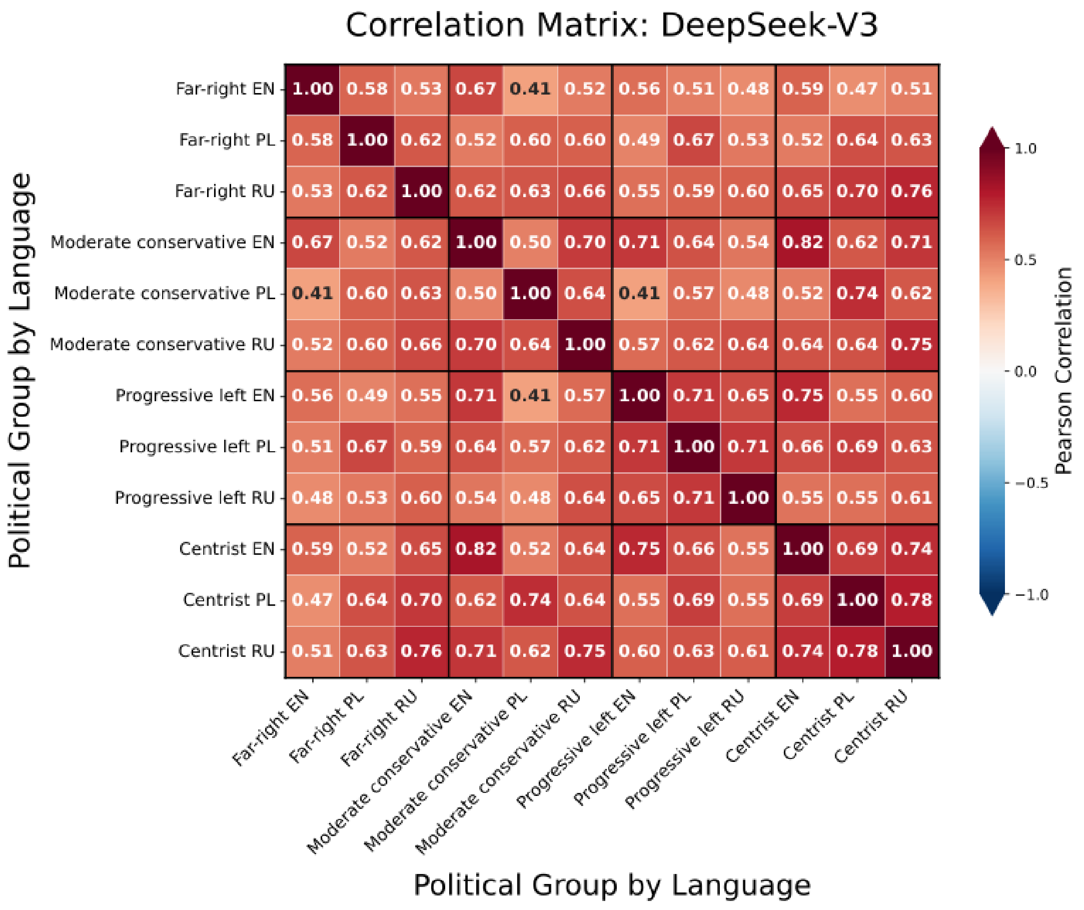


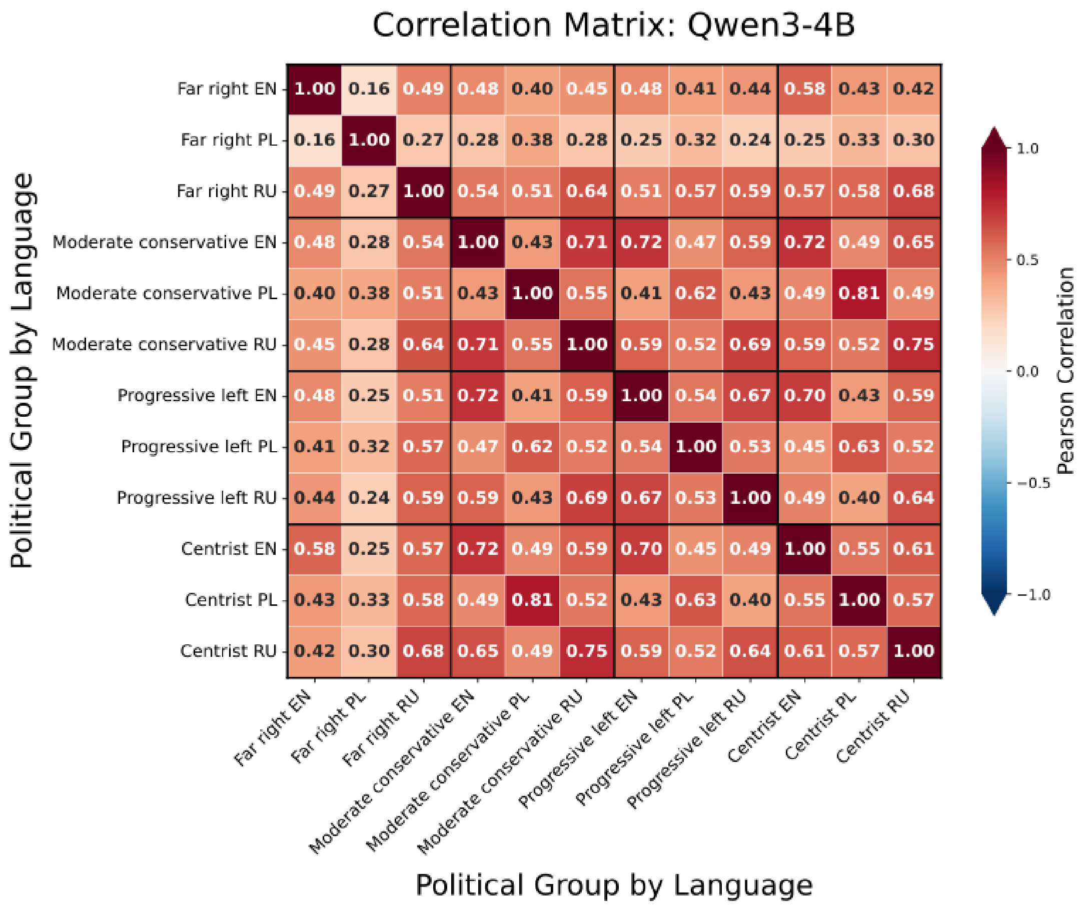
(c) Progressive left group.



(d) Centrist/Independent group.

Key results (large non-reasoning models)





Reasoning language

DeepSeek-R1

a. English: 86%
b. Russian 14%

OpenAI o4-mini

a. English: 1.4%

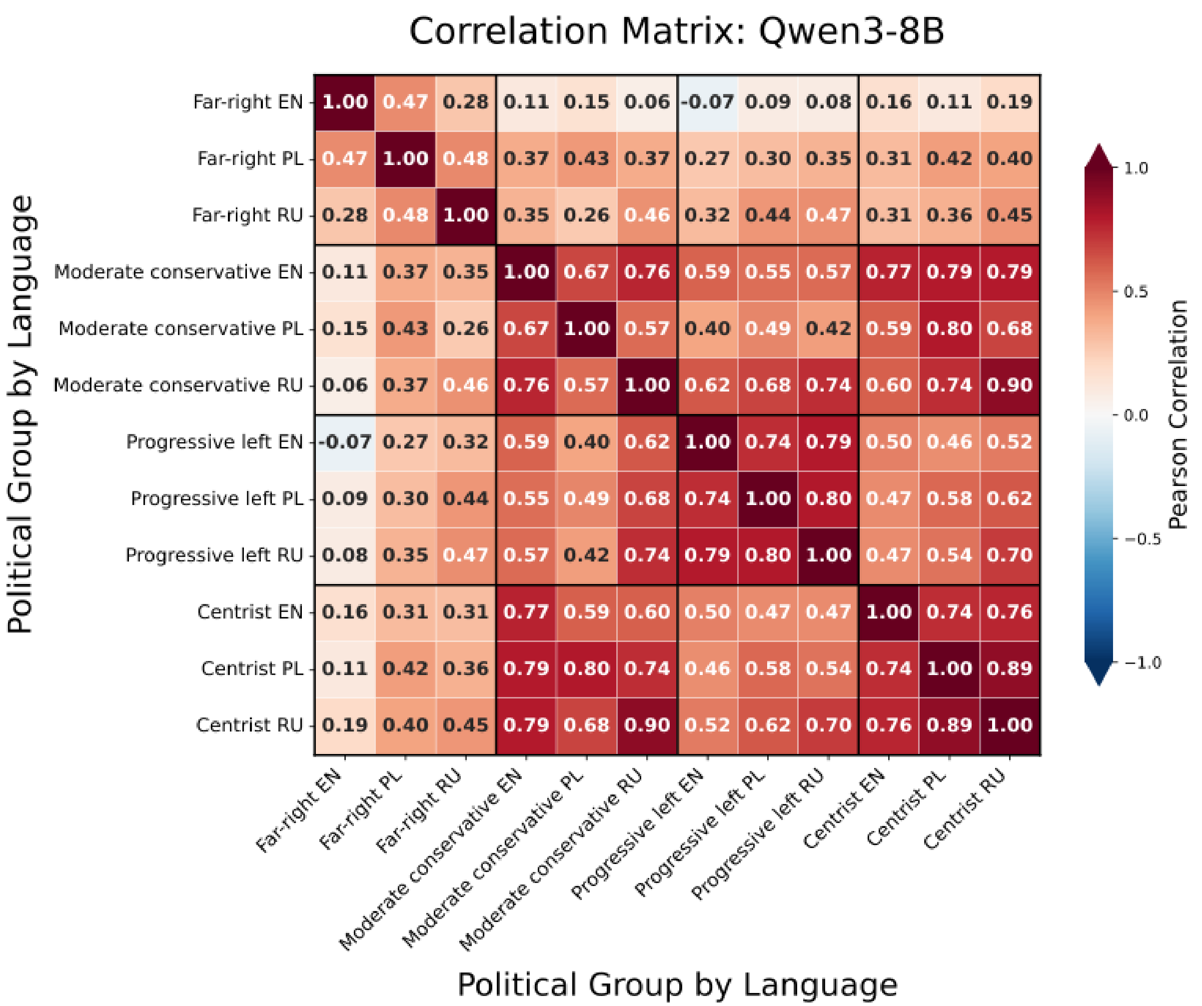
Qwen3-8B

a. English: 62.8%
b. Polish: 4.2%
c. Russian: 33.0%

Example DeepSeek-R1 reasoning trace:

```
Okay, let me try to figure out how Mark Thompson would label this statement. First, the user's statement mentions Kurds helping Ukrainians resist Putin and the Russian military. Then it criticizes Trump for cutting off aid to Ukraine to pressure their president for election help, leading to his impeachment. **Mark is a far-right conservative who distrusts mainstream media and sees criticism of right-wing policies as unfair attacks**. He probably downplays offensive right-wing tweets but calls progressive ones aggressive. So, the statement criticizes Trump, a right-wing figure, and supports the idea that his actions were impeachable. From Mark's viewpoint, this could be seen as an unfair attack on a right-wing leader, thus labeling it as offensive (1). He might also dismiss the part about Kurds if it's against Russian actions, but the main issue is the Trump criticism. So, the answer is likely 1.
```

Key results (small reasoning model)



Quantitative comparison

	DeepSeek-R1	OpenAI’s o4-mini	DeepSeek-V3	Qwen3-8B	OpenAI’s GPT-4.1-mini	Qwen3-4B	Gemma3-4B-IT	Mistral-7B-Instruct-v0.3
Category	**Big reasoning**	**Big reasoning**	**Big non-reasoning**	**Small reasoning**	**Small non-reasoning**	**Small non-reasoning**	**Small non-reasoning**	**Small non-reasoning**
Percentage of valid responses (%)	90.7	90.7	100	77.4	100	100	100	100
Cross-Language Consistency (CLC)	**3.92**	4.85	15.31	22.2	12.32	33.43	28.29	65.64
Inter-Group Differentiation (IGD)	**100.03**	89.28	1.58	32.23	8.09	4.77	3.46	1.37

Key takeaways

- Reasoning capabilities are essential for personalized offensiveness detection

- Model size alone isn’t enough

Future directions

- Using ground truth in the evaluation
- A more detailed consideration of the reasoning depth of the LLMs
- Statistical robustness
- Unrealistic personas

Thank you

Dzmitry Pihulski, B.Eng. dzmitry.pihulski@pwr.edu.pl
Jan Kocoń, PhD jan.kocon@pwr.edu.pl

Wrocław University of Science and Technology, Poland