Book of Abstracts of the 5th Workshop on Computational Linguistics for Political Text Analysis (CPSS-2025)
Hildesheim, Germany

Workshop chairs: Dennis Assenmacher, Christopher Klamm, Gabriella Lapesa,
Ines Rehbein, Simone Paolo Ponzetto, Indira Sen

Sep 10-11, 2025

What is Democracy? Exploiting Situation Entities to Uncover Democracy Frames in German Political Discourse

Julian Schlenker,[1,*] **Ines Rehbein,**[1] **and Simone Paolo Ponzetto**[1]
[1]*Data and Web Science Group, University of Mannheim, Germany*
*Corresponding author: *julian.schlenker@uni-mannheim.de*

Political discourse frequently involves contested and contextually evolving concepts. The notion of democracy (*Demokratie* in German) exemplifies such a concept, exhibiting significant variability in framing across political parties and over time [1]. As a work in progress, this abstract presents preliminary insights into utilizing Situation Entities to investigate how the concept of democracy is framed within German parliamentary debates.

Situation Entities (SitEnts) are clause-level semantic annotations that introduce *situations* to the discourse [2]. Smith [2] distinguishes between Events, States, Generics, Generalizing sentences, Facts, Propositions, Questions and Imperatives. From a political text analysis perspective, two SitEnt types, Generalizing and Generic are especially valuable as they often encapsulate stereotypes or conceptual definitions. In this abstract, we focus on the latter and hypothesize that SitEnts can be leveraged to uncover frames that seek to define *democracy*.

To test this, we construct a human-annotated dataset derived from speeches of the 19th legislative term of the German Bundestag. As the annotation of SitEnts is challenging and only yields moderate agreement (see, e.g., [3]), we consider all annotations as valid labels and formulate the task as a multi-label classification task. Using this dataset, we fine-tune a pre-trained German BERT model[1] and train a classification head to perform sequence classification. A preliminary evaluation of our model, using a probability threshold of 0.5 for predictions, yielded promising F1 scores ranging from 0.57 to 0.96 across all labels (macro-average F1: 0.74). For the label of interest, Generic, we obtain an F1 score of 0.73.

The trained classifier is subsequently employed to automatically label a comprehensive corpus of Bundestag debates spanning from the 1st to the 19th legislative term [4], specifically extracting passages containing the term *Demokratie*. Subsequently, instances labeled as Generic are selected for detailed semantic exploration. For each instance, we extract the syntactic subtree governed by the verb evoking the SitEnt, encode them using a multilingual sentence transformer model[2] and cluster them via the Fast Clustering algorithm from the Sentence Transformers library [5].

Initial qualitative analysis confirms that several clusters contain consistent patterns indicative of explicit definitions of democracy, underscoring the potential of this method. For instance, specific clusters recurrently highlight democracy in terms of core values, emphasizing principles such as participation, governance structures and fundamental rights (see Table 1 for Examples). Among the clustered instances, the two largest clusters stand out: Cluster 1 predominantly defines democracy by highlighting what it requires or benefits from, while Cluster 2 defines it through what it endangers or stands in contrast to. In the terminology of Subramonian et al. [6], these clusters reflect the overarching themes of *Necessary/Beneficial* and *Dangers*, respectively.

These findings provide preliminary evidence supporting the viability of leveraging SitEnts as a tool for nuanced political text analysis at scale. Moving forward, we plan to systematically explore how democracy frames differ according to party affiliation and how they evolve over time. Furthermore, our analysis will extend beyond democracy to encompass additional political concepts, and we will investigate the full range of SitEnt types.

[1] `https://huggingface.co/deepset/gbert-large`
[2] `https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1`

References

[1] Alexander Alekseev. The (changing) concept of democracy in (transforming) European populist radical right discourses: the case of polish law and justice. *Journal of Contemporary European Studies*, 33(1):121–141, 2025. doi: 10.1080/14782804.2024.2360635. URL https://doi.org/10.1080/14782804.2024.2360635.

[2] Carlota S. Smith. *Modes of Discourse: The Local Structure of Texts*. Cambridge Studies in Linguistics. Cambridge University Press, 2003.

[3] Annemarie Friedrich and Alexis Palmer. Situation entity annotation. In Lori Levin and Manfred Stede, editors, *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 149–158, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. doi: 10.3115/v1/W14-4921. URL https://aclanthology.org/W14-4921/.

[4] Andreas Blaette. Germaparl. Corpus of plenary protocols of the German Bundestag, 2017. TEI files, available at: https://github.com/PolMine/GermaParlTEI.

[5] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

[6] Arjun Subramonian, Vagrant Gautam, Dietrich Klakow, and Zeerak Talat. Understanding "democratization" in NLP and ML research. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3151–3166, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.184. URL https://aclanthology.org/2024.emnlp-main.184/.

[7] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022. URL https://arxiv.org/abs/2203.05794.

A Clustering Examples

Cluster ID (n **Elements)**	**Top-3 Keywords**	**Verb Subtree Examples**
Cluster 1 (491 Elements)	*live, mean, need*	*Democracy lives through everyone participating and contributing.* *Democracy means distribution of power.* *Democracy needs firm principles, but also compromise.*
Cluster 2 (75 Elements)	*destroy, lose, harm*	*to destroy our democracy* *In the end, democracy loses when peaceful protest is not possible.* *to harm democracy*
Cluster 10 (15 Elements)	*parliamentary, parliament, election campaign*	*A democracy only functions with a working parliamentary opposition.* *Parliaments without influence always weaken democracy.* *Election campaigns are part of parliamentary democracy.*
Cluster 16 (13 Elements)	*liberal, freedom, freedom of expression*	*A liberal democracy thrives on open and public discussion.* *The freedom of those who think differently is a defining feature of democracy.* *Democracy, the temporary delegation of power, is only conceivable through freedom of expression.*
Cluster 19 (13 Elements)	*freedom of the press, press, free*	*Without freedom of the press, there is no democracy.* *Freedom of the press and opinion are cornerstones of any democracy.* *Free art and free press are pillars of democracy.*

Table 1: Top-3 keywords and example verb subtrees for selected clusters. We retrieve 40 clusters in total using a minimum cluster size of 10. Out of 2139 GENERICS, 1012 are clustered into one of the clusters. We use c-TF-IDF [7] to extract top-k keywords. All keywords and verb subtree examples have been translated from German into English using GPT-4o accessed through ChatGPT (`https://chatgpt.com`).

Are Epistemic Norms in Decline? A Comparative Analysis of Parliamentary Discourse

Segun T. Aroyehun,[1*] Fabio Carrella,[2] and Stephan Lewandowsky,[3] and David Garcia[1]

[1] *University of Konstanz, Konstanz, Germany*
[2] *University of Campinas, Campinas, Brazil*
[3] *University of Bristol, United Kingdom*
*Corresponding author: *segun.aroyehun@uni-konstanz.de*

Truth is fundamental for governance, accountability, and informed decision-making in democratic societies. It also fosters intellectual honesty and social cohesion [1]. Truth exists along a multidimensional continuum [2, 3]. This continuum encompasses evidence-based reasoning, rooted in concrete facts and data, at one end, and intuition, driven by gut feelings and subjective interpretations, at the other. Recent computational analysis of congressional discourse in the United States (from 1879 to 2022) shows a decline in evidence-based language since the mid-1970s [4], raising concerns about the weakening of epistemic norms such as truth-seeking and justification that underpin public deliberation in liberal democracies [5]. It also raises deeper questions about whether democratic institutions (particularly parliaments) are maintaining these norms as emblematic spaces of reasoned discourse, or whether parliamentary debate is increasingly dominated by appeals to intuition, emotion, and identity at the expense of evidence and justification. To the best of our knowledge, this question has received very little empirical treatment.

Accordingly, this study asks whether the declining trend in evidence-based language in parliamentary discourse is unique to the United States or indicative of a broader shift. We compare parliamentary discourse across three Western democracies—the United States, Germany, and Italy. These countries differ in institutional design, political culture, and democratic stability. While the United States has the longest uninterrupted democratic history, Germany and Italy experienced democratic breakdowns and postwar democratization. Despite these differences, analyzing discourse along epistemic dimensions provides insight into the quality of deliberations over time. We link this analysis to the Deliberative Democracy Index (DDI) from the Varieties of Democracy (V-Dem) project [6], which provides an expert-based assessment of the extent to which public discourse is grounded in reasoned justification.

We use transcripts of parliamentary speeches from the United States, Germany, and Italy (spanning 1861 to 2022) [7, 4, 8, 9, 10]. We adopt a computational text analysis approach proposed in [4]. We train temporal word embeddings on country-specific corpora which are segmented by major historical periods. For each speech, we compute an Evidence-Minus-Intuition (EMI) score, based on the cosine similarity between the speech embedding and the centroids of evidence- and intuition-related words. A positive EMI score reflects a more evidence-based orientation while a negative score suggests intuition-driven rhetoric.

We find that the decline in EMI score in recent decades is unique to the United States. In contrast, we observe rising trends in Germany and Italy (See Figure 1). Historical declines in Germany and Italy align with periods of democratic crisis (i.e., the rise of Fascism in Italy and Nazism in Germany), while the post-war upward trends coincide with democratic consolidation. The recent decline in the United States occurs in parallel with a drop in the DDI, suggesting erosion in deliberative quality. Further statistical analyses corroborate this interpretation. Using a lagged dependent variable model with country fixed effects to examine the relationship between the EMI and DDI series, we find that the DDI is a positive and significant predictor of the EMI. In alternative formulations with country-specific models, the DDI remains a positive predictor of EMI.

Across all three countries, we find a consistent association between higher EMI scores and greater perceived deliberative quality (DDI) in public discourse. These findings raise broader questions about how and when parliaments uphold norms of truth-seeking and justification, and whether similar patterns can be observed in other democratic contexts or institutional settings.

References

[1] E Tory Higgins, Maya Rossignac-Milon, and Gerald Echterhoff. "Shared reality: From sharing-is-believing to merging minds". In: *Current Directions in Psychological Science* 30.2 (2021), pp. 103–110.

[2] Stephan Lewandowsky. "Deliberate Ignorance: Choosing Not to Know". In: The MIT Press, 2020. Chap. Willful construction of ignorance: A tale of two ontologies.

[3] Binyamin Cooper et al. "Honest behavior: Truth-seeking, belief-speaking, and fostering understanding of the truth in others". In: *Academy of Management Annals* 17.2 (2023), pp. 655–683.

[4] Segun T Aroyehun et al. "Computational analysis of US congressional speeches reveals a shift from evidence to intuition". In: *Nature Human Behaviour* (2025), pp. 1–12.

[5] Aurélia Bardon et al. "Disaggregating civility: Politeness, public-mindedness and their connection". In: *British Journal of Political Science* 53.1 (2023), pp. 308–325.

[6] Michael Coppedge et al. *V-Dem Country-Year Dataset v15*. `https://www.v-dem.net`. Version 15. 2025.

[7] Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. *Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts*. 2018. URL: `https://data.stanford.edu/congress_text`.

[8] Giuseppe Abrami et al. "German Parliamentary Corpus (GerParCor)". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 1900–1906. URL: `https://aclanthology.org/2022.lrec-1.202/`.

[9] Florian Richter et al. "Open discourse: towards the first fully comprehensive and annotated corpus of the parliamentary protocols of the German Bundestag". In: *OSF preprint* (2023).

[10] Valentino Frasnelli and Alessio Palmero Aprosio. "There's Something New about the Italian Parliament: The IPSA Corpus". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 16037–16046. URL: `https://aclanthology.org/2024.lrec-main.1394/`.

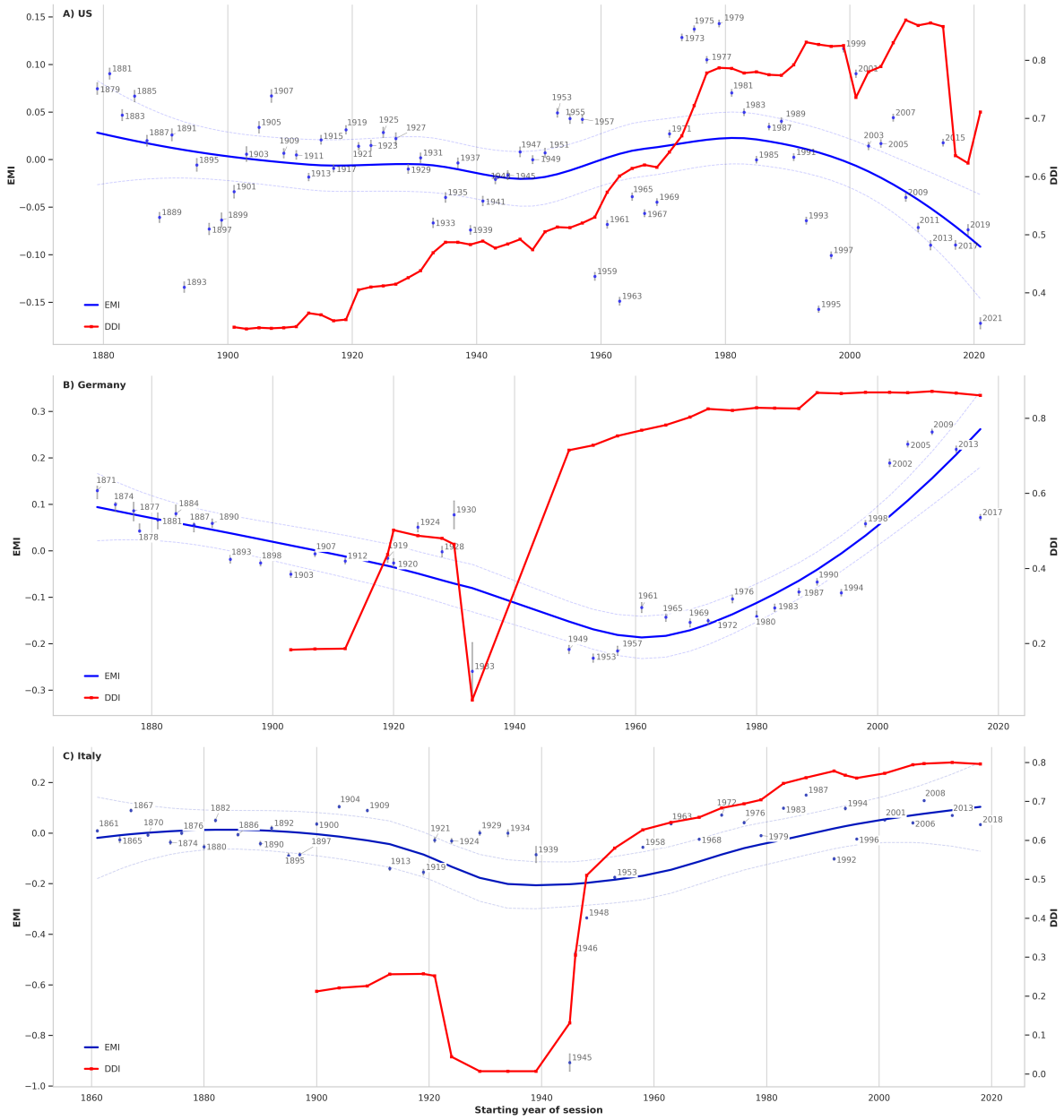

Figure 1: Trends in Evidence-Minus-Intuition (EMI) scores and the Deliberative Democracy Index (DDI) across parliamentary sessions in the United States (A), Germany (B), and Italy (C). Blue dots represent the average EMI score for each parliamentary session, with vertical bars indicating 95% confidence intervals for the session mean. The blue trend line shows a LOESS smoothed fit of EMI scores, with dashed lines indicating 95% confidence intervals of the fit. Red lines indicate the DDI trend, based on the DDI value for the year marking the beginning of each session.

How Stable Are Political Ideology Classifiers Over Time? An Empirical Evaluation of Temporal Robustness in NLP Models

Mohsin Khan,[1,*] **Praveen Gupta**[2]
*Corresponding author: *this.mohsin@gmail.com*

Most political NLP models are tested using random splits that ignore how language changes over time. We are examining what happens when you train ideology classifiers on old political texts and test them on newer ones. Our work will use manifestos and parliamentary speeches from several countries to see how classifier performance degrades as the time gap grows.

Our study will analyze political texts from party manifestos (2000-2020) and parliamentary speeches (2005-2022) across five countries. We will implement strict temporal splitting where models are trained on data before a cutoff year and tested on data from subsequent years with varying time gaps (1-8 years).

We will compare three model architectures: BERT-based classifiers, RoBERTa-based classifiers, and Logistic regression baselines. Performance will be measured using accuracy, F1 score, and class-specific metrics across different temporal gaps.

We expect to find substantial accuracy drops over longer periods based on preliminary analysis. The degradation will likely appear roughly linear with time, and transformer models should maintain better performance than traditional approaches but still show significant decline.

Feature analysis will be performed using SHAP values to track how specific political terms shift meaning - words like "reform" and "security" don't stay associated with the same ideological camps. We will examine what percentage of discriminative features remain stable versus those that show complete reversal in association.

Different ideological classes will be analyzed for varying temporal patterns.

This research exposes problems with how we currently evaluate political text classifiers and questions whether models trained on historical data can reliably analyze contemporary politics. The findings will push for better evaluation methods that account for language evolution in political discourse and provide methodological recommendations for more robust evaluation of political NLP systems.

References

1. Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica, 87*(4), 1307–1340. `https://doi.org/10.3982/ECTA16566`
2. Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297.
3. Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.
4. Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review, 97*(2), 311–331.
5. Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication, 29*(2), 205–231.
6. Jiang, H., Zhang, D., Cao, T., Yin, B., & Zhao, T. (2021). Named entity recognition with small strongly labeled and large weakly labeled data. *arXiv preprint arXiv:2106.08977*.
7. Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond binary labels: Political ideology prediction of Twitter users. In *ACL 2017*, 729–740.
8. Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. *arXiv preprint arXiv:1806.03537*.
9. Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PLOS ONE, 11*(12), e0168843.
10. Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *TACL, 6*, 587–604.
11. Rauh, C., & Schwalbach, J. (2020). The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in nine representative democracies. *Harvard Dataverse, 1*(1).

Leveraging Survey-Informed LLM Personas to Identify Cultural References : Enhancing Inclusivity in News Reporting

Reshmi Pillai, Wouter van Atteveldt, Antske Fokkens
Vrije University Amsterdam
Corresponding author: r.pillai@vu.nl

In a globalized world, newsrooms face the challenge of making their reporting accessible to diverse audiences, including non-native residents. Especially in multi-cultural regions with ethnic and socio-economic diversity, media plays a critical role in 'bridging and bonding' the migrant groups with society [1]. However, individuals from migrant backgrounds often struggle to effectively comprehend and utilize the content of regional news outlets. News reports frequently assume that readers possess prior knowledge about local cultural norms, nuances, historical/prominent figures, locations and monuments and the non-native readers might lack this familiarity. Such cultural contexts embedded in news stories often complicate understanding for those unfamiliar with local nuances.

The first step towards achieving this inclusion would be to equip journalists with a mechanism to identify news content that might be unfamiliar to non-native readers. Ideally, journalists from diverse ethnic or cultural backgrounds can bring different perspectives and thus attract different audiences [2]. However, with the prevailing lack of demographic diversity in workforce, news organizations often struggle to effectively understand and address the needs of a multi-cultural, multi-ethnic audience. We seek to explore how can AI support newsrooms to address diversity in news audience, while conserving the agency of journalists in the news-making process.

Specifically, we explore this research question: **Can we use survey-informed LLM personas to simulate non-native readers in identifying culturally relevant information in news reports?**

We follow a novel solution approach. We address this research question through the use case of dutch language news reports. Given the scarce information available about the cultural hindrances of non-native readers in comprehending news, we conduct a survey using convenience sampling, to gather characteristics of such readers and the hurdles they face. The responses from this survey are analysed to create clusters of respondents each of which is further represented through a persona. Using NexisUni, a widely used online archive for news resources, we collected a dataset of 500 news reports published in the Netherlands during the period of 01.04.2023 to 01.04.2025. For the purpose of analysis in the current study, we randomly chose 100 news passages, from the reports, and report the annotations of culturally relevant phrases in each of these passages using 1) personas of the clusters implemented using three different Large Language Models (GPT-4o, LLaMA-3-8B and Mistral-7B-Instruct-v0.2) 2) a baseline method using Named Entity Recognition 3) and a human coder. Jaccard similarity analysis revealed higher consistency between GPT-4o and LLaMA models across personas, with Mistral showing relatively lower similarity with other LLMs. Further results in comparison with annotations of cultural references identified by a human coder showed that larger models such as GPT-4o achieved strong performance (precision = 0.75, recall = 0.98, F1-score = 0.85), while smaller models like Mistral-7B produced more variable outputs (e.g., precision = 0.32, recall = 0.98, F1-score = 0.48). While these findings demonstrate the potential of persona-driven prompting to influence model outputs, they also underscore limitations, including the small survey sample (39 respondents) and reliance on a single human annotation reference. Future steps (which are currently in progress) will expand the dataset, refine persona clusters, and incorporate annotations from multiple human coders to more accurately assess LLM behaviour across user types.

References

[1] Peeters, A. L., & d'Haenens, L. (2005). Bridging or bonding? relationships between integration and media use among ethnic minorities in the netherlands. (2005): 201-231

[2] Richardson, R. J. (2022). Local tv newsroom diversity: Race and gender of newscasters

and their managers. Journal of Broadcasting & Electronic Media, 66(5), 823–842.

Persona-driven Simulation of Voting Behavior in the European Parliament with Large Language Models

Anonymous Abstract for CPSS 2025

Large Language Models (LLMs) display remarkable capabilities in understanding or even producing political discourse [7, 3, 6, 9, 1, 10]. In this work, we analyze whether zero-shot persona prompting can accurately predict individual voting decisions and, by aggregation, accurately predict positions of European parliamentary groups on a diverse set of policies. We evaluate if predictions are stable towards different persona prompts, generation methods and models. We find that we can simulate voting behavior of *Members of the European Parliament (MEP)* reasonably well with a weighted F1 score of approximately 0.793.

Methodology: We use roll-call vote data from HowTheyVote[1]. Of all 1,688 roll-call votes that were held in the ninth European parliament in the year 2024, we further filter for votes that are associated with a press release and a debate, resulting in a total of 47 votes. To prevent data leakage, we choose a model with a training cut-off prior to the earliest vote in our dataset on 16.01.2024: *Llama3* [4]. In addition, as Language Models can show bias towards positions of their creators [2], we choose a second model that was not developed in the west: *Qwen2.5*[2] [11]. Our models are: Llama3-8B (`Llama-3.1-8B-Instruct`), Llama3-70B (`Llama-3.1-70B-Instruct`), Qwen-7B (`Qwen2.5-7B-Instruct`), Qwen-72B (`Qwen2.5-72B-Instruct`).

To create persona descriptions for each MEP, we consider two strategies. First, we create a attribute prompt, which includes full name, gender, age at the time of voting, birthplace, the country they represent, the European group, and the national party with which they are affiliated at the time of voting. Second, we use Llama3-70B to summarize the English Wikipedia article associated with each MEP.

We adopt the theory that a common group position is established by the group's policy experts and then later adopted by other MEPs [8]. We choose to use the speeches of each group representative as context of the vote. The speeches are presented to the LLM in randomized order, without disclosing the speaker's identity or group. Names of politicians, groups and parties appearing in speeches were anonymized by substituting them with placeholders.

We instruct the model to respond with exactly one of the options FOR, AGAINST or ABSTENTION. We further compare two different prompting strategies; *reasoning (r)*, where the model responds first with an open text reasoning chain before choosing one of the options and *no reasoning (nr)*, where the model answers with one of the options directly. We prompt the model three times for each persona and take the mean weighted F1 score as result. In all runs, we use a temperature of 0.6.

Results: We use the mean weighted F1 score across three runs as our evaluation measure. We report that the responses of all models are robust, i.e. in 87.1% of cases, the models predict the same vote for each persona across all three runs, leading to a low standard deviation of < 0.002 across all runs. Our main results are shown in Table 1. The highest weighted F1-score of 0.793 is achieved by Llama3-70B when employing reasoning with attribute-based persona prompts. Consistent with expectations, larger models surpass their smaller counterparts in performance. Furthermore, the western Llama models typically outperform Qwen models from China of comparable size. The reasoning chain prompting approach improves overall prediction performance for LLama3-70B, Llama3-8B and Qwen-72B. As can be seen in Table 2, all models predict the votes of center-left and progressive groups (S&D, Renew, Greens/EFA) the most accurate and perform worst for groups at the edge of the political spectrum (ID, ECR, GUE/NGL). Model voting behavior of the best model is shown in Figure 1. Our findings show that LLMs outperform traditional approaches, for example Random Forests [5]. With further advancements LLMs could make democratic decision processes explainable to voters.

[1] https://howtheyvote.eu/

[2] Knowledge cutoff date is not known. Based on results compared to Llama3, risk of data leakage is low.

References

[1] Jan Batzner, Volker Stocker, Stefan Schmid, and Gjergji Kasneci. Germanpartiesqa: Benchmarking commercial large language models for political bias and sycophancy. *arXiv preprint arXiv:2407.18008*, 2024.

[2] Maarten Buyl, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jefrey Lijffijt, et al. Large language models reflect the ideology of their creators. *arXiv preprint arXiv:2410.18417*, 2024.

[3] Ilias Chalkidis and Stephanie Brandl. Llama meets EU: Investigating the European political spectrum through the lens of LLMs. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 481–498, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[5] Marina Guadarrama Rios, Federico Zamberlan, Paris Mavromoustakos Blom, and Nevena Rankovic. Knowing our choices: unveiling true voting patterns through machine learning (ml) and natural language processing (nlp) in european parliament. *Social Network Analysis and Mining*, 15(1):24, 2025.

[6] Michael Heseltine and Bernhard Clemm von Hohenberg. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239, 2024.

[7] Hao Li, Ruoyuan Gong, and Hao Jiang. Political actor agent: Simulating legislative system for roll call votes prediction with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 388–396, 2025.

[8] Nils Ringe. *Who decides, and how?: Preferences, uncertainty, and policy choice in the European Parliament*. OUP Oxford, 2009.

[9] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in LLM simulations of debates. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 251–267, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[10] Patrick Y Wu, Jonathan Nagler, Joshua A Tucker, and Solomon Messing. Large language models can be used to estimate the latent positions of politicians. *arXiv preprint arXiv:2303.12057*, 2023.

[11] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Appendix

Persona descr.	Llama3-70B	Llama3-8B	Qwen-72B	Qwen-7B
all attributes (r)	**0.793**	0.728	0.789	0.670
wikipedia (r)	0.779	0.724	0.773	0.697
all attributes (nr)	0.772	0.687	0.773	0.688
wikipedia (nr)	0.770	0.701	0.760	0.712

Table 1: **Prediction performance of voting behavior with persona prompts.** We report the weighted F1 scores for voting prediction across all proposals for prompts with reasoning (r) and no reasoning (nr). Majority baseline achieves 0.666.

Llama3-70B	**Overall**	GUE/NGL	S&D	Greens/EFA	Renew	EPP	ECR	ID	NI
all attributes (r)	**0.793**	0.711	0.924	0.867	0.900	0.804	0.592	0.529	0.596
wikipedia (r)	0.779	0.669	0.919	0.849	0.894	0.795	0.543	0.538	0.602
all attributes (nr)	0.772	0.671	0.911	0.804	0.899	0.792	0.529	0.56	0.569
wikipedia (nr)	0.770	0.662	0.924	0.854	0.895	0.786	0.507	0.528	0.588
Qwen-72B	**Overall**	GUE/NGL	S&D	Greens/EFA	Renew	EPP	ECR	ID	NI
all attributes (r)	**0.789**	0.735	0.917	0.842	0.899	0.799	0.574	0.574	0.563
wikipedia (r)	0.773	0.703	0.918	0.855	0.892	0.787	0.502	0.548	0.576
all attributes (nr)	0.773	0.687	0.912	0.846	0.900	0.777	0.516	0.578	0.525
wikipedia (nr)	0.760	0.664	0.915	0.849	0.899	0.771	0.475	0.523	0.559

Table 2: **Prediction performance of voting behavior across groups.** Weighted F1 scores for voting prediction with Llama3-70B and Qwen-72B across all proposals for each group for prompts with reasoning (r), without reasoning (nr).

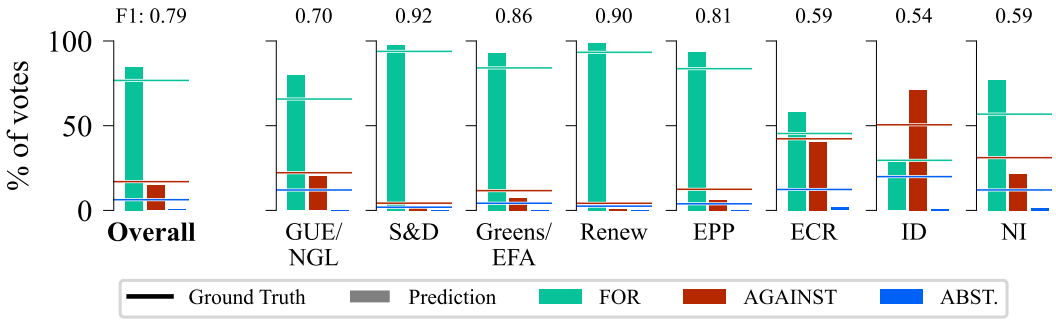

Figure 1: **Distribution of predicted votes per European group.** We display the voting predictions of the best performing model (Llama3-70B with attribute prompting and reasoning) compared to the ground truth. The weighted F1 score is displayed above each group.