

The BERT probabilities of the tokens *agree* and *disagree* are correlated and we can exploit it.

How Gender Interacts with Political Values

Adnan Al Ali and Jindřich Libovický

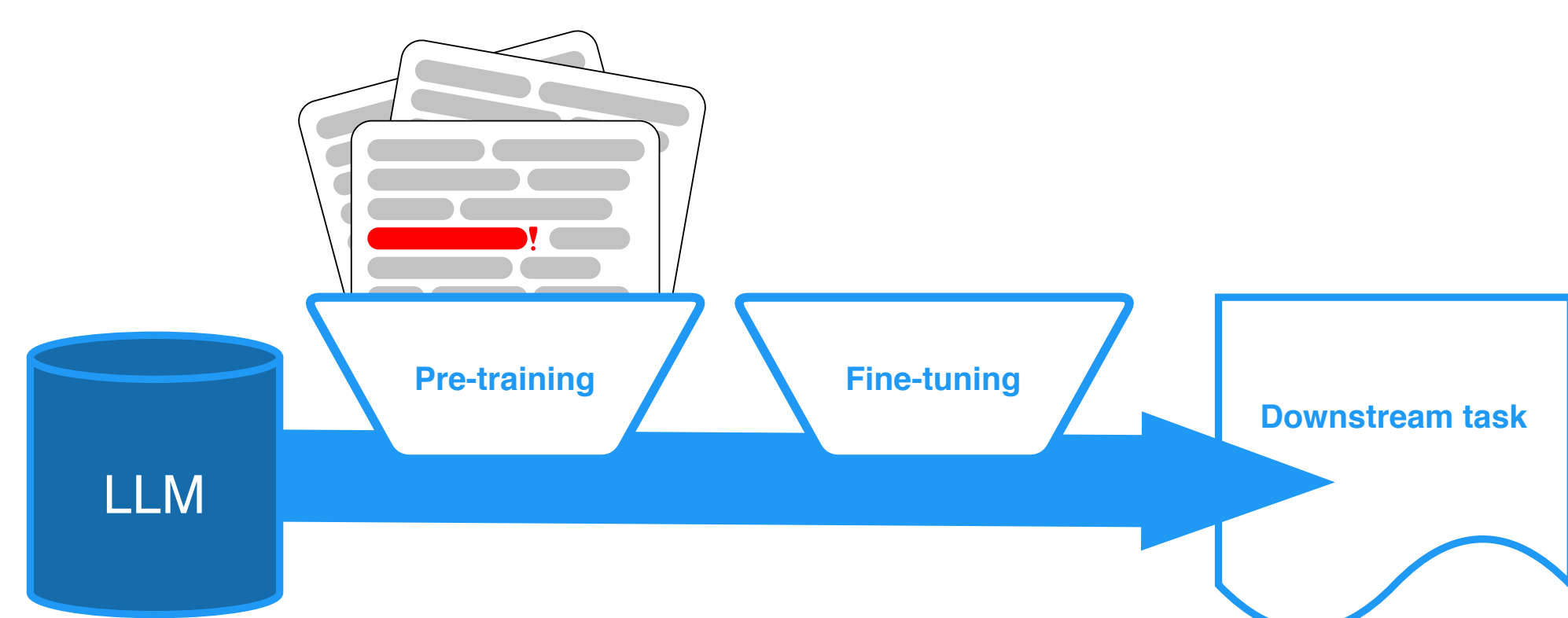
adnan@matfyz.cz libovicky@ufal.mff.cuni.cz

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic



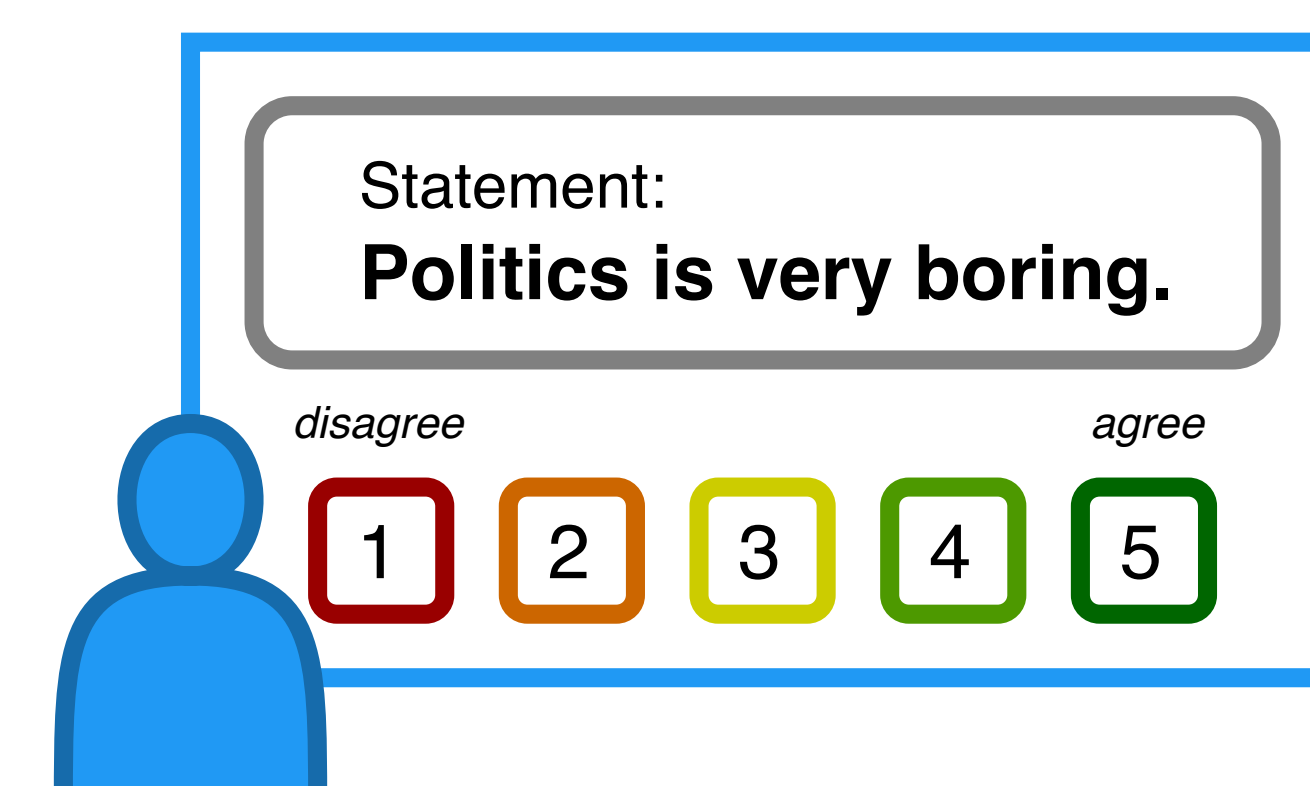
CHARLES UNIVERSITY

Motivation

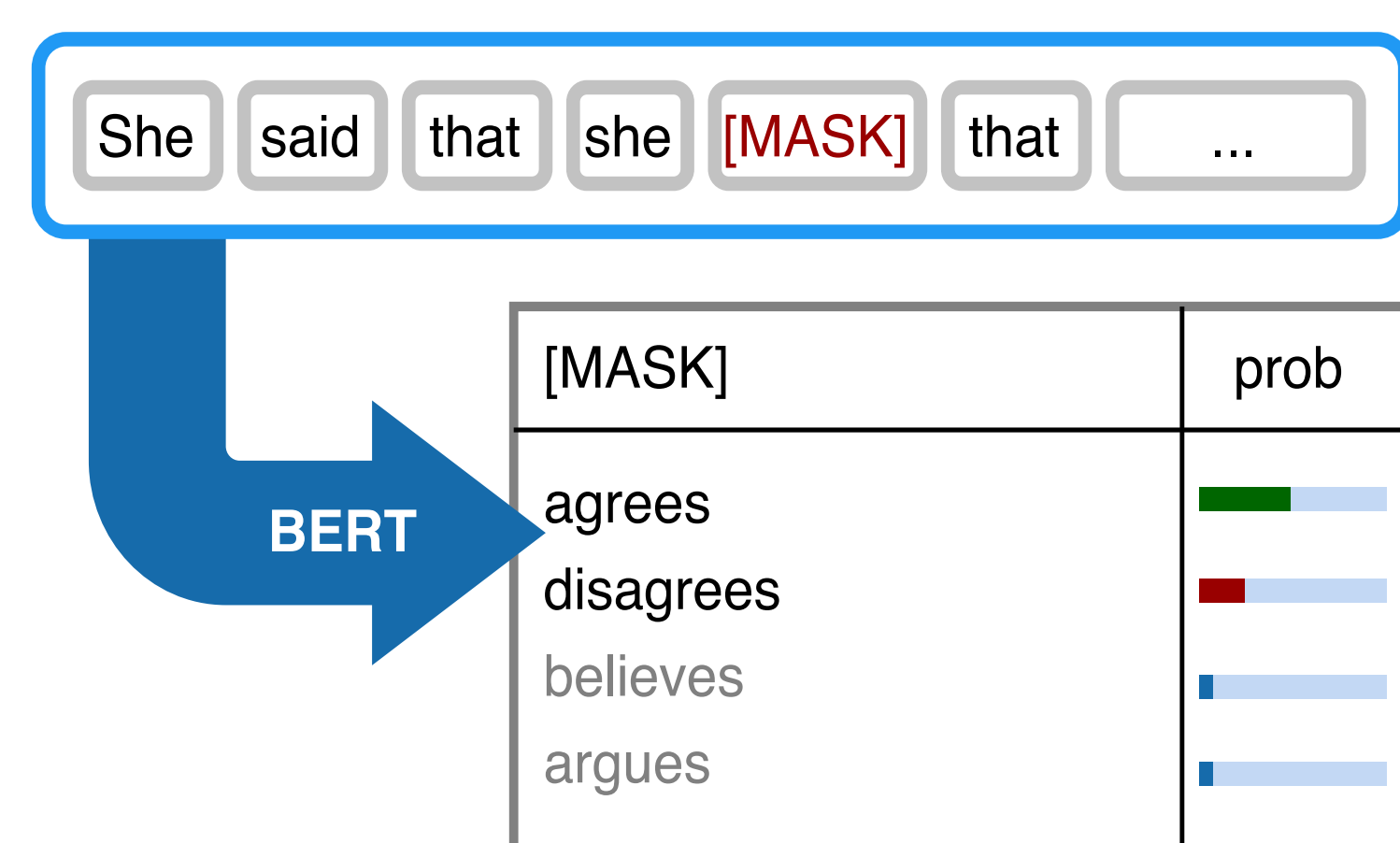


- Pre-training corpora often contain **unmoderated content**.
- **Gender bias and political values**: how do they interact?

Respondents in political surveys use a **1 – 5 dis/agree scale**. • We want to get comparable ratings from LMs. •



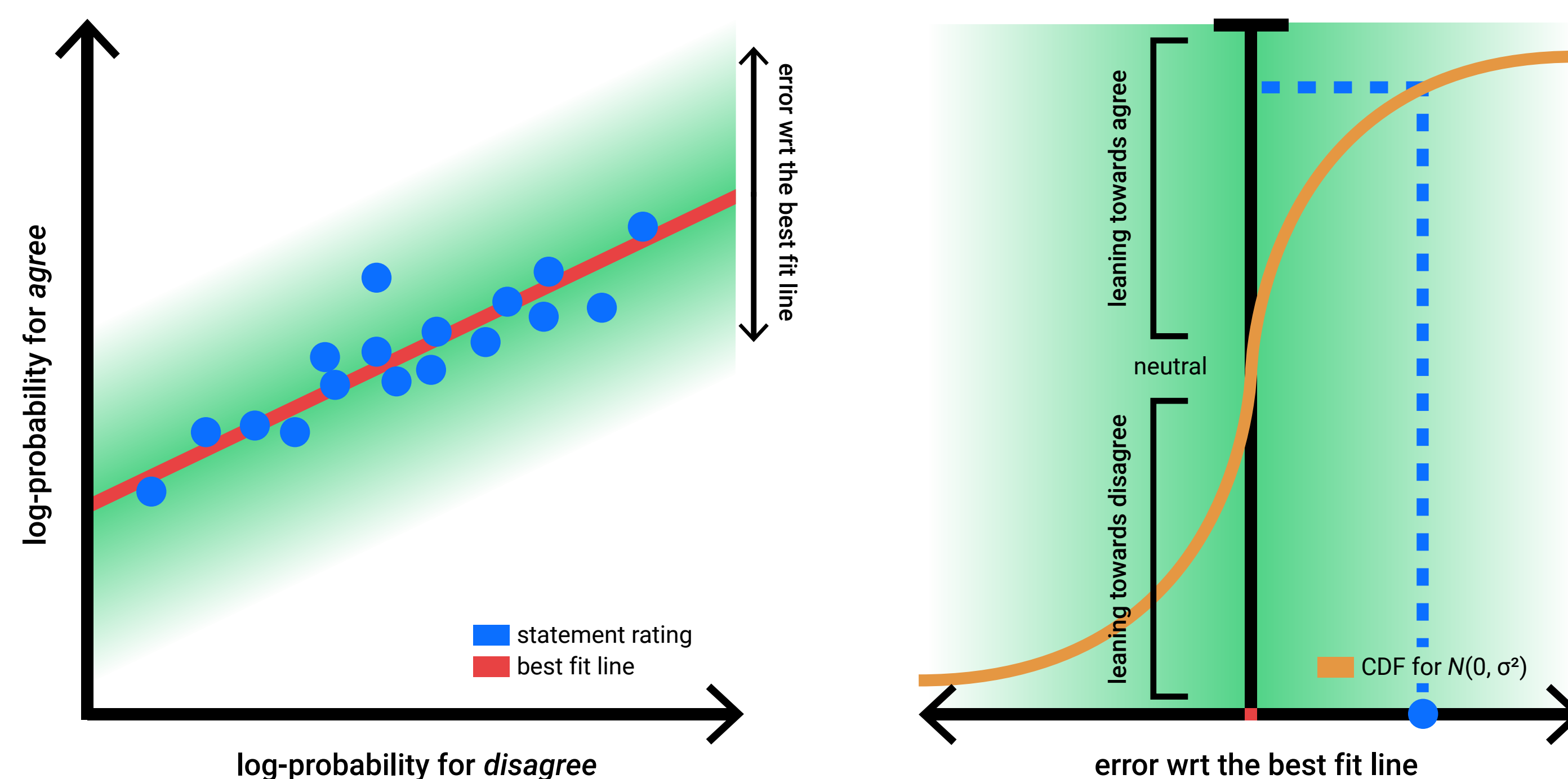
Methodology



- We propose **template-based prompts**.
- We **mask** the tokens for the words **agree/disagree**.
- **Four templates** per statement: each combination of feminine/masculine and agree/disagree.
- **Gender** only expressed **grammatically**.

	Fem
Agree	[CS] Rekla, že souhlasí s tím, že --- <i>She said that she agrees that ---</i>
Disagree	[CS] Rekla, že nesouhlasí s tím, že --- <i>She said that she disagrees that ---</i>

- We work with the **logarithms** of the probabilities.
- **Agree** is usually **rated higher** than disagree, regardless of the statement.
- The (log-)probabilities for agree and disagree are **correlated**.
- We introduce **apolitical calibration data** to estimate the **best fit line** and **variance** of the error w.r.t. the line.
- Ratings below and above the line are considered as disagreeing and agreeing respectively.
- The exact **rating** is calculated using the **CDF for $N(0, \sigma^2)$** .



Results

Model	Average Rating								Standard deviation							
	AntiAuth		CultLib		EconEq		Trib		AntiAuth		CultLib		EconEq		Trib	
	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
Survey mean*	3.3	3.8	3.9	4.0	3.1	3.1	3.2	3.2	0.8	0.7	0.7	0.6	0.6	0.6	0.6	0.5
RobeCzech	3.0	3.1	2.9	2.9	3.3	3.2	3.3	3.4	1.7	1.7	1.2	1.1	1.3	1.3	1.4	1.3
Czert	2.7	2.7	2.7	2.7	3.2	3.2	3.2	3.1	1.7	1.7	1.1	0.9	1.4	1.1	1.7	1.5
FERNET News	3.5	3.0	2.8	3.1	3.8	3.1	3.7	3.4	0.8	1.1	1.0	0.9	1.7	1.3	1.3	1.9
mBERT	3.8	3.9	2.3	2.4	3.9	3.8	3.5	3.6	1.2	1.2	1.2	1.3	0.8	0.8	1.4	1.3
Slavic BERT	3.9	3.9	3.2	3.0	3.5	3.7	3.0	3.1	1.0	1.0	1.3	1.3	1.0	0.9	0.9	0.9
XLm-R	3.3	3.3	3.2	3.2	4.1	4.1	3.3	3.3	0.7	0.6	1.5	1.6	1.2	1.3	1.1	1.2

- We **compare** the obtained ratings to a real-life **political values study** of polled on Czech-speaking people, divided by gender.
- Most models made **little distinction** between the **masculine** and the **feminine** sentences, although the ratings differ in the real-life data
- All models **underestimated** the rating of **cultural liberalism**.
- All models **overestimated** the rating of **economic equity**.
- **mBERT** had the **strongest opinions**.
- Many ratings are close to the **midpoint** of the scale, with a **large variance**.

Conclusions

Most models made little to **no distinction** between the **feminine** and the **masculine** sentences.
Most models rated the sentences corresponding to the same value **inconsistently**, leading to a **large variance**.

We did not find any significant systematic perceived political values in the models.



Presented at LREC-COLING 2024

The work was supported by the Charles University project PRIMUS/23/SCI/023. The work described herein has been using services provided by the LIN-DAT/CLARIAH-CZ Research Infrastructure (<https://lindat.cz>), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).