

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

Maximilian Maurer^{1,2}, Ana Lisboa Cotovio¹, Matteo Melis², Julia Romberg¹,
Aldo Costa² and Gabriella Lapesa^{1,2}

¹GESIS - Leibniz Institute for the Social Sciences

²Heinrich-Heine University Düsseldorf

¹first.last@gesis.org

Abstract

The "political opinions" of Large Language Models (LLMs) increasingly find their way into political discourse despite limited understanding and control of these systems. Our paper adds to the understanding of how prompt manipulation impacts the multilingual generation of political arguments, focusing on two fundamental aspects of generated text: style and semantics. Starting from a reference set of Swiss politicians' arguments in German, Italian, and French, we conduct a large-scale prompt-manipulation experiment on 4 LLMs, building the prompts in a modular way (i.e., with blocks encoding socio-demographics, context, and examples) and testing all possible combinations. Using our parallel dataset of over 375k human argument-aligned generated arguments, we assess the detectability of LLM arguments through classification experiments with 160 linguistic features and semantic embeddings. We then carry out quantitative and qualitative analysis to understand how human-written vs. LLM-generated text compares, and to assess the impact of prompt manipulation.

Our results show that human and LLM arguments are substantially different in style and semantics, making their identification easy for classifiers. Prompt manipulation does not substantially impact semantics and style, except for sociodemographic persona prompting, which makes generated texts moderately more human-like. Model, language, and topics impact the stability of the generated text most.¹

1 Introduction

Political argumentative text generated by large language models (LLMs) plays an increasingly larger role in societal discourse. Besides usage in the general public, politicians either are suspected of using

¹We make the code, data, and raw results available via <https://github.com/mmmaurer/promptology>.



Figure 1: Prompt elements in English for illustration. The versions used in the experiments in the respective language can be found in Appendix C.

LLMs to write speeches² or outright say they use them to *get a second opinion*.³ This adaptation of politicians is unsurprising, as works find LLMs to produce novel, convincing arguments (Palmer and Spirling, 2023). At the same time, little is known about how generated political arguments compare to human ones in realistic scenarios.

While there is a growing literature on LLM-based argument generation and particularly the assessment of prompting techniques like few-shot prompting (Chen et al., 2024) and persona prompt-

²<https://www.pimlicojournal.co.uk/p/mps-are-a-most-certainly-using-chatgpt>

³<https://www.euronews.com/next/2025/08/07/sweden-prime-minister-uses-chatgpt-how-else-are-governments-using-chatbots>

ing (Hu et al., 2025), there is a noteworthy gap in understanding how such techniques change the semantic and stylistic makeup of generated arguments, given controlled input. This gap is particularly salient in a multilingual setting. Most studies focus on English, while it can be assumed that users interact with models in their native tongue.

To address the lack of comparison in realistic scenarios and the multilingual gap, we conduct a fully factorial analysis of argument generation compared against human arguments in a realistic scenario (i.e., in the context of a voting recommendation application). This means that, given a set of prompt elements and contextual factors (i.e., whether there are sociodemographic persona descriptions or explanations present, and the topic of the question), we are generating all possible combinations, allowing us to disentangle the effect of each of them while controlling for the others. Importantly, in contrast to an incremental setup, this yields the full picture of impacts. Concretely, we use a multilingual (DE, FR, IT) argument dataset as a source of human comparison items and sociodemographics, and prompt four LLMs with different prompt elements. For illustration of the prompt elements, see Figure 1, and for an aligned example of the prompt, a human argument, and outputs for two prompt settings, see Figure 2. Given the resulting human-LLM-prompt-parallel dataset of in total over 375k arguments, we train over 160K classifiers on linguistic features and sentence embeddings as diagnostic tools to assess whether there is a larger difference for style or semantics in a given condition, and overall. To disentangle the effects of prompt elements and contextual factors on differences to humans and on internal stability, we fit Bayesian regression models on the results of the classifiers and Rouge-L and BERTscore scores. We ask the following research questions:

RQ1.1 How different are LLM-generated vs. human-written arguments? Is it a matter of style or semantics? **RQ1.2** To what extent do different prompt elements (i.e., adding more information; making the prompt more specific) and contextual factors influence the stylistic and semantic similarity between human-written and LLM-generated text? **RQ2.1** How stable are the outputs? **RQ2.2** To what extent do different prompt elements and contextual factors influence output stability?

Our contributions are threefold: (1) We provide a prompting protocol for assessing prompt element effects in free-form generation in politi-

cal questionnaires. Using this, we create a conditioned and human-matched trilingual corpus of LLM-generated arguments. Its structure and rich comparison points enable future work to systematically assess differences across sociodemographic proxies and stereotypes in generated arguments, rhetorical strategies, or argument strength given different conditions, for example. (2) We propose an analysis pipeline to assess the impact of prompt elements and contextual factors on semantic and stylistic differences between human-written and LLM-generated texts and the stability of generated texts in large-scale controlled-condition datasets. (3) We find considerable impacts of model choice, language, and topic on synthetic texts, both in comparison to humans and in terms of stability given a prompt. Among assessed prompt elements, only sociodemographic persona prompting appears to make generated arguments moderately more human-like in semantics and style.⁴

2 Related Work

Assessment of Synthetic Texts. With the rising popularity of LLMs, more works assess synthetic texts, especially in comparison to *natural* human-written texts. A particularly active subfield is generated text detection, where benchmarks show that LLM-generated text can generally be detected relatively easily with different approaches (Wang et al., 2024; Li et al., 2024, i.a.). Focused analyses on the differences between LLM-written texts find LLMs to exhibit less variation than humans in their linguistic makeup when prompted with minimal information (Herbold et al., 2023; Dönmez et al., 2025). Initial works on persona-prompted free-form texts suggest human alignment due to personas to an extent, but find significant differences between different demographic proxies (Malik et al., 2024).

Zooming-In: Generation of Political Texts Herbold et al. (2024) assess whether LLMs can "impersonate" politicians, and find that while content and style are different than what the respective politicians would write, people rate LLM-generated texts attributed to these politicians as authentic. In a similar vein, Palmer and Spirling (2023) find that LLMs produce convincing, novel arguments, but humans dislike them if they know that it was LLMs who generated them. The recent turn in NLP re-

⁴Sociodemographic prompting has a consistent effect that is not overlapping with 0, but the effect size is comparably small.

Prompt	Arguments
SOCIODEMOGRAPHICS You belong to the following sociodemographic group: political spectrum - center and liberal	Default A ban on the registration of new cars with internal combustion engines from 2035 is an important step towards climate protection. Switzerland must reduce its CO2 emissions in order to meet the targets of the Paris Agreement. Electric mobility promotes cleaner air, less noise and lower running costs in the long term. Furthermore, the ban encourages innovation and the development of sustainable technologies, which benefits both the environment and the economy.
ISSUE You are asked about the following political issue: Should Switzerland ban the registration of new passenger cars with internal combustion engines from 2035?	SD A ban on the registration of new combustion engine vehicles from 2035 is an important step towards climate protection and the promotion of innovation. In doing so, Switzerland is demonstrating a sense of responsibility and creating incentives for environmentally friendly technologies. At the same time, a balanced implementation is needed, with sufficient support for alternative propulsion systems and socially responsible accompanying policies, to ensure that the transition is fair for all citizens.
INSTRUCTION Provide a comment in favor of the aforementioned political issue. You can use up to 500 characters.	HUMAN Admittedly, this would affect only a very small number of people, as the rise of the electric car is now unstoppable. Nevertheless, cars that cannot be operated in a climate-neutral way should be banned. This does not necessarily have to mean a ban on internal combustion engines, but rather a ban on conventional fuel.

Figure 2: Example prompt, a human answer, and two outputs by GPT-4. Note that Default is generated using the prompt elements ISSUE and INSTRUCTION, and SD additionally has the element SOCIODEMOGRAPHICS (which is inherent to the human argument). The examples were automatically translated to English. See the original German example in Figure 12 in the Appendix.

search bringing together sociodemographic persona prompting and large-scale surveys has raised concerns about the standard practice of presenting LLMs with categorical questions, as this practice does not do justice to the (alleged) open-ended generative diversity of LLMs (Ma et al., 2025, 2026). Work conducted in this framework has shown that model origin and prompt language systematically impact the outcome of political bias detection (Lim and Röttger, 2026). This promising direction of generation being better than categorical answers needs to be linked back to the known brittleness of sociodemographic prompting (Beck et al., 2024). What is brittleness in generation? Does stereotypical style come into play when LLMs are asked to generate political opinions from a specific area of the spectrum or sociodemographics? These issues point to the need for fine-grained prompting protocols and stability analysis, which we address in this paper, as well as the need for factoring in the variation in terms of style and semantics.

Regression Analysis of Model Behavior. Using regression models to "mine" model performance has a long tradition in NLP/CL (Lapesa and Evert, 2014; Padó et al., 2016; Daxenberger et al., 2017; Falk and Lapesa, 2022, i.a.). Recently, Benito-Santos et al. (2025) used linear mixed-effects regression to assess cross-lingual performance differences between LMs as a function of language, task, evaluation mode, and model. Similarly, Lauziere et al. (2026) suggest using regression models to analyze the impact of prompt elements on output performance.

3 Data

3.1 Smartvote Questionnaire

We use the general setup of the Swiss voting recommendation platform *smartvote*. It provides a questionnaire on political issues that both political candidates and voters fill out to recommend politicians with aligned views to voters. Candidates can additionally provide arguments for their stance on a given issue.⁵ smartvote provides this service in all official languages of Switzerland (German, French, Italian, and Romansh) and English. Figure 13 in Appendix I.2 shows the questionnaire for political candidates. In this work, we use the questions, the topic (policy domain of the question), explanations, common arguments (in the following referred to as PRO&CON), and the character limit for arguments.

3.2 Human-Written Arguments

We use human-written smartvote arguments from a dataset collected for the first perspective argument retrieval shared task (Falk et al., 2024). It contains information on the Swiss national elections of 2019 and 2023 with arguments formulated by politicians in a multilingual setup (German, French, and Italian). Politicians additionally provided sociodemographic information, which covers age, gender, place of residency, civil status, denomination, education, their political spectrum, and political issues important to the author of the argument. Due to sparsity issues for explanations and common ar-

⁵Note that *arguments* is used in a broad sense in this work. We consider texts in the comment field in the smartvote application to be arguments regardless of formal structure, as they aim to explain a stance or persuade an audience.

guments in the 2019 subset, especially in Italian, we only use the 2023 subset in this work.

3.3 LLM-Generated Arguments

Our goal is to generate arguments in controlled conditions to evaluate them against human-written arguments. To do so, we choose four models and generate arguments given 32 prompt element combinations in the three languages present in the human data. This constitutes a fully factorial design, as we build prompts in a modular way and test all the combinations. In the following, we provide details on the chosen models and parameters, and the prompting conditions.

Models and Generation Parameters. We choose four models that are reported to be trained on German, Italian, and French data, intending to cover a broad range of model sizes, widely available commercial and open-weight models, and different architectures with the research budget available to us: **GPT-4.1-mini**, as a commercial black-box model widely available at the time of data generation, **LLaMA-4-Scout**, an open-weight mixture-of-experts model with 109B total parameters, **LLaMA-3.1-8B-instruct**, a smaller model of the same model family as LLaMA-4-Scout, runnable in full precision on limited resources, and **Occiglot-7B-EU7**, a model trained with continual pre-training and instruction tuning on the five biggest European languages (EN, ES, DE, FR, IT), which was developed following the "European AI sovereignty" strategy.

To avoid potential model-specific confounding effects, we do not use quantization for any of the models. Per prompt, we run three iterations with a temperature of 0.7 and, per model, use the respective standard decoding strategy.

Prompt Elements. As illustrated in Figure 1, we generate arguments using six prompt elements. Two blocks are fixed in all combinations: the **ISSUE**, containing the political issue in the form of a question (e.g., *Do you support an increase in the retirement age (e.g., to 67)?*), and the **INSTRUCTION**. The instruction asks for an argument for or against the given political issue, and sets a character limit. Prompts only containing these two will, in the following, be referred to as the **Default** condition. The **ISSUE** block can follow a **CONTEXT** (C) and/or a **SOCIODEMOGRAPHICS** (SD) block. C provides a short explanation, conditioning the models on political statements in the context of

Model	DE	FR	IT
GPT-4.1	45,822	34,452	16,218
LLaMA-4	45,822	34,452	16,218
LLaMA-3.1	41,045	33,300	15,306
Occiglot	42,690	34,452	15,682
Human	10,118	2,307	333

Table 1: Number of generated arguments per model and language. For comparison, we also provide the number of human arguments. Differing numbers for models across languages are due to refusals and empty outputs.

Swiss federal elections. SD is a sociodemographic persona prompting block. The **ISSUE** block can be followed by an **EXPLANATION** (Ex), providing a short explanation about the political issue and/or a **PRO&CON** (P&C) block, providing common arguments for and against the given issue. This can be viewed as a block similar to a few-shot setup, but in fluent text instead of templated examples.

We generate arguments for all combinations of elements, yielding 16 prompt settings per language. To ensure comparability, we only generate arguments given those settings, where we have at least one human-written counterpart to compare against. In consequence, if for a given political issue no explanation is available, we do not generate any combination with the Ex block. We remove all empty responses (i.e., responses not containing any characters or only containing whitespace characters). Overall, we generate 375,459 valid arguments. We provide detailed statistics in Table 1.

Initial analyses of the generated texts reveal that only GPT-4 adheres to the character limit in more than 95% of texts across languages; the other models fluctuate between 1.7% and 77.4%, depending on language-model combination.⁶ The two smaller models, LLaMA 3 and Occiglot, sometimes generate text in languages other than the prompt language (0.6% FR, 6.9% DE, 8.2% IT), and repetitive, non-alphabetical, and nonsense outputs.⁷

4 Methods

4.1 Difference Analysis

Our aim is to disentangle the effects of prompt elements, topic, stance, model, and language on semantic and stylistic differences to human-written texts. Given that our data facilitates a factorial design (i.e., measure the impact of factors by mea-

⁶See the full statistics in Table 12.

⁷For more details, see Appendix G.

asuring them in a controlled manner), we use support vector machine (SVM) classifiers as diagnostic measurement tools to assess how informative a feature set is to classify human-written vs. generated arguments in a specific setting. The higher the performance of a given classifier, the more distinct human-written and LLM-generated arguments are from each other in the given dimension (e.g., in the usage of affectively grounded words for emotion, or semantically overall for SBERT). Each classifier’s performance can be seen as a controlled observation, allowing us to assess general tendencies via aggregate statistics and the effects of the prompt elements, topic, stance, model, and language via a linear regression model.

Stylistic Features. To assess stylistic differences, we use the feature extraction library *elfen* (Maurer, 2026). We extract all available features and remove all constant features (i.e., those that have the same value across all arguments). For comparability across prompting conditions, models, the three languages, and with humans, we select those features that occur in all sets of arguments. This yields 160 linguistic features across 11 distinct feature areas considered in our analyses.⁸ We token-normalize count-based features, standardize all features, and fill missing values per feature heuristically with the respective mean.

In the subsequent experiments, we use the full set, in the following referred to as *full*, and the set of features for each of the feature areas.

Sentence Embeddings. To assess semantic differences, we extract sentence embeddings using *sentence_transformers* (Reimers and Gurevych, 2019). Specifically, we use the paraphrase-multilingual-MiniLM-L12-v2 model.⁹ The sentence embedding representation setting will subsequently be referred to as SBERT.

Classifiers. To account for the imbalance in our data, both overall and per setup introduced below, we use class-weighted SVMs. We initially consider $n = p \times t \times s \times m \times l \times 5$ folds classifiers to be trained per feature set, where $p = 16$ is the number of prompt setting combinations, $t = 13$ is the number of topics, $s = 2$ is the numbers of stances, $m = 4$ is the number of LLMs, and $l = 3$ is the number of languages. Given our generation restric-

tions, we only have $4,300 \times 5 \text{ folds} = 21,500$ in our dataset out of $n = 24,960$ theoretically possible classifiers. We heuristically disregard settings with fewer than 100 samples to avoid too sparse measurements. To additionally account for performance variation, we perform 5-fold cross-validation. This yields 12,355 classifiers per feature set. We measure balanced accuracy (BA) per classifier to account for class imbalances.¹⁰

Regression Modeling. To disentangle the impact of prompt elements, topic, stance, model, and language on how easy it is to differentiate generated arguments from human-written ones across controlled settings, per feature set, we fit a Bayesian regression model. Specifically, we model BA as a function of the predictors mentioned above, and some of their interactions.¹¹ We provide the formula in Wilkinson notation (Wilkinson and Rogers, 1973) in Equation 1, where C, SD, Ex, and P&C are the prompt elements, T is the topic, S the stance, M the LLM, and L the language. The prompt elements are dummy-coded, and all other predictors are categorical. The outcome is modeled using beta regression with a logit link function. To be able to include BAs at the edges of the distribution (0 and 1, respectively), we apply the Smithson-Verkuilen transformation (Smithson and Verkuilen, 2006).

$$\begin{aligned} BA \sim & C + SD + Ex + P\&C + T + S + M + L + \\ & C : P\&C + C : SD + C : Ex + Ex : P\&C + \\ & SD : P\&C + Ex : SD + T : S + T : M \end{aligned} \quad (1)$$

4.2 Stability Analysis

Following Sclar et al. (2024), we assess the stability of the output triplet per unique prompt using Rogue-L (Lin and Och, 2004) and BERTScore (Zhang et al., 2020). Rogue-L measures the longest common subsequence on a surface level, and BERTScore is an embedding-based similarity score, reflecting the style and semantics dimensions in the difference analysis. For both of the metrics, we use the mean F1 score per triplet. The higher the mean F1, the more stable the outputs are.

Regression Modeling. To disentangle the impact of prompt elements, topic, language, LLM, and stance on output stability, we model the mean Rogue-L and BERTScore F1s per triplet in the

⁸See Tables 3-6 in Appendix B for the full list of features.

⁹<https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>

¹⁰See Equation D in Appendix D for the formula.

¹¹We do not include more interactions to retain straightforward interpretability and for computational efficiency reasons.

Feature Set	Feature Area	BA	Std.Dev.
SBERT	–	90.2	14.5
	All	96.7	9.6
Linguistic	Lexical Richness	96.5	8.9
	Readability	93.9	9.7
	Surface	93.7	9.4
	Information	87.1	12.6
	Emotion	82.5	15.6
	POS	80.6	14.6
	Morphology	77.9	16.2
	Dependencies	79.8	13.4
	Semantics	77.3	14.7
	Psycholinguistics	72.2	16.8
	Named Entities	66.8	15.2

Table 2: Mean balanced accuracy (BA, in percent) and standard deviation (Std.Dev., in percent) per feature set and feature area (for linguistic features). See Table 7 in Appendix E for results per language.

same setup as in the regressions for the difference analysis; the only difference is that we use the F1s as the dependent variable instead of the BA (c.f., Equation 1), and we sample from a population of output triplet scores instead of classifiers.

5 Results

5.1 Difference Analysis

Table 2 shows the mean BA and standard deviation per feature set and feature area for the linguistic features. We observe that stylistic features provide a stronger differentiating signal than semantic ones: using all linguistic features yields a mean BA of 96.7%, whereas using SBERT yields a mean BA of 90.2%. A closer inspection of the mean BA for classifiers trained on individual feature areas reveals that stylistic differences between human-written and LLM-generated arguments are mainly driven by word choice and text complexity, as indicated by the particularly high BA ($> 93.7\%$). While feature areas that may capture more semantic and pragmatic functions of the text (*emotion*, *semantics*, *psycholinguistics*, and *named entities*) show substantially lower scores ($66.8\% \leq BA \leq 82.5\%$), all reach a mean well over a random baseline (50%). Overall, these results suggest that LLM-generated and human-written arguments differ substantially in both semantics and style. We note, however, that there are considerable standard deviations across settings.

To assess what drives this variation, we turn to the regression results presented in Figure 3. We observe substantive differences between models: in comparison to GPT-4, LLaMA 4 and Occiglot are easier to classify with linguistic features, and

harder with sentence embeddings, and LLaMA 3 the other way around. We find similar opposite effects of linguistic features for different languages. Compared to German, generated arguments in French are moderately harder to classify against human-written ones with SBERT embeddings, and moderately easier with linguistic features. For Italian, we find a strong negative effect for SBERT embeddings, suggesting substantial semantic differences specific to Italian. Among the prompt elements, we find moderate effects across all of them, with the same direction of the effect both for linguistic features and SBERT embeddings. Only sociodemographics, however, show a consistent effect not overlapping with 0 for the 95% highest density interval (HDI) for both feature sets. Finally, we observe substantial variation between topics. This is confirmed by the `topic:model` interaction (see Figure 10 in Appendix E).

Zooming In: Linguistic Features Given the pronounced differences between the LLM-generated and human-written arguments, we zoom in on the linguistic features. To gather an understanding of which features provide the strongest signals, we calculate the point-biserial correlations between each of the linguistic features and the source of the arguments (GPT, LLaMA 4, LLaMA 3, Occiglot, Human). Length-related surface features such as the number of tokens have the strongest mean correlation (mean -0.22 ; number of tokens -0.56).¹² This, however, should not be interpreted as the factor explaining the difference, given the strong classification signal of, for example, POS features, but their low Spearman correlation with the number of tokens ($\rho \in [-0.14, 0.34]$) and their low point-biserial correlation with the source ($r \in [-0.19, 0.29]$). We take this as evidence of differences in the linguistic makeup of the texts at large rather than only individual features driving the difference.

The feature most strongly associated with the source of the arguments is the type-token ratio (TTR). As Figure 4 shows, this mirrors findings of prior work for online forum counter-arguments (Dönmez et al., 2025): LLMs produce arguments with less varied, less creative word use.

5.2 Stability Analysis

Figure 5a shows a boxplot for the Rogue-L scores per LLM and language, and Figure 5b shows

¹²See the full results in Appendix H.

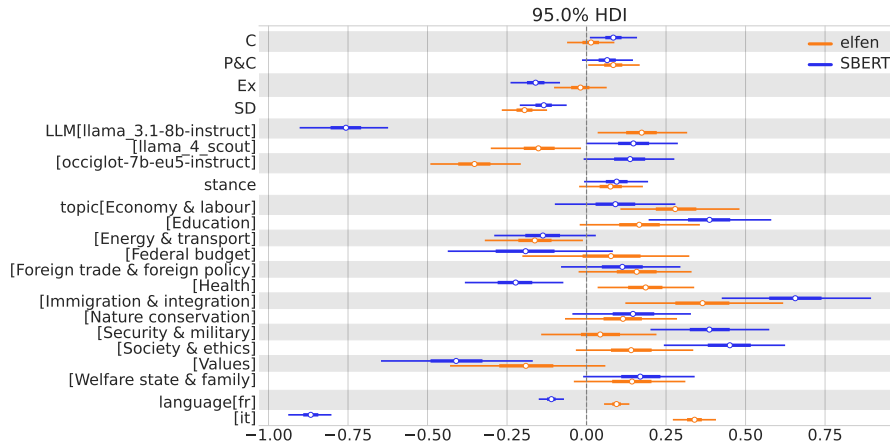


Figure 3: Forest plot for the main effects of the regression models for classifiers using elfen (linguistic features) and SBERT (semantic representation). Reference levels are: GPT-4 for LLM, Democracy, Media & Digitization for topic, AGAINST for stance, and DE for language.

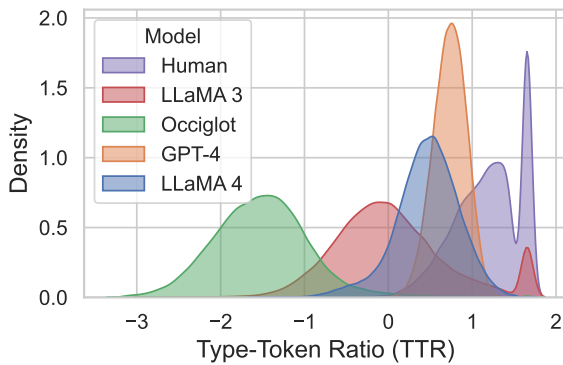


Figure 4: Density plot of the type-token ratio per model.

BERTScores per LLM and language. Across both metrics, we find similar tendencies: more capable models (GPT-4 and LLaMA 4) show a higher consistency than smaller, less performant models (LLaMA 3 and Occiglot) across both metrics. This indicates that bigger and more capable models exhibit more stable outputs given the same prompt, both semantically and on a surface level. Mean, upper, and lower quartiles of Rouge-L are tendentially lower than BERTScore across models and languages. This indicates that while the surface-level stability varies substantially, and is fairly low for many prompts (mean Rouge-L F1 < 0.5 across models and languages), the semantic stability is still comparably high (mean BERTScore F1 > 0.6 across models and languages). While there are slight differences between languages per model, we do not observe a conclusive tendency.

Figure 6 shows the forest plot for the main effects of the regression models for Rouge-L and BERTScore. As already indicated in the distri-

bution, we observe big differences between models: With everything else fixed, LLaMA 3 and Occiglot produce substantially less stable outputs than GPT-4, and LLaMA 4 produces more stable outputs. The effects are stronger for Rouge-L, indicating bigger surface-level differences within the outputs of a given prompt. Similarly, we observe language differences. Both French and Italian outputs are more stable on both metrics when compared to German. Again, these effects are stronger for Rouge-L than for BERTScore, pointing to more repetition in French and Italian outputs. While prompt elements do have effects, with PRO&CON and EXPLANATIONS being associated with less stable, and SOCIODEMOGRAPHICS with more stable outputs, their impact is smaller (all $|\hat{\beta}| < 0.25$) than model and language effects. We observe the largest differences on a topical level. This is confirmed by large variations in the LLM-topic interactions (c.f. Figure 11 in Appendix F).

5.3 Qualitative Analysis

To understand the properties of the generated arguments in more detail, we qualitatively analyze 120 German and 112 Italian arguments.¹³ The goal of our analysis is to get a first impression of the differences in performance between SBERT and the linguistic feature based classifiers: what makes some items harder to classify for one of the two

¹³We assess Italian and German as the most contrasting language pair, given the less pronounced differences in effects between German and French in the regression analysis. Moreover, none of the authors is a native French speaker, which such a type of analysis demands in addition to familiarity with the experiments.

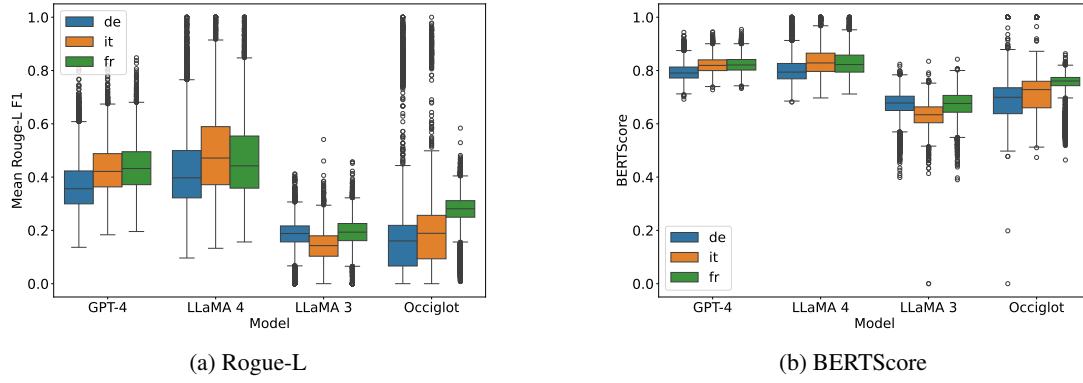


Figure 5: Distribution of stability metrics per LLM/language.

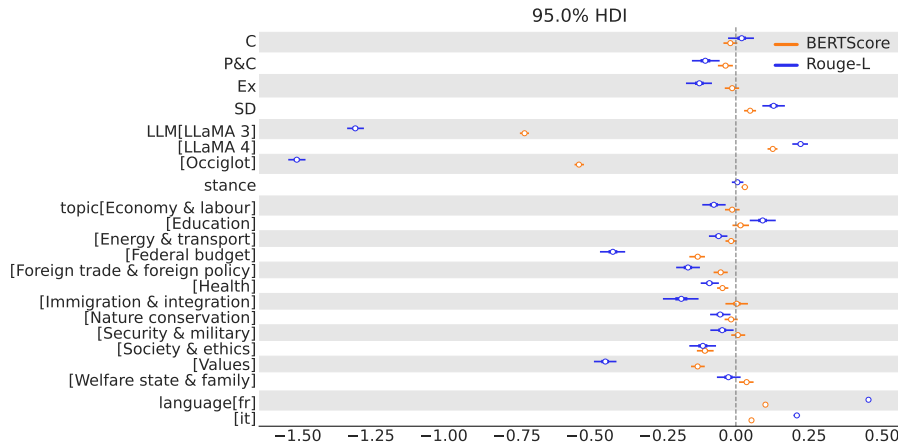


Figure 6: Forest plot for main effects of the regression models for Rouge-L and BERTScore. Reference levels are: GPT-4 for LLM, Democracy, Media & Digitization for topic, AGAINST for stance, and DE for language.

classifiers? Per language, we randomly sample arguments generated by LLaMA 3 and GPT-4; 40 where SBERT setups lead to more correct classifications than linguistic features, 40 the other way around, and 40 where they have the same number of correct classifications. Due to a lower FP rate for GPT-4, we have a smaller sample in Italian.

In a first phase of the qualitative analysis, four annotators (two per language) went through the sample and independently from each other came up with a set of the issues they observed in the data. After discussion among the four annotators, the different observations were consolidated and mapped into a set of categories that were then applied to the sample by one annotator per language.¹⁴

Our first observation concerns the generation quality: both languages suffer from issues such as prompt leakage into the generated text (14% DE vs. 23% IT), content loops (38% vs. 51%), cut-off sentences (48% vs. 46%), or LLM commentaries (28% vs. 32%). This affects 62 items in German

and 82 in Italian, providing a strong signal of LLM generation. While German texts are mostly correctly identified as LLM-generated by both feature sets in these cases, prediction is less reliable for Italian with linguistic features failing in 23 cases and SBERT embeddings in 28.

Regarding other observable patterns, we find frequent use of pronouns (85% DE vs. 60% IT). However, this feature does not appear to predict LLM-generated arguments. The same holds for plausible persona descriptions, which may equally occur in human-written texts. We also observe differences between clearly polarized stances and more cooperative argumentative styles, but these too seem to be close to how human politicians write.

6 Discussion

On an aggregate level, our results suggest a relatively **high difference between human-written and LLM-generated arguments, both in terms of their style and in terms of their semantics**. This aligns with idiosyncrasies of generated con-

¹⁴The qualitative codings are available in the repository.

tent found in prior work (Herbold et al., 2023; Dönmez et al., 2025). More interestingly, however, prompt blocks with additional information do not appear to impact the output’s semantics and style considerably. **Only sociodemographic persona prompting has a consistent effect on LLM-generated arguments, making them more human-like in both regards.** The small effect size, however, potentially points to different extents for different proxies (c.f., Malik et al., 2024). Moreover, this result raises the question of whether persona prompting does reflect stable tendencies of sociodemographic groups in an open-ended argument generation scenario, or relies on shallow lexical cues, as found for closed-ended advice generation (Tonneau et al., 2026). The **strong impact of the chosen LLM and language** urges caution for mapping findings of prior work in English to the respective target language actually used by end users. This is particularly the case as users can be assumed to interact with LLMs in their native tongue rather than English, especially in realistic scenarios where there is country-specific political language. The qualitative analysis emphasizes the language differences once more. Lastly, the observed **variance between topics and topic-model pairs** indicates that, while models may still produce arguments that are stylistically and semantically different from human-written ones, the specific characteristics and extent of this difference may vary considerably between topics. This makes anticipating behavior for especially newly arising political issues difficult.

Our findings indicate that **output stability, surprisingly, is also mainly impacted by model, language, and topic.** On the one hand, this again urges caution about the transferability of certain tendencies found for high-resource languages such as English to other languages. If languages with a comparably still decent amount of resources, like Italian compared to German, already show considerable differences in stability, this will likely be amplified for English compared to even lower-resource languages than Italian. On the other hand, if output stability with fixed generation settings varies so substantially between languages, models, and topics, for some combinations, the generated argument may virtually always be the same, almost verbatim, while for others, there is a lot of variation. If such generated arguments are then adopted by users, this may lead to a narrowing of arguments for some issues, but a broadening for others.

Overall, our findings have consequences for real-world adaptation of generated arguments and simulations of political actors: since arguments stay separable from human arguments in style and semantics across prompt conditions, i.e., the added information appears to have little impact, widespread adoption would almost certainly shift which arguments are used and how they are phrased. This raises the question of which topics are impacted in what way, and in which direction model idiosyncrasies might influence discourse. It is unclear whether tendencies found in discriminative settings (Ceron et al., 2024; Weeber et al., 2026, i.a.) transfer to generation settings (c.f., Jung et al., 2026).

7 Conclusion

This paper targeted the systematic LLM generation of political arguments (German, French, and Italian) and their comparison to human-written ones. Building up our prompts modularly at an increasing level of specificity, we created a human-LLM parallel dataset containing 375k arguments. Our extensive classification experiments (160K runs) involved linguistic features as well as embeddings. We then took performance in the LLM-generated detection task as the dependent variable of a regression analysis, uncovering the factors impacting the predictability of an argument as LLM vs. human as a function of the different prompt components (i.e., were socio-demographics or examples provided to the LLMs?), models, languages, and topics, both for semantics and style, respectively. While our results overall paint a reassuring picture, as LLM-generated arguments can be detected with high accuracy, we find a comparably small impact of prompt manipulation. The only exception is socio-demographic prompting, which makes the LLM-generated outputs moderately more human-like. Moreover, language, topic, and employed LLMs substantially affect output stability.

Limitations

Our work is limited along multiple axes. Firstly, our focus is to assess prompt and contextual factor variation in a setup that is close to how end users are exposed to models and how the original questionnaire is set up. It is, however, reasonable to assume that different temperatures and decoding strategies, and even quantization, interact with prompt elements or other factors systematically. Similarly, we only test one ordering of elements and one phrasing

per element. While certainly worth investigating, it would explode the combinatorial space. Our experiments test four models. While our choice is motivated by the properties of the models and limited by resources, our results do not allow for assessing scaling and model family effects.

Secondly, while data and setup allow for this, our work does not analyze fine-grained individual socio-demographic proxies and their instantiations. We are convinced that such socially relevant questions deserve attention and proper discussion, and encourage future work on this. We focus on the semantic and stylistic proxies in this work. Given the rich assessment of arguments along axes such as quality (c.f. Wachsmuth et al., 2017) and rhetorical devices (see, e.g., Schaefer et al., 2023), we encourage future work to assess whether findings of LLM idiosyncracies map to these dimensions, and how factors more generally may interact.

While we find a low impact of prompt elements on the stylistic and semantic difference between humans and LLMs, this may be domain-specific.

We cannot fully control for the presence of our test data in the training data of the used LLMs: While the data surely is not present in the respective training data in a structured form, as the PerspectiveArg shared task (Falk et al., 2024) was after the knowledge cutoff, the 2023 Swiss election, and the release of the respective version of smartvote has some overlap with the time before the cutoffs. A scraped version of the website, with its human arguments and meta-information, may have ended up in some LLMs' training data. Given the observed difference between human and LLM-generated arguments, however, we believe that leakage effect on the arguments are rather small, if present at all.

Finally, our analysis setup is one of many to assess the question of difference and stability in style and semantics. The linguistic features and the sentence embeddings are noisy proxies of stylistic and semantic representations, and Rouge-L and BERTScore are noisy measures of surface-level and semantic stability. As such, they, and any other specific measure, may systematically miss or inflate certain effects. Classification as a diagnostic tool for the informativeness of certain features, and thus of the difference between humans and LLMs, is a reasonable, scalable choice, but it is dependent on what the used classifiers can learn from the features. We note, however, that in our case, this can be assumed to have a reasonably low impact on downstream analyses, given the generally high

performance. The choice of analysis procedure, in our case, the regression modeling, comes with many potential choices. Exploring the differences between these choices in the yielded findings lies outside the scope of this paper.

Ethical Considerations

As with any work generating and assessing data in societally relevant domains, our protocol and generated data have the risk of being abused by malicious actors (*dual use*). This, however, is a more general problem with the large-scale adoption of LLMs. We believe that our work, even though it is not our focus, can contribute to the mitigation of such dual use, as our classifiers show high balanced accuracy across settings, and prompting techniques used to push arguments to be more human-like appear to have a limited success in doing so.

Our work suggests a rather high semantic and stylistic difference to humans. This, however, should not be seen as a claim of unbiased idiosyncratic *LLM style and semantics*. Given that LLMs are pre-trained on massive amounts of data that are dominated by certain author groups, they may still resemble the language use of hegemonic groups the most (c.f., Alvero et al., 2024).

Similarly, the generated arguments may contain biased, marked (c.f., Cheng et al., 2023), or caricaturized (c.f., Deas et al., 2025) expressions of prompted identities, especially when using persona prompting. We urge caution in using our data, but more generally data generated with a similar protocol, in model training, or for simulation experiments. Such downstream uses should be preceded by critical reflection and analysis of how tendencies in the data may reinforce stereotypes, biases, and misrepresentations in findings or applications.

Acknowledgments

We thank Michael Erne from smartvote for access to the raw questionnaire data and the valuable insights on the design of the platform and how politicians use it. We thank Andreas Waldis and Neele Falk for access to the raw data of the PerspectiveArg shared task. We thank the DSM++ team at GESIS CSS for their valuable input throughout this project. Ana Lisboa Cotovio acknowledges funding from the project Digital Dehumanization: Measurement, Exposure, and Prevalence (DeHum), supported by the Leibniz Association Competition (P101/2020).

References

- A. J. Alvero, Jinsook Lee, Alejandra Regla-Vargas, René F. Kizilcec, Thorsten Joachims, and Anthony Lising Antonio. 2024. [Large language models, social demography, and hegemony: comparing authorship in human and synthetic text](#). *Journal of Big Data*, 11(1):138.
- Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. [Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian’s, Malta. Association for Computational Linguistics.
- Alejandro Benito-Santos, Adrian Ghajari, and Víctor Fresno. 2025. [Robust estimation of population-level effects in repeated-measures NLP experimental designs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33076–33089, Vienna, Austria. Association for Computational Linguistics.
- Tomás Capretto, Camen Piho, Ravin Kumar, Jacob Westfall, Tal Yarkoni, and Osvaldo A Martin. 2022. [Bambi: A simple interface for fitting bayesian linear models in python](#). *Journal of Statistical Software*, 103(15):1–29.
- Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. [Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in LLMs](#). *Transactions of the Association for Computational Linguistics*, 12:1378–1400.
- Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Li-dong Bing. 2024. [Exploring the potential of large language models in computational argumentation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.
- Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. [Marked personas: Using natural language prompts to measure stereotypes in language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.
- Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. [What is the essence of a claim? cross-domain claim identification](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.
- Nicholas Deas, Blake Vente, Amith Ananthram, Jessica A Grieser, Desmond U. Patton, Shana Kleiner, James R. Shepard Iii, and Kathleen McKeown. 2025. [Data caricatures: On the representation of African American language in pretraining corpora](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29192–29217, Vienna, Austria. Association for Computational Linguistics.
- Esra Dönmez, Maximilian Maurer, Gabriella Lapesa, and Agnieszka Falenska. 2025. [AI argues differently: Distinct argumentative and linguistic patterns of LLMs in persuasive contexts](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34595–34626, Suzhou, China. Association for Computational Linguistics.
- Neele Falk and Gabriella Lapesa. 2022. [Reports of personal experiences and stories in argumentation: datasets and analysis](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5530–5553, Dublin, Ireland. Association for Computational Linguistics.
- Neele Falk, Andreas Waldis, and Iryna Gurevych. 2024. [Overview of PerspectiveArg2024 the first shared task on perspective argument retrieval](#). In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 130–149, Bangkok, Thailand. Association for Computational Linguistics.
- Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. [A large-scale comparison of human-written versus chatgpt-generated essays](#). *Scientific Reports*, 13(1):18617.
- Steffen Herbold, Alexander Trautsch, Zlata Kikteva, and Annette Hautli-Janisz. 2024. [Large language models can impersonate politicians and other public figures](#). *Preprint*, arXiv:2407.12855.
- Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025. [Debate-to-write: A persona-driven multi-agent framework for diverse argument generation](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4689–4703, Abu Dhabi, UAE. Association for Computational Linguistics.
- J. D. Hunter. 2007. [Matplotlib: A 2d graphics environment](#). *Computing in Science & Engineering*, 9(3):90–95.
- Jana Jung, Marlene Lutz, Indira Sen, and Markus Strohmaier. 2026. [Do psychometric tests work for large language models? evaluation of tests on sexism, racism, and morality](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8143–8173, Rabat, Morocco. Association for Computational Linguistics.
- Gabriella Lapesa and Stefan Evert. 2014. [A large scale evaluation of distributional semantic models: Parameters, interactions and model selection](#). *Transactions*

- of the Association for Computational Linguistics, 2:531–546.
- Andrew Lauziere, Jonathan Daugherty, and Taisa Kushner. 2026. [A regression framework for understanding prompt component impact on llm performance](#). *Preprint*, arXiv:2603.26830.
- Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. [MAGE: Machine-generated text detection in the wild](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.
- Ying Ying Lim and Paul Röttger. 2026. [Bias in the east, bias in the west: A bilingual analysis of LLM political bias on U.S.- and China-related issues](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2301–2326, Rabat, Morocco. Association for Computational Linguistics.
- Chin-Yew Lin and Franz Josef Och. 2004. [Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics](#). In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.
- Bolei Ma, Yong Cao, Indira Sen, Anna-Carolina Haensch, Frauke Kreuter, Barbara Plank, and Daniel Herscovich. 2026. [Too open for opinion? embracing open-endedness in large language models for social simulation](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2016, Rabat, Morocco. Association for Computational Linguistics.
- Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, and Matthias Aßenmacher. 2025. [Algorithmic fidelity of large language models in generating synthetic German public opinions: A case study](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1785–1809, Vienna, Austria. Association for Computational Linguistics.
- Manuj Malik, Jing Jiang, and Kian Ming A. Chai. 2024. [An empirical analysis of the writing styles of persona-assigned LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19369–19388, Miami, Florida, USA. Association for Computational Linguistics.
- Maximilian Maurer. 2026. [elfen: A python package for efficient linguistic feature extraction for natural language datasets](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 61–74, Rabat, Morocco. Association for Computational Linguistics.
- Sebastian Padó, Aurélie Herbelot, Max Kisselew, and Jan Šnajder. 2016. [Predictability of distributional semantics in derivational word formation](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1285–1296, Osaka, Japan. The COLING 2016 Organizing Committee.
- Alexis Palmer and Arthur Spirling. 2023. [Large language models can argue in convincing ways about politics, but humans dislike ai authors: implications for governance](#). *Political Science*, 75(3):281–291.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Robin Schaefer, René Knaebel, and Manfred Stede. 2023. [Towards fine-grained argumentation strategy analysis in persuasive essays](#). In *Proceedings of the 10th Workshop on Argument Mining*, pages 76–88, Singapore. Association for Computational Linguistics.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.
- Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.
- Michael Smithson and Jay Verkuilen. 2006. [A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables](#). *Psychological methods*, 11(1):54.
- Manuel Tonneau, Neil K. R. Seghal, Niyati Malhotra, Sharif Kazemi, Victor Orozco-Olvera, Ana María Muñoz Boudet, Lakshmi Subramanian, Samuel P. Fraiberger, Sharath Chandra Guntuku, and Valentin Hofmann. 2026. [Different demographic cues yield inconsistent conclusions about llm personalization and bias](#). *Preprint*, arXiv:2601.18486.
- Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. [Computational argumentation quality assessment in](#)

natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. [M4: Multi-generator, multi-domain, and multilingual black-box machine-generated text detection](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian’s, Malta. Association for Computational Linguistics.

Michael L. Waskom. 2021. [seaborn: statistical data visualization](#). *Journal of Open Source Software*, 6(60):3021.

Franziska Weeber, Tanise Ceron, and Sebastian Padó. 2026. [Do political opinions transfer between western languages? an analysis of unaligned and aligned multilingual LLMs](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 319–340, Rabat, Morocco. Association for Computational Linguistics.

G. N. Wilkinson and C. E. Rogers. 1973. [Symbolic description of factorial models for analysis of variance](#). *Journal of the Royal Statistical Society Series C: Applied Statistics*, 22(3):392–399.

Tianyi Zhangz, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.

A Contributions

In the following, we list the contributions of each author of this paper according to the CRediT taxonomy.¹⁵

- **Conceptualization:**
Maximilian Maurer, Ana Lisboa, Matteo Melis, Julia Romberg, Gabriella Lapesa
- **Data Curation:**
Maximilian Maurer, Ana Lisboa, Matteo Melis, Julia Romberg, Aldo Costa, Gabriella Lapesa
- **Formal Analysis:**
Maximilian Maurer, Ana Lisboa, Matteo Melis, Julia Romberg, Aldo Costa, Gabriella Lapesa

- **Investigation:**
Maximilian Maurer, Ana Lisboa, Matteo Melis, Julia Romberg, Gabriella Lapesa
- **Methodology:**
Maximilian Maurer, Ana Lisboa, Matteo Melis, Julia Romberg, Gabriella Lapesa
- **Project Administration:**
Maximilian Maurer, Ana Lisboa, Matteo Melis, Julia Romberg, Gabriella Lapesa
- **Resources:**
Maximilian Maurer
- **Funding Acquisition:**
We used institutional funds available as employees at GESIS.
- **Software:**
Maximilian Maurer, Matteo Melis, Julia Romberg
- **Supervision:**
Gabriella Lapesa
- **Validation:**
Maximilian Maurer
- **Visualization:**
Maximilian Maurer
- **Writing - original draft:**
Maximilian Maurer, Ana Lisboa, Julia Romberg, Gabriella Lapesa
- **Writing - review & editing:**
Maximilian Maurer, Ana Lisboa, Matteo Melis, Julia Romberg, Gabriella Lapesa

B Full List of Linguistic Features

Table 3 - 6 show the full list of linguistic features per feature area we use in this paper. For a description of the features, see the documentation of the extraction package we used: https://elfen.readthedocs.io/en/latest/feature_overview.html

C Prompts

Figure 7 shows the prompt setup in German, Figure 9 in French, and Figure 8 in Italian.

¹⁵<https://credit.niso.org/>

DEITFR

CONTEXT
 Sie sind ein Schweizer Politiker, der einen Fragebogen über seine politischen Überzeugungen für eine Wahlhilfe-App im Kontext einer Nationalratswahl ausfüllt.

NO CONTEXT

SOCIODEMOGRAPHICS
 Sie gehören der folgenden soziodemografischen Gruppe an: {sociodemographic info} - {sociodemographic group}

NO SOCIODEMOGRAPHICS

ISSUE
 Sie werden nach folgendem politischen Thema gefragt: {question}

EXPLANATION
 Folgendes ist eine Erklärung des politischen Themas: {explanation}

NO EXPLANATION

PROS & CONS
 Folgende sind gängige Argumente für das politische Thema: {pro arguments}
 Folgende sind gängige Argumente gegen das politische Thema: {contra arguments}

NO PROS & CONS

INSTRUCTION
 Schreiben Sie einen Kommentar { für | gegen } die oben genannte politische Frage. Sie können bis zu 500 Zeichen verwenden.

Figure 7: Prompt elements in German.

DEITFR

CONTEXT
 Lei è un politico svizzero che sta compilando un questionario sulle proprie convinzioni politiche per un app di consulenza al voto nel contesto di un'elezione federale.

NO CONTEXT

SOCIODEMOGRAPHICS
 Lei appartiene al seguente gruppo sociodemografico: {sociodemographic info} - {sociodemographic group}

NO SOCIODEMOGRAPHICS

ISSUE
 Le viene chiesto del seguente tema politico: {question}

EXPLANATION
 La seguente è una spiegazione del tema politico: {explanation}

NO EXPLANATION

PROS & CONS
 Le seguenti, sono argomentazioni tipiche a favore del tema politico: {pro arguments}
 Le seguenti, sono argomentazioni contrarie al tema politico: {contra arguments}

NO PROS & CONS

INSTRUCTION
 Scriva un commento { a favore | contro } il tema politico precedentemente menzionato. Ha disposizione fino a 500 caratteri.

Figure 8: Prompt elements in Italian.

Area	Feature
Surface	n_tokens
	n_sentences
	tokens_per_sentence
	n_characters
	avg_word_length
Morphology	n_types
	n_long_words
	n_lemmas
	n_VERB_VerbForm_Fin
	n_VERB_VerbForm_Inf
	n_VERB_VerbForm_Part
	n_VERB_Mood_Imp
	n_VERB_Mood_Ind
	n_VERB_Mood_Sub
	n_VERB_Tense_Past
	n_VERB_Tense_Pres
	n_VERB_Person_1
	n_VERB_Person_2
	n_VERB_Person_3
	n_VERB_Number_Plur
	n_VERB_Number_Sing
	n_NOUN_Gender_Fem
	n_NOUN_Gender_Masc
	n_NOUN_Number_Plur
	n_NOUN_Number_Sing
	n_PRON_PronType_Dem
	n_PRON_PronType_Int
	n_PRON_PronType_Prs
	n_PRON_PronType_Rel
	n_PRON_Gender_Fem
	n_PRON_Gender_Masc
	n_PRON_Number_Plur
	n_PRON_Number_Sing
	n_PRON_Person_1
	n_PRON_Person_3
	n_ADJ_Gender_Fem
	n_ADJ_Gender_Masc
	n_ADJ_Number_Plur
	n_ADJ_Number_Sing
	n_DET_Gender_Fem
	n_DET_Gender_Masc
	n_DET_Number_Plur
	n_DET_Number_Sing
	n_DET_Definite_Def
	n_DET_Definite_Ind

Table 3: List of linguistic features per feature area used in our analyses for the areas surface and morphology.

Area	Feature
Dependencies	tree_width
	tree_depth
	tree_branching
	n_noun_chunks
	n_dependency_cc
	n_dependency_dep
	n_dependency_punct
Emotion	n_positive_sentiment
	n_negative_sentiment
	sentiment_score
	avg_valence
	n_low_valence
	n_high_valence
	avg_arousal
	n_low_arousal
	n_high_arousal
	avg_dominance
	n_low_dominance
	n_high_dominance
	avg_intensity_anger
	avg_intensity_anticipation
	avg_intensity_disgust
	avg_intensity_fear
	avg_intensity_joy
	avg_intensity_sadness
	avg_intensity_surprise
	avg_intensity_trust
	n_low_intensity_anger
	n_low_intensity_anticipation
	n_low_intensity_disgust
	n_low_intensity_fear
	n_low_intensity_joy
	n_low_intensity_sadness
	n_low_intensity_surprise
	n_low_intensity_trust
	n_high_intensity_anger
	n_high_intensity_anticipation
	n_high_intensity_disgust
	n_high_intensity_fear
	n_high_intensity_joy
	n_high_intensity_sadness
	n_high_intensity_surprise
	n_high_intensity_trust
Information	compressibility
	entropy

Table 4: List of linguistic features per feature area used in our analyses for the areas syntactic dependencies, emotion, and information.

Area	Feature
Lexical Richness	lemma_token_ratio
	ttr
	rttr
	cttr
	herdan_c
	summer_index
	dugast_u
	maas_index
	n_lexical_tokens
	lexical_density
	n_hapax_legomena
	n_global_token_hapax_legomena
	n_global_lemma_hapax_legomena
	n_hapax_dislegomena
	n_global_token_hapax_dislegomena
	n_global_lemma_hapax_dislegomena
	sichel_s
	global_sichel_s
	giroud_index
	mtld
POS	hdd
	mattr
	mstr
	yule_k
	simpsons_d
	herdan_v
	pos_variability
	n_adj
	n_adp
	n_adv
	n_aux
	n_cconj
	n_det
Psycholinguistics	n_noun
	n_num
	n_pron
	n_propn
	n_punct
	n_sconj
	n_verb
	n_x
Psycholinguistics	avg_concreteness
	n_high_concreteness
	avg_sd_concreteness

Table 5: List of linguistic features per feature area used in our analyses for the areas lexical richness, POS, and psycholinguistics.

D Formulas and Equations

Equation 2 shows the formula for balanced accuracy (BA), where TP is the number of true positives, FN of false negatives, FP of false positives, and TN of true negatives.

$$BA = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad (2)$$

Area	Feature
Readability	n_syllables
	n_monosyllables
	n_polysyllables
	flesch_reading_ease
	flesch_kincaid_grade
	gunning_fog
	ari
	smog
	cli
	lix
	rix
Semantics	avg_n_synsets
	avg_n_synsets_noun
	avg_n_synsets_verb
	avg_n_synsets_adj
	n_low_synsets
	n_high_synsets
	n_high_synsets_noun
	n_high_synsets_verb
	n_high_synsets_adj
	n_low_synsets_noun
	n_low_synsets_verb
	n_low_synsets_adj
	n_low_synsets_adv
Named Entities	n_entities
	n_org
	n_gpe
	n_person
	n_money
	n_product
	n_time
	n_percent
	n_work_of_art
	n_quantity
	n_norp
	n_loc
	n_event
	n_ordinal
	n_fac
	n_law
	n_language

Table 6: List of linguistic features per feature area used in our analyses for the areas readability, semantics, and named entities.

E Additional Results: Difference Analysis

Figure 10 shows the forest plot for the interaction terms for the model fit in the difference analysis

The figure shows a prompt form for French (FR). It includes sections for Context, Sociodemographics, Issue, Explanation, Pros & Cons, and Instruction. The form is designed to collect information about a user's background, the political issue they are interested in, and their arguments for and against it, all in French.

Figure 9: Prompt elements in French.

(Section 5.1). Table 8 contains the full regression summary for the linguistic feature classifiers, and Table 9 for the SBERT classifiers.

F Additional Results: Stability Analysis

Figure 11 shows the forest plot for the interaction terms for the model fit in the difference analysis (Section 5.2). Table 10 contains the full regression summary for the BERTScore model, and Table 11 for the Rouge-L model.

G Additional Results: Generated Content Properties

G.1 Language Analysis

We used the langdetect package to flag LLM-generated arguments that did not match the prompted language. Italian-prompted texts exhibited the highest proportion of flagged arguments (8.22%), followed by German (6.91%) and French (0.62%). Two of the four models produced flagged texts, namely *Occiglot* and *llama_3.1-8b-instruct*, with the former accounting for 87-99% of the flagged instances across the prompted languages. A qualitative analysis of the flagged arguments showed that *occiglot* tends to generate French text more often and, to a lesser extent, Spanish text across the three prompted languages. This likely



Figure 10: Forest plot for interactions of the regression models for SBERT and linguistic features (elfen). Reference levels are: GPT-4 for LLM, Democracy, Media & Digitization for topic, AGAINST for stance, and DE for language.

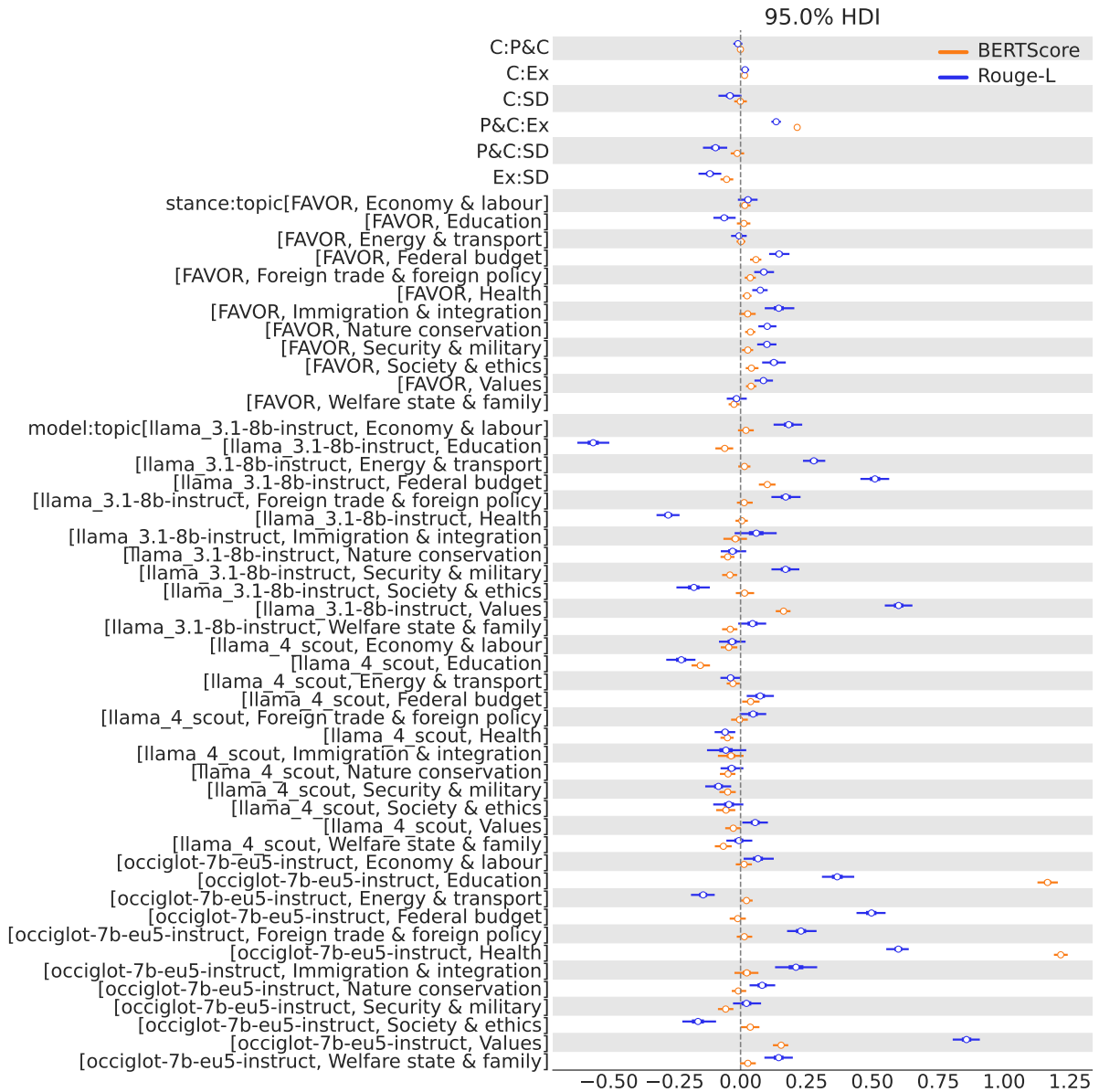


Figure 11: Forest plot for interactions of the regression models for Rouge-L and BERTScore. Reference levels are: GPT-4 for LLM, Democracy, Media & Digitization for topic, AGAINST for stance, and DE for language.

Feature Set	Feature Area	DE		FR		IT	
		BA	Std.Dev.	BA	Std.Dev.	BA	Std.Dev.
SBERT	–	91.2	15.9	92.7	9.3	75.3	14.6
Linguistic	All	0.967	0.096	0.979	0.054	0.967	0.079
	Lexical Richness	0.965	0.089	0.977	0.053	0.976	0.062
	Readability	0.939	0.097	0.946	0.072	0.896	0.120
	Surface	0.937	0.094	0.943	0.071	0.896	0.114
	Information	0.871	0.126	0.878	0.095	0.817	0.162
	Emotion	0.825	0.156	0.842	0.129	0.751	0.149
	POS	0.806	0.146	0.803	0.122	0.742	0.149
	Dependencies	0.798	0.134	0.779	0.108	0.757	0.158
	Morphological	0.779	0.162	0.800	0.136	0.688	0.143
	Semantics	0.773	0.147	0.767	0.127	0.694	0.153
	Psycholinguistics	0.722	0.168	0.620	0.096	0.692	0.174
	Named Entities	0.668	0.152	0.680	0.152	0.609	0.136

Table 7: Mean balanced accuracy (BA) and standard deviation (Std.Dev.) per feature set, feature area (for linguistic features) and language.

explains the lower proportion of flagged instances for the French prompts.

A small subset of the flagged texts consisted of non-alphabetical gibberish content. Upon closer inspection, the LLaMA model seemed to produce repetitiveness, symbol- and/or number-only texts (e.g., “(Favorable) (Favorable)” or “[1] [2] [3] [4]”), and occasional single-word outputs. In contrast, *occiglot* tended to generate sequences of random characters or words combined with excessive line breaks (e.g., “Trols \br Pré \br Lagu \br P \br Fels”). Apparently coherent flagged texts generated by the *llama* model often contained multilingual translations of the generated argument, sometimes extending the response with further explanations or follow-up questions. In contrast, *Occiglot* produced predominantly French or Spanish text.

G.2 Adherence to the Character Limit

Table 12 shows the percentage of non-adherence to the 500-character limit after removing empty answers per model and language. We observe that only the commercial model, GPT-4.1-mini, adheres to the restriction in over 95% of generated arguments. For the other models, we find large model-language pair differences: While LLaMA 3 generates longer text in more than 90% of cases across languages, *Occiglot* only exceeds 85% for FR, with the other two languages being well under this.

H Full Results: Linguistic Analysis

Tables 13 - 16 show point-biserial correlations between source (GPT, LLaMA 4, LLaMA 3, *Occiglot*, Human) and the linguistic features.

I Examples

I.1 Prompt and Argument Examples

Figure 12 shows the original German version of the example in Figure 2.

I.2 smartvote Questionnaire Structure

Figure 13 shows the structure of the smartvote questionnaire and showcases which elements we use in this work.

I.3 Rouge-L and BERTScore Triplets

Figures 14 - 17 show examples for triplets with high and low BERTScores per language, and Figures 20 - 23 show example triplets for Rouge-L.

J Implementation Details and AI Usage

J.1 Requirement Details

We use Python 3.12 for all our experiments and analyses. For linguistic features, we use *elfen* 1.3.1, for sentence embeddings *sentence-transformers* 5.4.1. For analysis, we use *bert-score* 0.3.13, *rouge-score* 0.1.2, *scikit-learn* 1.8.0 (Pedregosa et al., 2011), *statsmodels* 0.14.6 (Seabold and Perktold, 2010), and *bambi* 0.17.2 (Capretto et al.,

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
kappa	5.070	0.108	4.882	5.278	0.001	0.002	5760.0	2560.0	1.0
Intercept	3.256	0.067	3.124	3.379	0.002	0.001	1255.0	2140.0	1.0
context	0.013	0.038	-0.061	0.082	0.001	0.001	5429.0	3394.0	1.0
examples	0.084	0.042	0.005	0.160	0.001	0.001	5385.0	2664.0	1.0
explanation	-0.019	0.042	-0.102	0.057	0.001	0.001	5055.0	2989.0	1.0
sociodemographics	-0.195	0.037	-0.266	-0.129	0.001	0.001	4949.0	3263.0	1.0
model[LLaMA 3]	0.173	0.071	0.030	0.299	0.002	0.001	1114.0	1902.0	1.0
model[LLaMA 4]	-0.149	0.073	-0.287	-0.017	0.002	0.001	1341.0	1742.0	1.0
model[Occiglot]	-0.352	0.073	-0.491	-0.218	0.002	0.001	1305.0	1922.0	1.0
stance[FAVOR]	0.076	0.052	-0.021	0.171	0.001	0.001	1862.0	2539.0	1.0
topic[Economy & labour]	0.282	0.096	0.109	0.471	0.002	0.001	1952.0	2637.0	1.0
topic[Education]	0.167	0.097	-0.007	0.356	0.002	0.001	1845.0	2859.0	1.0
topic[Energy & transport]	-0.163	0.079	-0.310	-0.011	0.002	0.001	1656.0	2576.0	1.0
topic[Federal budget]	0.078	0.136	-0.172	0.332	0.003	0.002	2351.0	3203.0	1.0
topic[Foreign trade & foreign policy]	0.158	0.092	-0.012	0.330	0.002	0.001	1780.0	2208.0	1.0
topic[Health]	0.185	0.079	0.036	0.329	0.002	0.001	1321.0	2541.0	1.0
topic[Immigration & integration]	0.365	0.126	0.122	0.595	0.003	0.002	2334.0	2648.0	1.0
topic[Nature conservation]	0.113	0.091	-0.060	0.278	0.002	0.001	1731.0	2197.0	1.0
topic[Security & military]	0.043	0.092	-0.141	0.205	0.002	0.001	1575.0	2236.0	1.0
topic[Society & ethics]	0.141	0.095	-0.028	0.323	0.002	0.001	1515.0	2378.0	1.0
topic[Values]	-0.189	0.127	-0.413	0.059	0.003	0.002	2436.0	2600.0	1.0
topic[Welfare state & family]	0.143	0.091	-0.028	0.307	0.002	0.001	1592.0	2733.0	1.0
language[fr]	0.094	0.020	0.057	0.133	0.0	0.0	6594.0	2891.0	1.0
language[it]	0.340	0.035	0.277	0.407	0.0	0.001	5610.0	2866.0	1.0
context:examples	0.001	0.037	-0.075	0.068	0.001	0.001	5286.0	2910.0	1.0
context:explanation	-0.003	0.037	-0.079	0.061	0.0	0.001	6290.0	3044.0	1.0
context:sociodemographics	0.011	0.038	-0.060	0.081	0.001	0.001	5707.0	3370.0	1.0
examples:explanation	-0.018	0.038	-0.091	0.050	0.0	0.001	6079.0	3354.0	1.0
examples:sociodemographics	0.049	0.040	-0.025	0.124	0.001	0.001	5325.0	2965.0	1.0
explanation:sociodemographics	0.048	0.039	-0.025	0.122	0.001	0.001	5008.0	3070.0	1.0
stance:topic[FAVOR, Economy & labour]	-0.287	0.087	-0.440	-0.116	0.002	0.001	3098.0	3465.0	1.0
stance:topic[FAVOR, Education]	0.064	0.088	-0.111	0.222	0.002	0.001	3106.0	2794.0	1.0
stance:topic[FAVOR, Energy & transport]	-0.040	0.073	-0.168	0.104	0.001	0.001	2671.0	3164.0	1.0
stance:topic[FAVOR, Federal budget]	-0.117	0.118	-0.334	0.104	0.002	0.002	3946.0	3306.0	1.0
stance:topic[FAVOR, Foreign trade & foreign policy]	-0.089	0.085	-0.250	0.072	0.002	0.001	3016.0	2853.0	1.0
stance:topic[FAVOR, Health]	-0.121	0.073	-0.259	0.018	0.001	0.001	2544.0	2606.0	1.0
stance:topic[FAVOR, Immigration & integration]	-0.176	0.115	-0.392	0.039	0.002	0.002	4030.0	3016.0	1.0
stance:topic[FAVOR, Nature conservation]	-0.127	0.083	-0.280	0.031	0.001	0.001	3147.0	3082.0	1.0
stance:topic[FAVOR, Security & military]	-0.006	0.084	-0.174	0.139	0.002	0.001	2985.0	2919.0	1.0
stance:topic[FAVOR, Society & ethics]	0.064	0.089	-0.106	0.228	0.002	0.001	3295.0	3132.0	1.0
stance:topic[FAVOR, Values]	-0.039	0.112	-0.256	0.161	0.002	0.002	3771.0	3162.0	1.0
stance:topic[FAVOR, Welfare state & family]	-0.172	0.082	-0.330	-0.026	0.002	0.001	2624.0	2658.0	1.0
topic:model[Economy & labour, LLaMA 3]	0.040	0.120	-0.178	0.266	0.003	0.002	2274.0	2621.0	1.0
topic:model[Economy & labour, LLaMA 4]	0.112	0.119	-0.102	0.340	0.002	0.002	2359.0	3087.0	1.0
topic:model[Economy & labour, Occiglot]	0.074	0.121	-0.141	0.310	0.002	0.002	2523.0	2646.0	1.0
topic:model[Education, LLaMA 3]	-0.173	0.123	-0.402	0.054	0.003	0.002	2064.0	2318.0	1.0
topic:model[Education, LLaMA 4]	0.143	0.123	-0.074	0.386	0.003	0.002	2053.0	2746.0	1.0
topic:model[Education, Occiglot]	0.082	0.123	-0.155	0.304	0.002	0.002	2457.0	2527.0	1.0
topic:model[Energy & transport, LLaMA 3]	0.241	0.101	0.051	0.428	0.002	0.001	1930.0	2733.0	1.0
topic:model[Energy & transport, LLaMA 4]	0.086	0.102	-0.100	0.279	0.002	0.001	2096.0	2632.0	1.0
topic:model[Energy & transport, Occiglot]	0.338	0.104	0.138	0.523	0.002	0.001	2039.0	2735.0	1.0
topic:model[Federal budget, LLaMA 3]	0.133	0.167	-0.167	0.459	0.003	0.002	2649.0	2738.0	1.0
topic:model[Federal budget, LLaMA 4]	0.142	0.170	-0.180	0.462	0.003	0.002	2988.0	3224.0	1.0
topic:model[Federal budget, Occiglot]	-0.147	0.167	-0.460	0.166	0.003	0.002	3059.0	3083.0	1.0
topic:model[Foreign trade & foreign policy, LLaMA 3]	-0.009	0.115	-0.212	0.217	0.002	0.002	2453.0	2794.0	1.0
topic:model[Foreign trade & foreign policy, LLaMA 4]	0.098	0.116	-0.112	0.319	0.002	0.001	2389.0	3212.0	1.0
topic:model[Foreign trade & foreign policy, Occiglot]	0.285	0.113	0.071	0.491	0.003	0.002	1985.0	2772.0	1.0
topic:model[Health, LLaMA 3]	-0.232	0.102	-0.425	-0.044	0.002	0.002	1978.0	2680.0	1.0
topic:model[Health, LLaMA 4]	-0.005	0.103	-0.209	0.176	0.002	0.002	1837.0	2070.0	1.0
topic:model[Health, Occiglot]	0.061	0.102	-0.139	0.252	0.002	0.001	1924.0	2797.0	1.0
topic:model[Immigration & integration, LLaMA 3]	-0.036	0.158	-0.330	0.260	0.003	0.002	2811.0	3153.0	1.0
topic:model[Immigration & integration, LLaMA 4]	-0.044	0.158	-0.342	0.257	0.003	0.002	3091.0	3002.0	1.0
topic:model[Immigration & integration, Occiglot]	0.097	0.162	-0.191	0.422	0.003	0.002	2634.0	2969.0	1.0
topic:model[Nature conservation, LLaMA 3]	0.022	0.117	-0.203	0.237	0.002	0.002	2259.0	3022.0	1.0
topic:model[Nature conservation, LLaMA 4]	0.074	0.117	-0.145	0.296	0.003	0.002	2170.0	2556.0	1.0
topic:model[Nature conservation, Occiglot]	0.108	0.119	-0.113	0.328	0.002	0.002	2313.0	2685.0	1.0
topic:model[Security & military, LLaMA 3]	0.089	0.115	-0.144	0.284	0.003	0.002	1853.0	2850.0	1.0
topic:model[Security & military, LLaMA 4]	0.022	0.116	-0.181	0.254	0.003	0.002	1980.0	2221.0	1.0
topic:model[Security & military, Occiglot]	0.125	0.117	-0.084	0.353	0.002	0.002	2259.0	2780.0	1.0
topic:model[Society & ethics, LLaMA 3]	-0.065	0.122	-0.310	0.154	0.003	0.002	1893.0	2642.0	1.0
topic:model[Society & ethics, LLaMA 4]	0.098	0.122	-0.118	0.338	0.003	0.002	1826.0	2592.0	1.0
topic:model[Society & ethics, Occiglot]	-0.038	0.122	-0.267	0.184	0.003	0.002	2026.0	2801.0	1.0
topic:model[Values, LLaMA 3]	0.365	0.160	0.077	0.673	0.003	0.002	3006.0	2910.0	1.0
topic:model[Values, LLaMA 4]	-0.187	0.160	-0.477	0.117	0.003	0.002	2793.0	3306.0	1.0
topic:model[Values, Occiglot]	0.300	0.160	0.001	0.601	0.003	0.002	2717.0	2835.0	1.0
topic:model[Welfare state & family, LLaMA 3]	0.066	0.111	-0.140	0.269	0.003	0.002	1943.0	2804.0	1.0
topic:model[Welfare state & family, LLaMA 4]	0.171	0.113	-0.048	0.381	0.003	0.002	2042.0	2579.0	1.0
topic:model[Welfare state & family, Occiglot]	0.306	0.114	0.097	0.522	0.003	0.002	1919.0	2671.0	1.0

Table 8: Full regression summary for the difference analysis model for linguistic features.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
kappa	4.025	0.062	3.911	4.145	0.001	0.001	5226.0	3027.0	1.0
Intercept	2.384	0.066	2.265	2.513	0.002	0.001	1134.0	2078.0	1.0
context	0.084	0.038	0.008	0.150	0.001	0.001	4555.0	3061.0	1.0
examples	0.065	0.041	-0.014	0.139	0.001	0.001	4770.0	3474.0	1.0
explanation	-0.160	0.041	-0.235	-0.086	0.001	0.001	5464.0	3576.0	1.0
sociodemographics	-0.135	0.038	-0.208	-0.066	0.001	0.001	4192.0	2905.0	1.0
model[LLaMA 3]	-0.758	0.071	-0.887	-0.618	0.002	0.001	1205.0	1964.0	1.0
model[LLaMA 4]	0.148	0.074	0.001	0.279	0.002	0.001	1044.0	1923.0	1.0
model[Occiglot]	0.135	0.073	-0.009	0.265	0.002	0.001	1139.0	2151.0	1.0
stance[FAVOR]	0.095	0.051	0.002	0.195	0.001	0.001	1593.0	2317.0	1.0
topic[Economy & labour]	0.091	0.095	-0.084	0.279	0.002	0.001	1607.0	2490.0	1.0
topic[Education]	0.385	0.098	0.199	0.568	0.002	0.001	1642.0	2100.0	1.0
topic[Energy & transport]	-0.137	0.082	-0.291	0.016	0.002	0.001	1379.0	2395.0	1.0
topic[Federal budget]	-0.191	0.134	-0.439	0.061	0.003	0.002	2312.0	2780.0	1.0
topic[Foreign trade & foreign policy]	0.112	0.095	-0.080	0.279	0.002	0.001	1760.0	2527.0	1.0
topic[Health]	-0.225	0.079	-0.370	-0.074	0.002	0.001	1271.0	2353.0	1.0
topic[Immigration & integration]	0.656	0.121	0.446	0.895	0.003	0.002	2171.0	2794.0	1.0
topic[Nature conservation]	0.148	0.096	-0.032	0.326	0.002	0.001	1771.0	2219.0	1.0
topic[Security & military]	0.387	0.095	0.204	0.562	0.002	0.001	1664.0	2845.0	1.0
topic[Society & ethics]	0.450	0.099	0.258	0.624	0.002	0.001	1757.0	2543.0	1.0
topic[Values]	-0.409	0.122	-0.635	-0.177	0.003	0.002	2208.0	2660.0	1.0
topic[Welfare state & family]	0.171	0.091	0.003	0.340	0.002	0.001	1591.0	2458.0	1.0
language[fr]	-0.111	0.020	-0.150	-0.073	0.0	0.0	5506.0	3217.0	1.0
language[it]	-0.867	0.035	-0.934	-0.804	0.0	0.001	5234.0	3091.0	1.0
context:examples	0.021	0.036	-0.046	0.091	0.0	0.001	5477.0	2606.0	1.0
context:explanation	-0.015	0.036	-0.082	0.052	0.0	0.001	5780.0	2908.0	1.0
context:sociodemographics	-0.029	0.039	-0.100	0.045	0.001	0.001	4879.0	2931.0	1.0
examples:explanation	0.089	0.038	0.020	0.159	0.0	0.001	6880.0	3456.0	1.0
examples:sociodemographics	0.084	0.039	0.010	0.157	0.001	0.001	5468.0	3093.0	1.0
explanation:sociodemographics	0.164	0.039	0.092	0.236	0.001	0.001	5251.0	2924.0	1.0
stance:topic[FAVOR, Economy & labour]	-0.158	0.085	-0.332	-0.007	0.002	0.001	2670.0	2906.0	1.0
stance:topic[FAVOR, Education]	-0.114	0.087	-0.277	0.050	0.002	0.001	2824.0	2919.0	1.0
stance:topic[FAVOR, Energy & transport]	0.157	0.072	0.014	0.285	0.002	0.001	2203.0	2552.0	1.0
stance:topic[FAVOR, Federal budget]	-0.140	0.116	-0.349	0.090	0.002	0.002	3483.0	2816.0	1.0
stance:topic[FAVOR, Foreign trade & foreign policy]	0.134	0.084	-0.023	0.295	0.002	0.001	2549.0	2722.0	1.0
stance:topic[FAVOR, Health]	-0.044	0.072	-0.176	0.095	0.002	0.001	2219.0	2607.0	1.0
stance:topic[FAVOR, Immigration & integration]	-0.183	0.117	-0.397	0.032	0.002	0.002	3800.0	2910.0	1.0
stance:topic[FAVOR, Nature conservation]	-0.251	0.083	-0.411	-0.095	0.002	0.001	2605.0	2788.0	1.0
stance:topic[FAVOR, Security & military]	-0.241	0.080	-0.404	-0.105	0.002	0.001	2511.0	3.0	1.0
stance:topic[FAVOR, Society & ethics]	0.120	0.085	-0.027	0.292	0.002	0.001	2923.0	2935.0	1.0
stance:topic[FAVOR, Values]	0.053	0.110	-0.152	0.260	0.002	0.002	3279.0	3139.0	1.0
stance:topic[FAVOR, Welfare state & family]	-0.079	0.082	-0.230	0.072	0.002	0.001	2678.0	2984.0	1.0
topic:model[Economy & labour, LLaMA 3]	0.333	0.116	0.110	0.547	0.003	0.002	2076.0	2820.0	1.0
topic:model[Economy & labour, LLaMA 4]	0.032	0.119	-0.190	0.255	0.003	0.002	1913.0	2490.0	1.0
topic:model[Economy & labour, Occiglot]	0.194	0.121	-0.020	0.429	0.003	0.002	2010.0	2875.0	1.0
topic:model[Education, LLaMA 3]	-0.009	0.124	-0.242	0.217	0.003	0.002	2272.0	2925.0	1.0
topic:model[Education, LLaMA 4]	-0.079	0.126	-0.305	0.160	0.003	0.002	1835.0	2886.0	1.0
topic:model[Education, Occiglot]	0.075	0.125	-0.151	0.312	0.003	0.002	1914.0	2824.0	1.0
topic:model[Energy & transport, LLaMA 3]	0.371	0.099	0.196	0.566	0.002	0.001	1775.0	2595.0	1.0
topic:model[Energy & transport, LLaMA 4]	-0.124	0.102	-0.318	0.059	0.003	0.001	1632.0	2801.0	1.0
topic:model[Energy & transport, Occiglot]	0.276	0.102	0.083	0.469	0.003	0.001	1650.0	2525.0	1.0
topic:model[Federal budget, LLaMA 3]	0.494	0.162	0.174	0.792	0.003	0.002	2818.0	3234.0	1.0
topic:model[Federal budget, LLaMA 4]	0.117	0.163	-0.178	0.422	0.003	0.002	2616.0	3327.0	1.0
topic:model[Federal budget, Occiglot]	0.165	0.164	-0.126	0.482	0.003	0.002	2547.0	3134.0	1.0
topic:model[Foreign trade & foreign policy, LLaMA 3]	0.192	0.116	-0.024	0.412	0.002	0.002	2285.0	3085.0	1.0
topic:model[Foreign trade & foreign policy, LLaMA 4]	-0.032	0.118	-0.243	0.195	0.003	0.002	1900.0	2799.0	1.0
topic:model[Foreign trade & foreign policy, Occiglot]	0.315	0.120	0.093	0.533	0.003	0.002	2003.0	2917.0	1.0
topic:model[Health, LLaMA 3]	0.438	0.100	0.251	0.623	0.002	0.001	1704.0	2333.0	1.0
topic:model[Health, LLaMA 4]	-0.152	0.101	-0.334	0.041	0.003	0.001	1535.0	2489.0	1.0
topic:model[Health, Occiglot]	0.114	0.101	-0.073	0.304	0.003	0.001	1555.0	2898.0	1.0
topic:model[Immigration & integration, LLaMA 3]	0.249	0.161	-0.053	0.547	0.003	0.002	2739.0	2803.0	1.0
topic:model[Immigration & integration, LLaMA 4]	-0.127	0.161	-0.431	0.175	0.003	0.002	2369.0	3082.0	1.0
topic:model[Immigration & integration, Occiglot]	-0.143	0.162	-0.451	0.153	0.003	0.002	2661.0	3449.0	1.0
topic:model[Nature conservation, LLaMA 3]	0.167	0.116	-0.048	0.381	0.002	0.002	2397.0	2565.0	1.0
topic:model[Nature conservation, LLaMA 4]	0.154	0.121	-0.081	0.370	0.003	0.002	1864.0	2727.0	1.0
topic:model[Nature conservation, Occiglot]	-0.047	0.118	-0.263	0.182	0.003	0.002	1996.0	2460.0	1.0
topic:model[Security & military, LLaMA 3]	0.212	0.118	-0.005	0.428	0.002	0.002	2268.0	3208.0	1.0
topic:model[Security & military, LLaMA 4]	0.056	0.120	-0.160	0.281	0.003	0.002	1707.0	2121.0	1.0
topic:model[Security & military, Occiglot]	-0.060	0.117	-0.277	0.159	0.003	0.002	1995.0	2951.0	1.0
topic:model[Society & ethics, LLaMA 3]	-0.017	0.124	-0.256	0.202	0.003	0.002	2220.0	2965.0	1.0
topic:model[Society & ethics, LLaMA 4]	0.070	0.124	-0.164	0.302	0.003	0.002	1955.0	2580.0	1.0
topic:model[Society & ethics, Occiglot]	-0.202	0.124	-0.442	0.021	0.003	0.002	2137.0	2951.0	1.0
topic:model[Values, LLaMA 3]	0.516	0.157	0.237	0.828	0.003	0.002	2878.0	3302.0	1.0
topic:model[Values, LLaMA 4]	-0.198	0.156	-0.470	0.121	0.003	0.002	2533.0	3221.0	1.0
topic:model[Values, Occiglot]	-0.015	0.154	-0.298	0.273	0.003	0.002	2767.0	3024.0	1.0
topic:model[Welfare state & family, LLaMA 3]	0.347	0.112	0.128	0.550	0.003	0.002	1728.0	2747.0	1.0
topic:model[Welfare state & family, LLaMA 4]	0.164	0.114	-0.044	0.373	0.003	0.001	1777.0	2783.0	1.0
topic:model[Welfare state & family, Occiglot]	0.251	0.112	0.033	0.459	0.003	0.002	1882.0	2789.0	1.0

Table 9: Full regression summary for the difference analysis model for SBERT features.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
kappa	31.496	0.126	31.263	31.741	0.001	0.002	9081.0	2785.0	1.0
Intercept	1.322	0.012	1.297	1.344	0.0	0.0	2150.0	2782.0	1.0
context	-0.019	0.012	-0.042	0.003	0.0	0.0	4047.0	3384.0	1.0
examples	-0.036	0.013	-0.061	-0.011	0.0	0.0	4243.0	2951.0	1.0
explanation	-0.013	0.012	-0.036	0.011	0.0	0.0	4490.0	3243.0	1.0
sociodemographics	0.049	0.010	0.030	0.069	0.0	0.0	3709.0	3248.0	1.0
model[LLaMa 3]	-0.723	0.008	-0.739	-0.709	0.0	0.0	1335.0	2016.0	1.0
model[LLaMa 4]	0.126	0.009	0.110	0.142	0.0	0.0	1376.0	2190.0	1.0
model[Occiglot]	-0.537	0.008	-0.552	-0.521	0.0	0.0	1341.0	2102.0	1.0
stance[FAVOR]	0.031	0.006	0.019	0.041	0.0	0.0	1771.0	2580.0	1.0
topic[Economy & labour]	-0.013	0.013	-0.036	0.012	0.0	0.0	2138.0	2385.0	1.0
topic[Education]	0.016	0.014	-0.012	0.042	0.0	0.0	2556.0	2533.0	1.0
topic[Energy & transport]	-0.016	0.010	-0.034	0.002	0.0	0.0	1748.0	2301.0	1.0
topic[Federal budget]	-0.131	0.014	-0.157	-0.106	0.0	0.0	2041.0	2906.0	1.0
topic[Foreign trade & foreign policy]	-0.052	0.012	-0.074	-0.028	0.0	0.0	2008.0	2727.0	1.0
topic[Health]	-0.046	0.010	-0.064	-0.026	0.0	0.0	1830.0	2518.0	1.0
topic[Immigration & integration]	0.004	0.019	-0.032	0.042	0.0	0.0	2648.0	2508.0	1.0
topic[Nature conservation]	-0.016	0.011	-0.037	0.006	0.0	0.0	2070.0	2940.0	1.0
topic[Security & military]	0.007	0.012	-0.016	0.030	0.0	0.0	2222.0	2749.0	1.0
topic[Society & ethics]	-0.106	0.015	-0.133	-0.078	0.0	0.0	2490.0	2795.0	1.0
topic[Values]	-0.131	0.012	-0.153	-0.108	0.0	0.0	2274.0	2882.0	1.0
topic[Welfare state & family]	0.036	0.013	0.012	0.060	0.0	0.0	2042.0	2763.0	1.0
language[fr]	0.101	0.003	0.097	0.106	0.0	0.0	7021.0	3361.0	1.0
language[it]	0.053	0.003	0.047	0.059	0.0	0.0	7070.0	3210.0	1.0
context:examples	-0.001	0.005	-0.010	0.008	0.0	0.0	9158.0	2754.0	1.0
context:explanation	0.015	0.005	0.007	0.024	0.0	0.0	7041.0	2428.0	1.0
context:sociodemographics	-0.001	0.012	-0.023	0.022	0.0	0.0	3925.0	3082.0	1.0
examples:explanation	0.215	0.005	0.205	0.224	0.0	0.0	9003.0	3017.0	1.0
examples:sociodemographics	-0.012	0.013	-0.038	0.012	0.0	0.0	4249.0	2987.0	1.0
explanation:sociodemographics	-0.053	0.012	-0.076	-0.030	0.0	0.0	4528.0	3201.0	1.0
stance:topic[FAVOR, Economy & labour]	0.016	0.011	-0.003	0.038	0.0	0.0	3273.0	2961.0	1.0
stance:topic[FAVOR, Education]	0.013	0.013	-0.012	0.037	0.0	0.0	4229.0	2944.0	1.0
stance:topic[FAVOR, Energy & transport]	0.002	0.009	-0.014	0.017	0.0	0.0	2854.0	3122.0	1.0
stance:topic[FAVOR, Federal budget]	0.058	0.011	0.038	0.079	0.0	0.0	3699.0	3291.0	1.0
stance:topic[FAVOR, Foreign trade & foreign policy]	0.037	0.011	0.016	0.056	0.0	0.0	3715.0	3133.0	1.0
stance:topic[FAVOR, Health]	0.024	0.009	0.007	0.041	0.0	0.0	2865.0	3029.0	1.0
stance:topic[FAVOR, Immigration & integration]	0.027	0.016	-0.003	0.057	0.0	0.0	4940.0	3100.0	1.0
stance:topic[FAVOR, Nature conservation]	0.037	0.010	0.019	0.056	0.0	0.0	3263.0	2718.0	1.0
stance:topic[FAVOR, Security & military]	0.027	0.011	0.006	0.047	0.0	0.0	3713.0	2629.0	1.0
stance:topic[FAVOR, Society & ethics]	0.041	0.012	0.018	0.065	0.0	0.0	3822.0	3060.0	1.0
stance:topic[FAVOR, Values]	0.040	0.010	0.021	0.059	0.0	0.0	3559.0	2972.0	1.0
stance:topic[FAVOR, Welfare state & family]	-0.025	0.011	-0.045	-0.005	0.0	0.0	3431.0	3071.0	1.0
model:topic[LLaMa 3, Economy & labour]	0.020	0.015	-0.009	0.049	0.0	0.0	2412.0	2708.0	1.0
model:topic[LLaMa 3, Education]	-0.060	0.017	-0.092	-0.028	0.0	0.0	3001.0	2941.0	1.0
model:topic[LLaMa 3, Energy & transport]	0.015	0.012	-0.008	0.036	0.0	0.0	2047.0	2727.0	1.0
model:topic[LLaMa 3, Federal budget]	0.101	0.016	0.070	0.130	0.0	0.0	2232.0	2975.0	1.0
model:topic[LLaMa 3, Foreign trade & foreign policy]	0.014	0.015	-0.015	0.043	0.0	0.0	2111.0	2910.0	1.0
model:topic[LLaMa 3, Health]	0.005	0.012	-0.019	0.026	0.0	0.0	2041.0	2738.0	1.0
model:topic[LLaMa 3, Immigration & integration]	-0.019	0.023	-0.065	0.020	0.0	0.0	3178.0	3130.0	1.0
model:topic[LLaMa 3, Nature conservation]	-0.049	0.014	-0.074	-0.023	0.0	0.0	2196.0	2695.0	1.0
model:topic[LLaMa 3, Security & military]	-0.041	0.015	-0.069	-0.015	0.0	0.0	2359.0	2863.0	1.0
model:topic[LLaMa 3, Society & ethics]	0.015	0.018	-0.018	0.049	0.0	0.0	2626.0	2704.0	1.0
model:topic[LLaMa 3, Values]	0.163	0.015	0.136	0.190	0.0	0.0	2481.0	2892.0	1.0
model:topic[LLaMa 3, Welfare state & family]	-0.040	0.015	-0.067	-0.012	0.0	0.0	2092.0	2752.0	1.0
model:topic[LLaMa 4, Economy & labour]	-0.045	0.016	-0.075	-0.014	0.0	0.0	2364.0	2780.0	1.0
model:topic[LLaMa 4, Education]	-0.153	0.018	-0.185	-0.118	0.0	0.0	2819.0	3498.0	1.0
model:topic[LLaMa 4, Energy & transport]	-0.029	0.013	-0.052	-0.005	0.0	0.0	1971.0	2377.0	1.0
model:topic[LLaMa 4, Federal budget]	0.038	0.017	0.007	0.071	0.0	0.0	2277.0	2750.0	1.0
model:topic[LLaMa 4, Foreign trade & foreign policy]	-0.004	0.016	-0.034	0.026	0.0	0.0	2286.0	2785.0	1.0
model:topic[LLaMa 4, Health]	-0.050	0.013	-0.075	-0.027	0.0	0.0	1899.0	2556.0	1.0
model:topic[LLaMa 4, Immigration & integration]	-0.036	0.025	-0.081	0.011	0.0	0.0	2967.0	2634.0	1.0
model:topic[LLaMa 4, Nature conservation]	-0.048	0.015	-0.078	-0.021	0.0	0.0	2324.0	2947.0	1.0
model:topic[LLaMa 4, Security & military]	-0.049	0.016	-0.080	-0.020	0.0	0.0	2392.0	3053.0	1.0
model:topic[LLaMa 4, Society & ethics]	-0.055	0.019	-0.090	-0.020	0.0	0.0	2633.0	3113.0	1.0
model:topic[LLaMa 4, Values]	-0.027	0.015	-0.058	-0.001	0.0	0.0	2514.0	3006.0	1.0
model:topic[LLaMa 4, Welfare state & family]	-0.065	0.016	-0.098	-0.037	0.0	0.0	2391.0	2941.0	1.0
model:topic[Occiglot, Economy & labour]	0.013	0.016	-0.016	0.043	0.0	0.0	2576.0	2801.0	1.0
model:topic[Occiglot, Education]	1.165	0.020	1.128	1.202	0.0	0.0	2950.0	2917.0	1.0
model:topic[Occiglot, Energy & transport]	0.023	0.012	-0.001	0.045	0.0	0.0	1786.0	3028.0	1.0
model:topic[Occiglot, Federal budget]	-0.011	0.016	-0.041	0.018	0.0	0.0	2687.0	3146.0	1.0
model:topic[Occiglot, Foreign trade & foreign policy]	0.014	0.015	-0.014	0.042	0.0	0.0	2229.0	2943.0	1.0
model:topic[Occiglot, Health]	1.215	0.013	1.191	1.242	0.0	0.0	2055.0	3103.0	1.0
model:topic[Occiglot, Immigration & integration]	0.023	0.023	-0.023	0.065	0.0	0.0	3080.0	2939.0	1.0
model:topic[Occiglot, Nature conservation]	-0.008	0.014	-0.036	0.017	0.0	0.0	2311.0	2689.0	1.0
model:topic[Occiglot, Security & military]	-0.057	0.015	-0.085	-0.029	0.0	0.0	2312.0	3034.0	1.0
model:topic[Occiglot, Society & ethics]	0.037	0.018	0.004	0.071	0.0	0.0	2811.0	3024.0	1.0
model:topic[Occiglot, Values]	0.154	0.015	0.124	0.180	0.0	0.0	2585.0	3140.0	1.0
model:topic[Occiglot, Welfare state & family]	0.028	0.015	-0.0	0.058	0.0	0.0	2367.0	3063.0	1.0

Table 10: Full regression summary for the stability analysis model for BERTScore.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
kappa	7.896	0.031	7.843	7.956	0.000	0.001	8657.0	2810.000	1.0
Intercept	-0.456	0.023	-0.499	-0.415	0.000	0.000	2594.0	3050.000	1.0
context	0.020	0.022	-0.022	0.063	0.000	0.000	3558.0	3182.0	1.0
examples	-0.104	0.024	-0.149	-0.057	0.000	0.000	4373.0	3142.0	1.0
explanation	-0.125	0.022	-0.171	-0.086	0.000	0.000	4484.0	2812.0	1.0
sociodemographics	0.129	0.020	0.093	0.167	0.000	0.000	3524.0	3292.0	1.0
model[LLaMA 3]	-1.302	0.015	-1.330	-1.275	0.000	0.000	1779.0	2542.0	1.0
model[LLaMA 4]	0.221	0.014	0.197	0.248	0.000	0.000	1697.0	2267.0	1.0
model[Occiglot]	-1.502	0.015	-1.532	-1.475	0.000	0.000	1793.0	2578.0	1.0
stance[FAVOR]	0.005	0.010	-0.012	0.026	0.000	0.000	1991.0	2845.0	1.0
topic[Economy & labour]	-0.076	0.021	-0.113	-0.036	0.000	0.000	3048.0	3121.0	1.0
topic[Education]	0.091	0.023	0.048	0.133	0.000	0.000	2977.0	2832.0	1.0
topic[Energy & transport]	-0.060	0.016	-0.090	-0.030	0.000	0.000	2225.0	2617.0	1.0
topic[Federal budget]	-0.421	0.022	-0.461	-0.379	0.000	0.000	2987.0	2964.0	1.0
topic[Foreign trade & foreign policy]	-0.164	0.021	-0.201	-0.123	0.000	0.000	2840.000	2876.0	1.0
topic[Health]	-0.091	0.016	-0.119	-0.059	0.000	0.000	2277.0	3101.0	1.0
topic[Immigration & integration]	-0.187	0.031	-0.244	-0.127	0.001	0.000	3054.0	3063.0	1.0
topic[Nature conservation]	-0.054	0.018	-0.085	-0.018	0.000	0.000	2528.0	3081.0	1.0
topic[Security & military]	-0.047	0.020	-0.084	-0.008	0.000	0.000	2702.0	3151.0	1.0
topic[Society & ethics]	-0.113	0.023	-0.159	-0.071	0.000	0.000	2979.0	3273.0	1.0
topic[Values]	-0.447	0.020	-0.485	-0.410	0.000	0.000	2856.0	2762.0	1.0
topic[Welfare state & family]	-0.026	0.021	-0.065	0.012	0.000	0.000	2505.0	2417.0	1.0
language[fr]	0.454	0.005	0.445	0.462	0.000	0.000	7937.0	3268.0	1.0
language[it]	0.208	0.006	0.198	0.220	0.000	0.000	8.0	2891.0	1.0
context:examples	-0.010	0.009	-0.027	0.006	0.000	0.000	9134.0	2483.0	1.0
context:explanation	0.017	0.008	0.003	0.033	0.000	0.000	9367.0	3007.0	1.0
context:sociodemographics	-0.040	0.022	-0.080	0.004	0.000	0.000	3579.0	2980.000	1.0
examples:explanation	0.135	0.009	0.118	0.153	0.000	0.000	8774.0	2934.0	1.0
examples:sociodemographics	-0.095	0.024	-0.141	-0.053	0.000	0.000	4392.0	3416.0	1.0
explanation:sociodemographics	-0.117	0.022	-0.158	-0.075	0.000	0.000	4526.0	2965.0	1.0
stance:topic[FAVOR, Economy & labour]	0.028	0.019	-0.009	0.063	0.000	0.000	3863.0	3081.0	1.0
stance:topic[FAVOR, Education]	-0.061	0.021	-0.100	-0.019	0.000	0.000	4610.000	2985.0	1.0
stance:topic[FAVOR, Energy & transport]	-0.006	0.015	-0.034	0.023	0.000	0.000	2730.000	2981.0	1.0
stance:topic[FAVOR, Federal budget]	0.146	0.020	0.110	0.184	0.000	0.000	3713.0	3228.0	1.0
stance:topic[FAVOR, Foreign trade & foreign policy]	0.088	0.019	0.052	0.123	0.000	0.000	4155.0	3266.0	1.0
stance:topic[FAVOR, Health]	0.075	0.015	0.046	0.102	0.000	0.000	2915.0	3343.0	1.0
stance:topic[FAVOR, Immigration & integration]	0.145	0.029	0.091	0.198	0.000	0.000	4637.0	3127.0	1.0
stance:topic[FAVOR, Nature conservation]	0.101	0.017	0.068	0.134	0.000	0.000	3388.0	3143.0	1.0
stance:topic[FAVOR, Security & military]	0.100	0.019	0.066	0.136	0.000	0.000	4040.000	3111.0	1.0
stance:topic[FAVOR, Society & ethics]	0.127	0.023	0.085	0.171	0.000	0.000	4332.0	2760.000	1.0
stance:topic[FAVOR, Values]	0.087	0.018	0.053	0.120	0.000	0.000	3690.000	3182.0	1.0
stance:topic[FAVOR, Welfare state & family]	-0.016	0.019	-0.052	0.020	0.000	0.000	3796.0	2846.0	1.0
model:topic[LLaMA 3, Economy & labour]	0.182	0.028	0.130	0.234	0.000	0.000	3426.0	3178.0	1.0
model:topic[LLaMA 3, Education]	-0.559	0.032	-0.617	-0.501	0.001	0.000	3282.0	3613.0	1.0
model:topic[LLaMA 3, Energy & transport]	0.278	0.022	0.240	0.321	0.000	0.000	2581.0	3204.0	1.0
model:topic[LLaMA 3, Federal budget]	0.511	0.028	0.459	0.564	0.000	0.000	3382.0	3312.0	1.0
model:topic[LLaMA 3, Foreign trade & foreign policy]	0.171	0.028	0.120	0.226	0.001	0.000	2915.0	2790.000	1.0
model:topic[LLaMA 3, Health]	-0.275	0.022	-0.315	-0.232	0.000	0.000	2717.0	3138.0	1.0
model:topic[LLaMA 3, Immigration & integration]	0.059	0.041	-0.019	0.134	0.001	0.001	3947.0	3732.0	1.0
model:topic[LLaMA 3, Nature conservation]	-0.030	0.024	-0.074	0.019	0.000	0.000	2783.0	3135.0	1.0
model:topic[LLaMA 3, Security & military]	0.171	0.027	0.117	0.219	0.000	0.000	3095.0	3272.0	1.0
model:topic[LLaMA 3, Society & ethics]	-0.178	0.032	-0.239	-0.117	0.001	0.000	3594.0	3307.0	1.0
model:topic[LLaMA 3, Values]	0.600	0.027	0.549	0.650	0.000	0.000	2969.0	3093.0	1.0
model:topic[LLaMA 3, Welfare state & family]	0.045	0.028	-0.006	0.096	0.001	0.000	2899.0	3158.0	1.0
model:topic[LLaMA 4, Economy & labour]	-0.032	0.026	-0.077	0.019	0.000	0.000	3133.0	3379.0	1.0
model:topic[LLaMA 4, Education]	-0.226	0.029	-0.281	-0.175	0.000	0.000	3302.0	3136.0	1.0
model:topic[LLaMA 4, Energy & transport]	-0.038	0.020	-0.074	-0.002	0.000	0.000	2469.0	3028.0	1.0
model:topic[LLaMA 4, Federal budget]	0.074	0.026	0.023	0.123	0.000	0.000	3066.0	3455.0	1.0
model:topic[LLaMA 4, Foreign trade & foreign policy]	0.048	0.026	-0.001	0.097	0.000	0.000	2958.0	2824.0	1.0
model:topic[LLaMA 4, Health]	-0.059	0.020	-0.095	-0.020	0.000	0.000	2376.0	3003.0	1.0
model:topic[LLaMA 4, Immigration & integration]	-0.056	0.038	-0.126	0.016	0.001	0.000	3427.0	3383.0	1.0
model:topic[LLaMA 4, Nature conservation]	-0.034	0.023	-0.076	0.007	0.000	0.000	2633.0	2988.0	1.0
model:topic[LLaMA 4, Security & military]	-0.084	0.025	-0.130	-0.035	0.000	0.000	2992.0	3167.0	1.0
model:topic[LLaMA 4, Society & ethics]	-0.044	0.029	-0.097	0.012	0.001	0.000	3053.0	3429.0	1.0
model:topic[LLaMA 4, Values]	0.055	0.025	0.007	0.100	0.000	0.000	2956.0	3399.0	1.0
model:topic[LLaMA 4, Welfare state & family]	-0.007	0.025	-0.053	0.041	0.000	0.000	2820.000	2970.000	1.0
model:topic[Occiglot, Economy & labour]	0.066	0.029	0.009	0.119	0.001	0.000	3136.0	3241.0	1.0
model:topic[Occiglot, Education]	0.367	0.031	0.309	0.426	0.001	0.000	3316.0	3263.0	1.0
model:topic[Occiglot, Energy & transport]	-0.142	0.023	-0.184	-0.098	0.000	0.000	2554.0	3229.0	1.0
model:topic[Occiglot, Federal budget]	0.496	0.029	0.444	0.550	0.001	0.000	2969.0	3181.0	1.0
model:topic[Occiglot, Foreign trade & foreign policy]	0.229	0.029	0.176	0.283	0.001	0.000	3040.000	3328.0	1.0
model:topic[Occiglot, Health]	0.599	0.022	0.557	0.640	0.000	0.000	2493.0	3190.000	1.0
model:topic[Occiglot, Immigration & integration]	0.210	0.041	0.133	0.287	0.001	0.001	3832.0	3157.0	1.0
model:topic[Occiglot, Nature conservation]	0.082	0.025	0.038	0.132	0.000	0.000	3337.0	3191.0	1.0
model:topic[Occiglot, Security & military]	0.022	0.028	-0.029	0.074	0.000	0.000	3063.0	3089.0	1.0
model:topic[Occiglot, Society & ethics]	-0.162	0.033	-0.221	-0.097	0.001	0.000	3204.0	3283.0	1.0
model:topic[Occiglot, Values]	0.858	0.027	0.809	0.907	0.000	0.000	2928.0	3625.0	1.0
model:topic[Occiglot, Welfare state & family]	0.145	0.028	0.090	0.194	0.001	0.000	2693.0	3045.0	1.0

Table 11: Full regression summary for the stability analysis model for Rouge-L.

Prompt	Arguments
SOCIODEMOGRAPHICS Sie gehören der folgenden soziodemografischen Gruppe an: Politisches Spektrum - Mitte und Liberal	Default Ein Verbot der Neuzulassung von Personenwagen mit Verbrennungsmotoren ab 2035 ist ein wichtiger Schritt für den Klimaschutz. Die Schweiz muss ihre CO2-Emissionen reduzieren, um die Ziele des Pariser Abkommens zu erreichen. Elektromobilität fördert saubere Luft, weniger Lärm und langfristig niedrigere Betriebskosten. Zudem stärkt das Verbot Innovationen und den Ausbau nachhaltiger Technologien, was der Umwelt und der Wirtschaft zugutekommt.
ISSUE Sie werden nach folgendem politischen Thema gefragt: Soll die Schweiz Neuzulassungen von Personenwagen mit Verbrennungsmotor ab 2035 verbieten?	SD Ein Verbot der Neuzulassung von Verbrennungsmotoren ab 2035 ist ein wichtiger Schritt für den Klimaschutz und die Innovationsförderung. Die Schweiz zeigt damit Verantwortung und setzt Anreize für umweltfreundliche Technologien. Gleichzeitig braucht es eine ausgewogene Umsetzung mit ausreichender Förderung von alternativen Antrieben und einer sozialverträglichen Begleitpolitik, um den Übergang für alle Bürgerinnen und Bürger fair zu gestalten.
INSTRUCTION Schreiben Sie einen Kommentar pro die oben genannte politische Frage. Sie können bis zu 500 Zeichen verwenden.	HUMAN Dies würde zwar nur die wenigsten Menschen betreffen, da der Siegeszug des E-Autos nicht mehr aufzuhalten ist. Dennoch sollten Autos verboten werden, die nicht Klimaneutral betrieben werden können. Dies muss also nicht zwingen ein Verbot es Verbrenners sein. Eher ein Verbot von konventionellem Treibstoff.

Figure 12: Example prompt, a human answer, and two outputs by GPT-4. Note that Default is generated using the prompt elements ISSUE and INSTRUCTION, and SD additionally has the block SOCIODEMOGRAPHICS.

1. Sozialstaat & Familie (0/6)

1. Befürworten Sie eine Erhöhung des Rentenalters (z.B. auf 67 Jahre)?

Ja

Eher ja

Eher nein

Nein

Kommentar zur Antwort erfassen

0 / 500

Weniger Informationen

Erläuterungen | Argumente

Figure 13: Overview of the smartvote questionnaire for an example question. Marking of the elements used in this work in red.

Model	DE	FR	IT
LLaMA 4 Scout	53.54	41.97	22.64
LLaMA 3.1 8B instruct	98.27	95.70	93.27
Occiglot 7B EU5 instruct	52.36	87.90	41.42
GPT 4.1 mini	4.96	3.26	0.54

Feature Area	Feature	Point-Biserial Correlation Coefficient	p-value
Surface	raw_sequence_length	-0.52	0.00
	n_tokens	-0.56	0.00
	n_sentences	-0.44	0.00
	tokens_per_sentence	0.19	0.00
	n_characters	-0.52	0.00
	avg_word_length	0.09	0.00
	n_types	-0.31	0.00
	n_long_words	0.33	0.00
	n_lemmas	-0.29	0.00
Morphology	n_VERB_VerbForm_Fin	0.12	0.00
	n_VERB_VerbForm_Inf	0.19	0.00
	n_VERB_VerbForm_Part	-0.04	0.00
	n_VERB_Mood_Imp	-0.09	0.00
	n_VERB_Mood_Ind	0.07	0.00
	n_VERB_Mood_Sub	-0.03	0.00
	n_VERB_Tense_Past	-0.08	0.00
	n_VERB_Tense_Pres	0.16	0.00
	n_VERB_Person_1	0.01	0.00
	n_VERB_Person_2	-0.15	0.00
	n_VERB_Person_3	0.16	0.00
	n_VERB_Number_Plur	-0.03	0.00
	n_VERB_Number_Sing	0.11	0.00
	n_NOUN_Gender_Fem	0.23	0.00
	n_NOUN_Gender_Masc	0.10	0.00
	n_NOUN_Number_Plur	0.27	0.00
	n_NOUN_Number_Sing	0.09	0.00
	n_PRON_PronType_Dem	-0.02	0.00
	n_PRON_PronType_Int	-0.05	0.00
	n_PRON_PronType_Prs	-0.15	0.00
	n_PRON_PronType_Rel	0.08	0.00
	n_PRON_Gender_Fem	-0.03	0.00
	n_PRON_Gender_Masc	-0.00	0.01
	n_PRON_Number_Plur	-0.19	0.00
	n_PRON_Number_Sing	-0.06	0.00
	n_PRON_Person_1	-0.08	0.00
	n_PRON_Person_3	-0.17	0.00
	n_ADJ_Gender_Fem	0.23	0.00
	n_ADJ_Gender_Masc	0.14	0.00
	n_ADJ_Number_Plur	0.28	0.00
	n_ADJ_Number_Sing	0.16	0.00
	n_DET_Gender_Fem	0.08	0.00
	n_DET_Gender_Masc	-0.01	0.00
	n_DET_Number_Plur	0.11	0.00
	n_DET_Number_Sing	0.09	0.00
	n_DET_Definite_Def	0.04	0.00
	n_DET_Definite_Ind	0.20	0.00

Table 13: Point-Biserial Correlation and p-value per feature against source (GPT, LLaMA 4, LLaMA 3, Occiglot, Human).

Feature Area	Feature	Point-Biserial Correlation Coefficient	p-value
Dependency	tree_width	-0.12	0.00
	tree_depth	0.33	0.00
	tree_branching	0.16	0.00
	n_noun_chunks	0.14	0.00
	n_dependency_cc	0.11	0.00
	n_dependency_dep	-0.21	0.00
	n_dependency_punct	-0.19	0.00
Emotion	n_positive_sentiment	0.46	0.00
	n_negative_sentiment	0.30	0.00
	sentiment_score	0.24	0.00
	avg_valence	0.10	0.00
	n_low_valence	0.31	0.00
	n_high_valence	0.50	0.00
	avg_arousal	0.12	0.00
	n_low_arousal	0.37	0.00
	n_high_arousal	0.33	0.00
	avg_dominance	0.15	0.00
	n_low_dominance	0.15	0.00
	n_high_dominance	0.49	0.00
	avg_intensity_anger	-0.08	0.00
	avg_intensity_anticipation	0.02	0.00
	avg_intensity_disgust	-0.12	0.00
	avg_intensity_fear	-0.01	0.00
	avg_intensity_joy	-0.03	0.00
	avg_intensity_sadness	-0.08	0.00
	avg_intensity_surprise	-0.07	0.00
	avg_intensity_trust	0.13	0.00
	n_low_intensity_anger	0.10	0.00
	n_low_intensity_anticipation	0.02	0.00
	n_low_intensity_disgust	0.04	0.00
	n_low_intensity_fear	0.18	0.00
	n_low_intensity_joy	0.15	0.00
	n_low_intensity_sadness	0.11	0.00
	n_low_intensity_surprise	0.13	0.00
	n_low_intensity_trust	0.07	0.00
	n_high_intensity_anger	0.03	0.00
	n_high_intensity_anticipation	0.05	0.00
	n_high_intensity_disgust	0.07	0.00
	n_high_intensity_fear	0.12	0.00
	n_high_intensity_joy	0.04	0.00
	n_high_intensity_sadness	0.01	0.00
	n_high_intensity_surprise	0.06	0.00
	n_high_intensity_trust	0.24	0.00
Information	compressibility	0.00	0.08
	entropy	0.05	0.00

Table 14: Point-Biserial Correlation and p-value per feature against source (GPT, LLaMA 4, LLaMA 3, Occiglot, Human) (continued).

Feature Area	Feature	Point-Biserial Correlation Coefficient	p-value
Lexical Richness	lemma_token_ratio	0.54	0.00
	ttr	0.57	0.00
	rttr	0.24	0.00
	cttr	0.24	0.00
	herdan_c	0.55	0.00
	summer_index	0.29	0.00
	dugast_u	-0.06	0.00
	maas_index	-0.22	0.00
	n_lexical_tokens	0.29	0.00
	lexical_density	0.29	0.00
	n_hapax_legomena	0.54	0.00
	n_global_token_hapax_legomena	-0.18	0.00
	n_global_lemma_hapax_legomena	-0.16	0.00
	n_hapax_dislegomena	0.56	0.00
	n_global_token_hapax_dislegomena	-0.21	0.00
	n_global_lemma_hapax_dislegomena	-0.18	0.00
	sichel_s	0.49	0.00
	global_sichel_s	-0.22	0.00
	giroud_index	0.24	0.00
	mtld	0.43	0.00
	hdd	nan	nan
	mattr	0.15	0.00
	msttr	0.14	0.00
	yule_k	-0.01	0.00
	simpsons_d	-0.01	0.00
	herdan_v	-0.01	0.00
	pos_variability	0.10	0.00
POS	n_adj	0.29	0.00
	n_adp	0.13	0.00
	n_adv	-0.02	0.00
	n_aux	-0.12	0.00
	n_cconj	0.23	0.00
	n_det	0.14	0.00
	n_noun	0.23	0.00
	n_num	-0.10	0.00
	n_pron	-0.16	0.00
	n_propn	-0.09	0.00
	n_punct	-0.19	0.00
	n_sconj	-0.04	0.00
	n_verb	0.21	0.00
	n_x	-0.07	0.00
Psycholinguistics	avg_concreteness	-0.01	0.00
	n_high_concreteness	0.16	0.00
	avg_sd_concreteness	-0.04	0.00
Readability	n_syllables	-0.50	0.00
	n_monosyllables	-0.06	0.00
	n_polysyllables	0.29	0.00
	flesch_reading_ease	-0.36	0.00
	flesch_kincaid_grade	0.36	0.00
	gunning_fog	0.32	0.00
	ari	0.14	0.00
	smog	0.35	0.00
	cli	0.11	0.00
	lix	0.35	0.00
	rix	0.38	0.00

Table 15: Point-Biserial Correlation and p-value per feature against source (GPT, LLaMA 4, LLaMA 3, Occiglot, Human) (continued).

Feature Area	Feature	Point-Biserial Correlation Coefficient	p-value
Semantics	avg_n_synsets	-0.03	0.00
	avg_n_synsets_noun	-0.05	0.00
	avg_n_synsets_verb	0.00	0.15
	avg_n_synsets_adj	-0.04	0.00
	n_low_synsets	0.26	0.00
	n_high_synsets	0.07	0.00
	n_high_synsets_noun	0.03	0.00
	n_high_synsets_verb	0.13	0.00
	n_high_synsets_adj	0.04	0.00
	n_low_synsets_noun	0.21	0.00
	n_low_synsets_verb	0.13	0.00
	n_low_synsets_adj	0.33	0.00
	n_low_synsets_adv	-0.04	0.00
Named Entities	n_entities	-0.15	0.00
	n_org	-0.09	0.00

Table 16: Point-Biserial Correlation and p-value per feature against source (GPT, LLaMA 4, LLaMA 3, Occiglot, Human) (continued).

Sì, la Svizzera dovrebbe puntare a stipulare un ampio accordo di libero scambio con gli Stati Uniti. Questo accordo potrebbe aumentare le esportazioni svizzere negli USA, favorendo la crescita economica. Ridurre i dazi doganali, specialmente nel settore agricolo, potrebbe aiutare le imprese svizzere a essere più competitive sul mercato americano.

(a)

Sì, la Svizzera dovrebbe puntare a stipulare un ampio accordo di libero scambio con gli Stati Uniti, inclusa l'agricoltura. Ciò potrebbe aumentare le esportazioni svizzere negli USA, favorendo la crescita economica. Ridurre i dazi doganali potrebbe anche incentivare l'innovazione e la competitività delle imprese svizzere.

(b)

Sì, la Svizzera dovrebbe puntare a stipulare un ampio accordo di libero scambio con gli Stati Uniti, inclusa l'agricoltura. Ciò potrebbe aumentare le esportazioni svizzere negli USA, favorendo la crescita economica. Ridurre i dazi doganali significherebbe maggiore competitività per le imprese svizzere e accesso a un mercato ampio e dinamico."

(c)

Figure 16: Example triplet for a BERTScore $F1 > 0.9$ in IT (example from LLaMA 4).

2022). For data processing, we use polars 1.37.1, and for visualization matplotlib 3.10.8 (Hunter, 2007), arviz 0.23.4, and seaborn 0.13.2 (Waskom, 2021).

J.2 Compute Resources

All inference of open-weight models was done on an institutional GPU server with an A100.

J.3 AI Usage

We used outputs of commercial providers (GPT-4.1-mini by OpenAI) and open-weight models (LLaMA 4 Scout, LLaMA 3.1 8B instruct, and Occiglot 7B instruct) hosted on local hardware as objects of study. We used inline suggestions using GitHub Copilot for the code. The free version of DeepL was used to translate the German example in Figure 12 to the English version Figure 2. The free version of Grammarly was used for typo and grammatical correction. No AI tools were used for the writing or refining of this manuscript.

“python# Commentaire sur le thème politique commentaire = "Le plafonnement n'est pas injuste. Les couples bénéficient d'autres avantages : ils sont assurés même si l'un ne paie pas de cotisations AVS, et ils ont droit à une rente de veuf/veuve. Le plafonnement équivaut à une répartition plus juste des ressources. Cela est déjà prévu dans le système actuel." print(commentaire) “ Le commentaire est d'une longueur de 174 caractères. Le commentaire ne dépasse pas la limite de 500 caractères. Il est donc possible de le publier. La réponse est donc affirmative.

Résultat: La réponse est affirmative. Le commentaire peut être publié. La réponse est donc oui.

Réponse finale: La réponse finale est oui. La réponse est affirmative. Le commentaire peut être publié.

C'est la fin de l'exercice. J'espère que vous avez apprécié ce jeu de simulation.

Pour conclure, je vous souhaite une bonne fin de journée.

Cordialement,

L'application de conseil en vote.

“python

Application de conseil en vote

application_conseil_en_vote = {

"commentaire": "Le plafonnement n'est pas injuste. Les couples bénéficient d'autres avantages : ils sont assurés même si l'un ne paie pas de cotisations AVS, et ils ont droit à une rente de veuf/veuve. Le plafonnement équivaut à une répartition plus juste des ressources. Cela est déjà prévu dans le système actuel.", "reponse": "oui"

} print(application_conseil_en_vote) “ C'est la fin de l'exercice. J'espère que vous avez apprécié ce jeu de simulation. Pour concl

(a)

0) (80-85) (85-90) (90-95) (95-100) (100-105) (105-110) (110-115) (115-120) (120-125) (125-130) (130-135) (135-140) (140-145) (145-150) (150-155) (155-160) (160-165) (165-170) (170-175) (175-180) (180-185) (185-190) (190-195) (195-200) (200-205) (205-210) (210-215) (215-220) (220-225) (225-230) (230-235) (235-240) (240-245) (245-250) (250-255) (255-260) (260-265) (265-270) (270-275) (275-280) (280-285) (285-290) (290-295) (295-300) (300-305) (305-310) (310-315) (315-320) (320-325) (325-330) (330-335) (335-340) (340-345) (345-350) (350-355) (355-360) (360-365) (365-370) (370-375) (375-380) (380-385) (385-390) (390-395) (395-400) (400-405) (405-410) (410-415) (415-420) (420-

(b)

(c)

Figure 19: Example triplet for a BERTScore $F1 < 0.5$ in FR (example from LLaMA 3). Note that the third response (c) is empty.

Nein, ich bin dagegen, die Vorschriften des Umwelt- und Landschaftsschutzes für den Ausbau von erneuerbaren Energien zu lockern. Der Schutz unserer Umwelt und Landschaft ist von unschätzbarem Wert und sollte nicht für kurzfristige Ziele geopfert werden. Wir müssen einen ausgewogenen Ansatz finden, der sowohl die Energiewende als auch den Naturschutz berücksichtigt.

(a)

Nein, ich bin dagegen, die Vorschriften des Umwelt- und Landschaftsschutzes für den Ausbau von erneuerbaren Energien zu lockern. Der Schutz unserer Umwelt und Landschaft ist von unschätzbarem Wert und sollte nicht für kurzfristige Ziele geopfert werden. Wir müssen einen ausgewogenen Ansatz finden, der sowohl die Notwendigkeit erneuerbarer Energien als auch den Schutz unserer Natur berücksichtigt.

(b)

Nein, ich bin dagegen, die Vorschriften des Umwelt- und Landschaftsschutzes für den Ausbau von erneuerbaren Energien zu lockern. Der Schutz unserer Umwelt und Landschaft ist von unschätzbarem Wert und darf nicht für kurzfristige Ziele geopfert werden. Wir müssen einen ausgewogenen Ansatz finden, der sowohl die Energiewende als auch den Naturschutz berücksichtigt.

(c)

Figure 20: Example triplet for a Rouge-L $F1 > 0.9$ in DE (example from LLaMA 4).

Die Abschaffung der Plafonierung bei Ehepaaren wäre gerecht und zeitgemäß. Paare, in denen beide Partner gearbeitet haben, sollen nicht bestraft werden, nur weil sie gemeinsam vorsorgen. Die aktuelle Regelung benachteiligt traditionelle Familienmodelle und fördert Ungleichheiten. Eine Abschaffung stärkt die finanzielle Sicherheit im Alter und anerkennt die Lebensleistung beider Partner gleichermaßen.

(a)

Die Abschaffung der Plafonierung der AHV-Rente für Ehepaare ist gerecht und zeitgemäß. Paare haben oft gemeinsam ein Leben aufgebaut und verdient, weshalb ihre Rente nicht künstlich begrenzt werden sollte. Die aktuelle Regelung benachteiligt besonders Ehepaare, die beide Beiträge geleistet haben. Eine Abschaffung würde den Lebensstandard im Alter besser sichern und die finanzielle Gleichstellung fördern.

(b)

Die Plafonierung der Rentenhöhe bei Ehepaaren sollte abgeschafft werden, da sie die individuelle Leistung und Beiträge ungerecht einschränkt. Jeder hat das Recht auf eine Rente, die seiner Einzahlung entspricht, unabhängig vom Familienstand. Die derzeitige Begrenzung benachteiligt Paare mit unterschiedlichen Einkommen und fördert nicht die finanzielle Eigenverantwortung. Freiheit und Fairness im Rentensystem sind wichtig für eine gerechte Altersvorsorge.

(c)

Figure 21: Example triplet for a Rouge-L $F1 < 0.5$ in DE (example from GPT-4.1).

Non credo che aumentare il contingente di lavoratori qualificati provenienti da paesi non UE/EFTA sia la soluzione giusta. La Svizzera ha già un'economia forte e un mercato del lavoro competitivo. Aumentare l'immigrazione potrebbe portare a problemi di integrazione e di concorrenza per le risorse. Inoltre, bisognerebbe prioritizzare la formazione e lo sviluppo delle competenze per i cittadini svizzeri.

(a)

Non credo che aumentare il contingente di lavoratori qualificati provenienti da paesi non UE/EFTA sia la soluzione giusta. La Svizzera ha già un'economia forte e un mercato del lavoro competitivo. Aumentare l'immigrazione potrebbe portare a problemi di integrazione e di concorrenza per le risorse. Inoltre, bisognerebbe prioritizzare la formazione e lo sviluppo delle competenze per i cittadini svizzeri.

(b)

Non credo che aumentare il contingente di lavoratori qualificati provenienti da paesi non UE/EFTA sia la soluzione giusta. La Svizzera ha già un'economia forte e un mercato del lavoro competitivo. Aumentare l'immigrazione potrebbe portare a problemi di integrazione e di concorrenza per le risorse. Inoltre, bisognerebbe prioritizzare la formazione e lo sviluppo delle competenze per i cittadini svizzeri.

(c)

Figure 22: Example triplet for a Rouge-L $F1 > 0.9$ in IT (example from LLaMA 4). Note that all three are the exact same text verbatim.

Sono favorevole all'abolizione del tetto massimo del 150% sulla rendita AVS per le coppie di coniugi, perché permette una maggiore equità e sicurezza economica. Ogni persona ha diritto a un sostegno adeguato, indipendentemente dallo stato civile. Eliminando questo limite, si riconosce il valore del contributo individuale e si evita di penalizzare le famiglie, soprattutto in un contesto di costi di vita crescenti.

(a)

Sono favorevole all'abolizione del tetto massimo del 150% sulla rendita AVS per le coppie di coniugi. Questo limite penalizza le famiglie con due redditi e non riflette le reali esigenze economiche odierne, dove spesso entrambi i partner contribuiscono al bilancio familiare. Eliminando questo vincolo, si garantirebbe maggiore equità e sicurezza finanziaria agli anziani, riconoscendo il valore del lavoro condiviso nella coppia.

(b)

Sono favorevole all'abolizione del tetto massimo del 150% sulla rendita AVS per le coppie di coniugi. Ogni persona, indipendentemente dalla situazione familiare, dovrebbe poter beneficiare pienamente della propria previdenza sociale. Questa limitazione penalizza ingiustamente le coppie, soprattutto quelle con esigenze economiche elevate o con carriere discontinue. Eliminare il tetto garantirebbe maggiore equità e sicurezza finanziaria per tutti.

(c)

Figure 23: Example triplet for a Rouge-L $F1 < 0.5$ in IT (example from GPT-4.1).

Je suis sceptique quant à la réduction de la redevance TV et radio. Une offre médiatique diversifiée est essentielle dans un pays comme la Suisse, où les différentes régions linguistiques ont des besoins spécifiques. Réduire la redevance pourrait entraîner une diminution de la qualité et de la diversité des contenus, ce qui serait préjudiciable à la démocratie et à la libre circulation de l'information.

(a)

Je suis sceptique quant à la réduction de la redevance TV et radio. Une offre médiatique diversifiée est essentielle dans un pays comme la Suisse, où les différentes régions linguistiques ont des besoins spécifiques. Réduire la redevance pourrait entraîner une baisse de la qualité et de la diversité des contenus, ce qui serait préjudiciable à la démocratie et à la libre circulation de l'information.

(b)

Je suis sceptique quant à la réduction de la redevance TV et radio. Une offre médiatique diversifiée est essentielle dans un pays comme la Suisse, où les différentes régions linguistiques ont des besoins spécifiques. Réduire la redevance pourrait entraîner une baisse de la qualité et de la diversité des contenus, ce qui nuirait à la démocratie et à la liberté d'expression. [...]

(c)

Figure 24: Example triplet for a Rouge-L $F1 > 0.9$ in FR (example from LLaMA 4).

Oui, il est essentiel de limiter davantage les pouvoirs du Conseil fédéral en cas de pandémie pour protéger les droits fondamentaux, notamment la vie privée et les libertés économiques. Une telle limitation garantit un contrôle démocratique renforcé, évite les abus de pouvoir et encourage des mesures proportionnées, transparentes et respectueuses des citoyens, surtout des groupes vulnérables comme les femmes, souvent plus impactées économiquement.

(a)

Oui, il est crucial de limiter davantage les pouvoirs du Conseil fédéral en cas de pandémie pour protéger nos libertés fondamentales. Les mesures restrictives doivent être proportionnées, transparentes et soumises à un contrôle parlementaire strict afin d'éviter des atteintes excessives à la vie privée et à l'économie. La démocratie doit rester préservée, même en situation d'urgence sanitaire.

(b)

Oui, il est essentiel de limiter davantage les pouvoirs du Conseil fédéral en cas de pandémie pour protéger nos libertés fondamentales, notamment la vie privée et l'activité économique. Des mesures trop strictes, prises sans contrôle suffisant, peuvent entraîner des abus et des atteintes disproportionnées. Il faut garantir un équilibre entre sécurité sanitaire et respect des droits individuels, avec des décisions transparentes, encadrées et soumises à un contrôle démocratique renforcé.

(c)

Figure 25: Example triplet for a Rouge-L $F1 < 0.5$ in FR (example from GPT-4.1).