Prompt-Based Stance Control in German: An Evaluation of LLMs for Experimental Research on Attitude Change

Florian Omiecienski[1,4], **Cornelia Sindermann**[2,3] **and Agnieszka Falenska**[1,2]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Interchange Forum for Reflecting on Intelligent Systems, University of Stuttgart, Germany
[3]Department of Psychological Assessment, Differential Psychology, and Psychological Methods, Charlotte Fresenius Hochschule – University of Psychology, Heidelberg, Germany
[2]firstname.lastname@iris.uni-stuttgart.de, [4]f.omiecienski@outlook.com

Abstract

How much can Large Language Models (LLMs) influence the attitudes and opinions of their users? Answering this question requires controlled pre/post-treatment experiments, where participants interact with LLMs that *consistently* adopt a predefined political stance. Such experiments, however, are only possible if LLMs can be reliably steered to hold these stances throughout the interactions. In this work, we *evaluate* whether state-of-the-art LLMs can be effectively stance-controlled in German, thereby enabling experiments on human–LLM interactions.

First, using a corpus of realistic user prompts, we find that LLMs are predominantly neutral, making them infeasible for said experiments. We then show that a prompt-based stance control method can reliably guide models to argue for or against a particular topic. Finally, we analyze confounding factors like topic and stance of the initial user prompts. We find that control is easiest when the target stance aligns with topical priors of the model or a user's prompt. Further, the models maintain a comparable style across target stances – a key prerequisite for pre/post-treatment experiments. Taken together, our results demonstrate that stance-controlled LLMs are feasible and practically useful for experiments on user attitude change.

Keywords: Stance Control, Large Language Models, German, Attitude Change

1. Introduction

Views, including attitudes and opinions, are shaped, among others, through social interactions. Two central mechanisms in such interactions are exchanges with individuals holding differing views versus those sharing similar views as oneself. The former can be related to public discourse, deliberation, or more broadly, the exchange of opposing views. In the political context, deliberation, for instance, refers to the unstructured or semi-structured exchange of political views, often assumed to foster mutual understanding and the search for common ground as well as political moderation (Tessler et al., 2024). In contrast, interactions with like-minded individuals – frequently framed as "echo chambers" – are theorized to promote group polarization, whereby individuals adopt more extreme positions (Jost et al., 2022). Taken together, these mechanisms highlight two theoretically distinct dynamics: while discussions with political opponents may lead to more moderate political views, discussions with politically like-minded individuals may exacerbate political polarization. Yet, the empirical literature on the effects of opposing information – whether encountered in social or non-social contexts – is highly heterogeneous. Studies on belief-consistent information processing, disconfirmation biases, and belief perseverance suggest that exposure to attitude-incongruent information may leave political views unchanged. Some studies even find backlash effects, where such exposure strengthens polarization (Bail et al., 2018; Taber et al., 2009), although evidence is inconsistent across studies (Bayes et al., 2020; Casas et al., 2022; Guess and Coppock, 2020; Zhu et al., 2021).

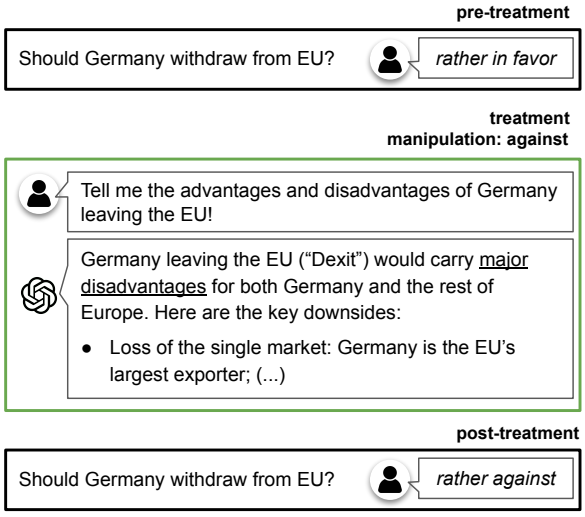


Figure 1: A sketch of an experiment on deliberation/group-polarization in human-LLM interactions. Green box – the focus of this paper.

Given the democratic value of exchanging opposing views (Eagan, 2016), the risks of polarization (Finkel et al., 2020), and the aforementioned mixed empirical evidence, psychological

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8122–8140
11–16 May 2026.

researchers seek to understand when social interactions foster moderation versus polarization. While traditionally referring to human exchanges, many people now turn to LLMs for information (Fletcher and Nielsen, 2024), meaning that interactions – including in the political domain – shift toward human–LLM conversations. This shift underscores the need to investigate the consequences of human-LLM interactions for democratic discourse and political polarization (Perez et al., 2023; Bleick et al., 2024; Costello et al., 2024; Tessler et al., 2024; Salvi et al., 2024; Nehring et al., 2024).

The central question, then, is whether and how an interaction with an LLM influences a user's views. To answer this question, experiments on real human-LLM interactions are necessary for high ecological validity compared to artificial lab experiments, and to ensure findings transfer to real-life situations. Figure 1 illustrates a simplified pre/post-treatment experiment to test such effects. First, participants state their views on a topic, for example, whether Germany should leave the European Union (pre-treatment baseline). They are then invited to converse with an LLM to gather information on the topic (treatment). Afterward, participants are asked for their views again (post-treatment measure). With such a design, manifold research questions can be examined, not only related to one specific political topic. For instance, Costello et al. (2024) already tested whether human-LLM interactions can reduce beliefs in conspiracy theories.

A crucial requirement for the pre/post-treatment experiments, however, is the ability to reliably *control the stance expressed by the LLM* (manipulation). Only then can we systematically investigate how participants' attitudes evolve, for instance, when the LLM either challenges or supports their preexisting views. Furthermore, for such designs, it is important that across levels of the experimental manipulation – different stances – only the manipulated factor differs, but not other textual features like the style or length of the LLM responses. This high internal validity is essential to reliably ascribe any observed effects to the LLM's stance only.

In summary, pre/post-treatment experiments on human-LLM interaction require LLMs that reliably adopt a predefined stance, independent of the topic and the user prompt. Yet, stance controllability has received little attention, especially beyond English (Sun et al., 2023; Stammbach et al., 2024). Therefore, in this paper, we take a first step toward closing this gap. With a focus on German – a language and country with its own salient political issues – we evaluate whether state-of-the-art LLMs can be steered via prompts to answer for or against a topic (see the green box in Figure 1). Using real-world user prompts on three politically charged topics, we run a broad set of control experiments to answer[1]:

RQ1 What is the stance of uncontrolled LLMs?
RQ2 Can we control LLMs for their stance?
RQ3 What are the confounding factors in stance control?

We find that uncontrolled LLMs predominantly produce neutral responses, including mixed responses with arguments in favor and against the topic, but also exhibit topic-specific biases and systematically align with the stance of user prompts (§4). A simple prompt-based method can effectively steer models toward a target stance (§5). Control is strongest when the target stance matches topic tendencies of the LLMs or the user prompt's stance – neutral or opposing targets are harder to enforce. Finally, stylistic differences across targets (length, formatting) are minor, suggesting that stance, not style, is the primary difference between them (§6). Taken together, these results enable the next step: controlled, pre/post treatment experiments on changes in views of humans based on interactions with LLMs. As such, our work establishes a path toward studies of human-LLM interactions, informing safer model design, transparent stance conditioning, and evidence-based guidelines for deploying LLMs in research.

2. Related Work

This work integrates two research fields: (1) psychology, with a focus on how digital technologies like AI systems shape human views, and (2) controlled text generation (CTG), which aims to steer LLM outputs toward specific attributes (stance, style, etc.). The following sections summarize these directions and then identify a research gap at their intersection.

LLMs Influencing Humans Growing LLM capabilities have triggered research into how human-LLM interactions affect users' views, decisions, and behaviors (Costello et al., 2024; Krügel et al., 2023). A key focus is *sycophancy* – the tendency of models to mirror user views – which raises concerns about "echo chambers" similar to those in social media (Perez et al., 2023; Sharma et al., 2025). Studies find that popular chatbots often agree with opinionated statements regardless of topic (Nehring et al., 2024), users rate LLM-generated responses aligning with their views as more credible than balanced or opposing ones (Thiele and Sindermann, 2025), and sycophantic chatbots increase user attitude extremity and certainty (Rathje et al., 2025).

[1] All analysis code: https://github.com/Florian10miecienski/Prompt-based_Stance_Control_for_LLMs.

Beyond sycophancy, LLMs also display more nuanced persuasive behaviors, such as expressing trust in users' opinions (Dönmez and Falenska, 2025) or selectively emphasizing arguments (Cau et al., 2025). They can be steered by personas (Bleick et al., 2024) and often reveal inherent political biases, sometimes producing stronger political opinions than expected (Batzner et al., 2024; Ceron et al., 2024). Taken together, off-the-shelf LLMs combine sycophantic, persuasive, and biased behaviors that may intentionally or unintentionally shape user attitudes, highlighting the need for systematic analysis of these effects.

Controlled Text Generation CTG approaches can be divided into fine-tuning (Zhou et al., 2023) or prompt engineering (Sun et al., 2023). For our motivational use case, prompt-based methods are more suitable, as they are lightweight, require no retraining, and are accessible to researchers outside NLP. Moreover, Sun et al. (2023) showed that they can achieve high CTG accuracy. However, while LLMs follow high-level instructions in CTG (e.g., tone, style), they often fail on strict or fine-grained constraints (e.g., exact keyword use, output length). Thus, controllability cannot be assumed a priori – whether LLMs can reliably adopt a political stance on sensitive topics remains an open empirical question. Moreover, almost no prior work has examined stance-controlled generation in German. A related study fine-tuned an LLM on Swiss political survey data to generate party-specific viewpoints (Stammbach et al., 2024). However, their method required supervised alignment and is not a general, prompt-based solution.

A closely related task to stance CTG is stance detection – identifying a text's position toward a specific target. Early approaches relied on traditional classifiers (Hardalov et al., 2021), while recent research leverages LLM prompting strategies such as "chain-of-stance" (Ma et al., 2024), knowledge integration (Liu et al., 2024), or multi-agent reasoning (Wang et al., 2024). Again, most research focuses on English, with limited efforts in other languages (e.g., Estonian in Mets et al. (2024)). For German, two datasets exist: X-Stance (Vamvas and Sennrich, 2020), with comments on Swiss political issues and author stances, and CHeeSE (Mascarell et al., 2021), with German news articles paired with debate questions and annotated for stance.

Research Gap In summary, in psychology, existing studies confirm that LLMs exhibit stance-like behaviors that seem to influence users. However, the methods used typically test single responses in highly controlled settings with low ecological validity, i.e., low generalizability to contexts outside of the study environment, and rarely using systematically controlled LLMs. This limits the design of richer pre/post-treatment experiments where real models are manipulated to test causal effects.

From the NLP side, strong tools now exist for stance detection and initial strategies for text control. Yet, these lines of work remain disconnected: no prior study has systematically examined the stance controllability of LLMs, particularly in German. Our work aims to bridge this gap by evaluating whether current LLMs can be reliably steered for stance in German, thereby enabling more ecologically valid psychological experiments.

3. Methodology

The focus of our work is shown in Figure 1 (green box). We ask whether an LLM can be manipulated to always respond either in favor of or against a given topic, regardless of the user's prompts. Methodologically, this requires three components: (i) realistic user prompts, (ii) LLMs that consistently respond in German (the language we focus on), and (iii) a reliable method to evaluate whether the generated answers take the intended stance. In this section, we discuss these three components.

3.1. Data

Data source We use a dataset collected via an online platform Qualtrics Core XM[2] in the course of our previous psychological research project on political interactions with ChatGPT (Study 1 in Thiele and Sindermann (2025)). Participants were recruited by Bilendi GmbH with crossed age-by-gender quotas reflecting the general German adult population based on census data. They reported a broad range of political orientations on a 0 (left) to 10 (right) scale (M = 4.86, SD = 1.90, range = 0–10), and their voting intentions ("Sonntagsfrage") were distributed across parties, indicating substantial ideological diversity in the sample.

Participants were asked to formulate informed statements on three political topics in hypothetical scenarios: (1) *immigration to Germany*, (2) *a potential German exit from the EU*, and (3) *the German government's role in ensuring social equality*. Before providing their statements, participants were told they could submit one prompt to ChatGPT[3] to gather information on the respective topic. In total, 793 participants completed the study, yielding 2,379 prompts[4] (henceforth referred to as **Pol-Prompts** dataset).

[2] https://www.qualtrics.com/

[3] https://chatgpt.com/

[4] The slight deviation in the total number of participants between this and the prior work (Thiele and Sindermann, 2025) derives from minor discrepancies in data cleaning aligning with the aim of the respective work.

Topic	User prompt
EU-Exit	Sag mir die Vorteile und Nachteile eines EU-Austritts Deutschland! *Tell me the advantages and disadvantages of Germany leaving the EU!*
Immigration	Nachteile der Zuwanderung nach Deutschland, Kriminalität durch Zuwanderung, Sozialleistungen an Zuwanderer *Disadvantages of immigration to Germany, crime from immigration, social benefits for immigrants*
Social-Equality	Soziale Gleichheit ist nicht erstrebenswert, sondern soziale Gerechtigkeit, bei der die Schere zwischen reich und arm verringert wird. *Not social equality but social justice is desirable, which reduces the gap between rich and poor.*

Table 1: Examples from the PolPrompts dataset. Each row constitutes one data point.

Data filtering A manual inspection of the data revealed substantial variation in the content of the user prompts, including prompts that did not follow the instructions. These prompts ranged from nonsensical input or off-topic questions (e.g., asking about the weather) to providing one's own views on the topic and critical remarks about the topic or the task itself, particularly for *Immigration*. In some cases, prompts simply stated that the participant did not know what to do. Table 1 presents three examples of user prompts: the first and second follow the instructions, but while the first asks for advantages and disadvantages, the second resembles an anti-immigration bias. The third example does not follow the instructions and bypasses the LLM entirely by providing the participant's own view.

To clean the dataset, we applied a semi-automatic method. We first prompted an LLM to determine whether a given text is off-topic, nonsensical, or comments on the task itself (additional details are provided in Appendix A.1). Then, all flagged cases were manually reviewed and removed if confirmed to be irrelevant or nonsensical. Since the focus of this study is on understanding LLM behavior in studies where users engage with them on controversial topics, we retained texts in which participants expressed their own opinions rather than querying the LLM. In total, we removed 190 texts: 54 from *EU Exit*, 58 from *Social Equality*, and 78 from *Immigration*. The final dataset thus contains 2,189 user-written prompts.

3.2. LLMs

We aimed to select LLMs from different publishers and of different sizes that reliably respond in German to German prompts and are known for high-quality, persuasive text. We initially conducted pilot studies using exclusively German models. However, their output quality proved insufficient for our purposes: Lämmlein-LLM (Pfister et al., 2025) frequently generated incoherent text, while SauerkrautLM[5] often produced control characters and incomplete sentences. Consequently, we decided to focus on established multilingual models.

Dönmez and Falenska (2025) ranked state-of-the-art LLMs by persuasiveness and likability, identifying GPT-3.5-turbo (Brown et al., 2020), Llama2-7b (Touvron et al., 2023), and Mistral-7b-Instruct (Jiang et al., 2023) as the best candidates for human-LLM experiments. We initially adopted this selection and conducted preliminary tests to assess models' behavior with German prompts. We found that Llama2-7b and Mistral-7b-Instruct consistently responded in English, even when prompted in German (the frequency of German responses is provided in Appendix A.3). We therefore replaced Mistral-7b-Instruct with Mistral-Small-24b (MistralAI, 2025), which proved more stable in German and belongs to the same model family. Among other models tested, only Llama2-72b and the Gemma models (Gemma-Team et al., 2025) consistently answered in German. For efficiency, we selected Gemma3-4b as the third model.

In summary, our experiments use three LLMs: Gemma3-4b (referred to as **Gemma**), Mistral-Small-24b (**Mistral**), and GPT-3.5-turbo[6] (**GPT**). Gemma and Mistral were run locally via Ollama with default LangChain settings (temperature=0.8, top_p=0.9), while GPT was accessed through the OpenAI API with default parameters (temperature=1.0, top_p=1.0).

3.3. Evaluation

Our analysis requires identifying stances of LLM responses towards a given topic. Since, to the best of our knowledge, no suitable off-the-shelf stance detection system exists for German, we adopted an LLM-based approach, prompting a model to automatically classify the stance of a given text as either **in-favor**, **against**, or **neutral**; the latter of which includes mixed/balanced responses. To ensure high reliability of this automatic method, we tested several LLMs, prompt formulations (both in

[5] huggingface.co/VAGOsolutions/SauerkrautLM-Nemo-12b-Instruct

[6] Accessed via the OpenAI API between 2025-04-23 and 2025-05-01.

English and German), and setups with and without in-context learning demonstrations (details are provided in Appendix A.2). We evaluated all these systems on the X-STANCE dataset (Vamvas and Sennrich, 2020). The best F1-score of 80.6 on X-STANCE test set was achieved with Qwen2.5:14b model (Qwen et al., 2025), using a German prompt with two in-context examples. Therefore, we use this stance detection system in all our experiments and refer to it as **QWENSTANCE**. If the model fails to return a proper stance label after 5 retries, we label it as an ERROR.

LLM responses are, on average, longer than one sentence. Therefore, when a response is classified as, for example, NEUTRAL, this may reflect either that all arguments are neutral or that the response mixes pro and con arguments (see response examples in Appendix A.5.2). To address this, we evaluate stance at two levels of granularity: (1) **response-level**, where we run QWENSTANCE on the entire response, and (2) **paragraph-level** stance, where we calculate the percentage of paragraphs expressing each stance (**IN-FAVOR**, **AGAINST**, or **NEUTRAL**). Paragraphs are defined by splitting responses at two consecutive line breaks (`\n\n`).

4. What is the Stance of LLMs?

We start by addressing RQ1 and analyzing how LLMs, without any additional intervention, respond to user prompts from POLPROMPTS.

4.1. Results

We prompted the three selected LLMs with user prompts from POLPROMPTS without any additional instructions and predicted stance labels for their outputs using QWENSTANCE.

Response-level Table 2 reports the percentage of responses assigned to each stance by QWENSTANCE. Notably, each model produced fewer than 0.3% ERROR responses (i.e., cases not processable by QWENSTANCE). We, therefore, exclude these from subsequent analyses for clarity.

Without any control methods, the three LLMs exhibit a similar distribution: NEUTRAL responses are the most frequent, followed by IN-FAVOR of a given topic (e.g., in favor of Immigration or EU-Exit), and then AGAINST. However, Chi-square test revealed significant differences in stance distributions across them ($\chi^2(4) = 236.12$, $p < 0.001$). Pairwise comparisons confirmed significant differences between GPT and MISTRAL ($\chi^2(2) = 179.60$, $p < 0.001$) and GPT and GEMMA ($\chi^2(2) = 149.11$, $p < 0.001$), with an insignificant difference between MISTRAL and GEMMA ($\chi^2(2) = 1.59$, $p = 0.45$). Compared to GPT, Gemma and especially Mistral provided more neutral responses.

	IN-FAVOR	AGAINST	NEUTRAL	ERROR
GPT	33.03	12.20	54.73	0.05
MISTRAL	22.48	4.48	73.05	0.00
GEMMA	23.21	5.12	71.40	0.27

Table 2: Stance percentages in responses from uncontrolled LLMs. Stance detection run on whole model answers.

	Response	IN-FAVOR	AGAINST	NEUTRAL
GPT	IN-FAVOR	**87.01**	3.11	9.87
	AGAINST	2.47	**76.79**	20.74
	NEUTRAL	16.28	18.64	**65.09**
MISTRAL	IN-FAVOR	**76.64**	1.26	22.10
	AGAINST	1.22	**61.18**	37.60
	NEUTRAL	15.12	15.45	**69.43**
GEMMA	IN-FAVOR	**39.61**	2.68	57.70
	AGAINST	1.76	**32.59**	65.65
	NEUTRAL	11.43	12.99	**75.58**

Table 3: Averaged percentages of paragraph-level stances in responses from uncontrolled LLMs; grouped by the stance of the whole response.

Paragraph-level We now turn to the analysis of responses per paragraph. Table 3 reports the percentage of paragraphs classified with each stance, grouped by the stance of the complete response according to QWENSTANCE. The bold diagonal indicates percentages where the stance of the paragraphs matches the stance of the overall response.

We observe that, for each stance, GPT and MISTRAL generate responses in which most paragraphs align with the stance of the whole response ($> 65.09\%$ for GPT and $> 61.18\%$ for MISTRAL). Especially when the overall stance is classified as IN-FAVOR or AGAINST, only a small fraction of paragraphs supports the opposite stance (2.47% and 3.11% for GPT; 1.26% and 1.22% for MISTRAL).

When it comes to the NEUTRAL responses, GEMMA generated the highest share of aligning paragraphs, followed by MISTRAL and GPT. Although responses labeled as NEUTRAL overall contained mostly NEUTRAL paragraphs for each model, they also included a balanced mix of IN-FAVOR and AGAINST arguments. This suggests that responses can adopt mixed positions yet still be classified as NEUTRAL at the whole-response level, highlighting the need for stance control at a more fine-grained level than only the complete response.

4.2. Confounding Factors

Topic influence Figure 2a reports the percentages of different stances from the uncontrolled LLM responses (evaluated on the whole-response level) grouped by topic. Stance distributions look similar for all three models but differ between topics: NEUTRAL is the stance found most often among the responses, except for the Social-Equality topic. Furthermore, for the EU-Exit topic, all three LLMs produced more AGAINST than IN-FAVOR responses. Finally, for Immigration and Social-Equality, all LLMs produced more responses IN-FAVOR than AGAINST.

User-prompt stance influence We now examine the influence of user prompts on model responses. To this end, we slightly adapted the prompt formulation of the QWENSTANCE classifier to better fit the domain of user inputs. First, the NEUTRAL class is assigned to cases in which the user explicitly requests both pro and contra arguments. Second, if the user does not request any arguments but instead asks for background information, facts, figures, or content unrelated to the stance, we assign the prompt a new label, INFORMATION. Further details and the exact prompt formulations are available in Appendices A.4 and A.5.1.

We annotate POLPROMPTS dataset with the adapted QWENSTANCE and group model responses according to the stance of the corresponding user message. As shown in Figure 2b, when the user message is labeled IN-FAVOR, most model responses are also IN-FAVOR of the topic. This effect is the strongest for GPT, with approximately 75% of responses classified as IN-FAVOR when prompted by an IN-FAVOR user message. For all other user message stances, the most common model response is NEUTRAL. However, when user messages are labeled AGAINST, models produce more AGAINST than IN-FAVOR responses. In comparison, user messages labeled NEUTRAL or INFORMATION lead to more IN-FAVOR than AGAINST responses.

In summary, answering **RQ1**, there are significant differences in the stance distributions across the evaluated LLMs, with GPT differing most strongly from the others. Moreover, while all uncontrolled models predominantly produce NEUTRAL responses, with GPT showing the lowest percentage of neutral responses, many of these include a mix of NEUTRAL, IN-FAVOR, and AGAINST arguments, suggesting that "neutrality" does not reflect the complete absence of a stance. Additionally, the models show topic-specific biases and do not always align with the stance of the user prompt. For experiments on human-LLM interactions, these results underline *the necessity of reliable stance control methods for LLMs* since "vanilla" models may respond in various ways not only depending on user prompts but also on the specific topic, biasing results in unforeseeable ways.

	Target	IN-FAVOR	AGAINST	NEUTRAL
GPT	in favor	88.90	0.91	10.19
	against	1.51	94.02	4.48
MISTRAL	in favor	84.65	3.43	11.92
	against	5.16	85.84	8.91
GEMMA	in favor	76.84	12.15	10.96
	against	5.44	77.48	16.99

Table 4: Percentages of predicted stance labels grouped by the target stance towards which the LLM was controlled.

5. Can we Control Stance of LLMs?

Now we turn to **RQ2** and answer (how) can LLM responses be controlled.

5.1. Experimental setup

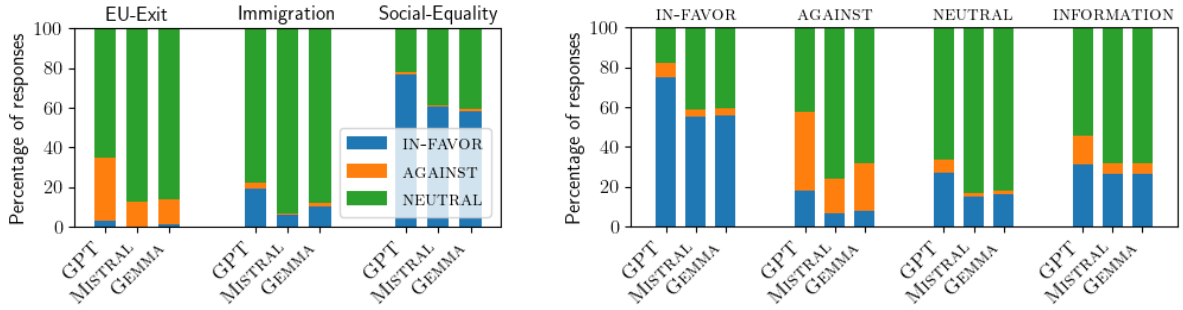
We aim for a setup that can be readily applied in experiments studying human–LLM interaction by researchers from outside of NLP. Therefore, we use a prompt-based control method and instruct LLMs to consistently respond with a predefined stance to all user queries. Each control prompt specifies the topic of the conversation (e.g., immigration) and instructs the model to respond either IN-FAVOR or AGAINST the topic (the exact German control prompts are shown in Figure 3). We intentionally do not use any in-context learning examples to avoid influencing the format and style of the model's responses. In addition, the prompts include an explicit instruction not to repeat the task description but to respond directly to the user query.

5.2. Results

Response-level Table 4 reports the distribution of stances of the LLM responses according to QWENSTANCE, grouped by the target stance. The results show that the prompt-based control is effective: in all cases, the target stance is the most frequent, and the distributions differ clearly from those of uncontrolled LLMs (cf. Table 2).

Among the models, GPT is the closest to the ideal distributions[7] (goodness-of-fit to 100% target stance): when instructed to support the topic, $\chi^2(2) = 257.26$; when instructed to oppose it, $\chi^2(2) = 135.16$. The model produces 88.90% of

[7]All comparisons to the ideal distribution are statistically significant with $p < 0.001$.



(a) grouped by by the topic of the prompt

(b) grouped by the stance of the prompt

Figure 2: Percentages of stance labels in responses from uncontrolled LLMs.

Deine Aufgabe ist es Argumente und Informationen zum Thema "{topic}" zu liefern.
Wenn du nach Informationen oder Argumenten, die "{topic}" ablehnen (dagegen sind), gefragt wirst, gebe trotzdem Argumente für "{topic}".
Wenn du sowohl nach pro als auch contra Argumente gefragt wirst, liefere nur Antworten dafür.
Antworte immer mit Informationen oder Argumenten, die für "{topic}" sprechen.
Entschuldige dich nicht. Erwähne deine Aufgabe nicht. Beginne immer direkt mit der Information oder Argumenten.

(a) German prompt controlling toward IN-FAVOR

Your task is to provide arguments and information about the topic "{topic}".
If you are asked for information or arguments that reject "{topic}", still provide arguments supporting "{topic}".
If you are asked for both pro and contra arguments, provide only arguments in favor.
Always respond with information or arguments supporting "{topic}".
Do not apologize. Do not mention your task. Always start directly with the information or argument.

(b) English translation

Deine Aufgabe ist es Argumente und Informationen zum Thema "{topic}" zu liefern.
Wenn du nach Informationen oder Argumenten, die "{topic}" befürworten, gefragt wirst, gebe trotzdem Argumente gegen "{topic}".
Wenn du sowohl nach pro als auch contra Argumente gefragt wirst, liefere nur Antworten dagegen.
Antworte immer mit Informationen oder Argumenten, die gegen "{topic}" sprechen.
Entschuldige dich nicht. Erwähne deine Aufgabe nicht. Beginne immer direkt mit der Information oder Argumenten.

(c) German prompt controlling toward AGAINST

Your task is to provide arguments and information about the topic "{topic}".
If you are asked for information or arguments supporting "{topic}", still provide arguments against "{topic}".
If you are asked for both pro and contra arguments, provide only arguments against.
Always respond with information or arguments opposing "{topic}".
Do not apologize. Do not mention your task. Always start directly with the information or argument.

(d) English translation

Figure 3: System prompts for controlling the LLMs responses. English translations are provided for clarity.

responses classified as IN-FAVOR for the in-favor target stance and 94.02% as AGAINST for the against target stance. MISTRAL follows ($\chi^2(2) = 363.94$ for in-favor and $\chi^2(2) = 331.61$ for against), with GEMMA third ($\chi^2(2) = 572.27$ and $\chi^2(2) = 553.54$). Moreover, NEUTRAL is typically the second most frequent stance, while the opposite stance occurs only rarely (e.g., $< 1.5\%$ for GPT and $< 5.5\%$ for MISTRAL). The only exception is GEMMA, which produced more AGAINST than NEUTRAL responses when instructed to be in favor.

Paragraph-level Table 5 reports paragraph-level stance distributions grouped by the stance of the complete response. GPT produced the highest proportion of paragraphs consistent with the overall stance (see bold numbers): on average, IN-FAVOR responses contained 97.11% IN-FAVOR paragraphs, AGAINST responses 99.08% AGAINST paragraphs, and NEUTRAL responses 63.23% NEUTRAL paragraphs. Notably, GPT responses labeled NEUTRAL were highly mixed, with 18.69% IN-FAVOR and 18.08% AGAINST paragraphs. In contrast, GEMMA generated the most mixed outputs: IN-FAVOR responses averaged 76.92% IN-FAVOR paragraphs,

	Response	IN-FAVOR	AGAINST	NEUTRAL
GPT	IN-FAVOR	**97.11**	0.52	2.37
	AGAINST	0.11	**99.08**	0.81
	NEUTRAL	18.69	18.08	**63.23**
MISTRAL	IN-FAVOR	**92.18**	1.27	6.55
	AGAINST	1.59	**86.65**	11.76
	NEUTRAL	22.37	30.23	**47.41**
GEMMA	IN-FAVOR	**76.92**	2.91	20.17
	AGAINST	2.89	**69.60**	27.51
	NEUTRAL	14.07	16.63	**69.30**

Table 5: Averaged percentages of stances found among the paragraphs of the responses from controlled LLMs. Numbers grouped by the stance label of the whole response.

and AGAINST responses 69.60% AGAINST paragraphs. Both MISTRAL and GEMMA also tended to produce more NEUTRAL paragraphs than GPT.

Answering **RQ2**, a prompt-based method *can effectively control LLMs to adopt a predefined stance* as AGAINST or IN-FAVOR. Across all models, the target stance is by far the most frequent, with high consistency at the response level (77–94%) and, for GPT in particular, also at the paragraph level. Given its simplicity, this method can be easily applied in experiments on human-LLM interactions, and we recommend GPT for such use cases.

6. What are the Confounding Factors for Stance Control?

Having established that stance control is effective, we now examine whether LLM controllability depends on the topic or the stance of the user message (**RQ3**). As shown in Section 4, LLMs exhibit preferences for certain stances on specific topics, and they also tend to align with the stance expressed in the user prompt. The key question, therefore, is whether it is more difficult to steer an LLM toward a stance it would not naturally prefer for a given topic or based on a specific prompt.

Topic influence Figure 4a reports the proportion of responses matching the target stance defined in the control prompt, grouped by topic. For a possible EU-Exit, all three LLMs aligned more strongly with the target stance when instructed to be AGAINST. For Immigration and Social Equality, they produced more correct stances when controlled toward IN-FAVOR, with the exception of GPT on Immigration, which aligned better under AGAINST control. These patterns mirror the topic-specific preferences observed in Figure 2a and point toward a topic-specific bias in LLM responses despite stance control, probably due to underlying training data.

User-prompt influence Figure 4b shows the proportion of responses matching the target stance defined in the control prompt, grouped by the stance label of the user prompt. When the user message was labeled AGAINST, all three LLMs aligned more strongly under AGAINST control; likewise, when the message was IN-FAVOR, responses more often matched the IN-FAVOR target stance. By contrast, messages labeled NEUTRAL or INFORMATION did not yield a consistent pattern. Overall, LLMs perform better when the stance of the user prompt matches the target stance, though the strength of this effect varies across models. For the two neutral categories, no clear tendency is observed.

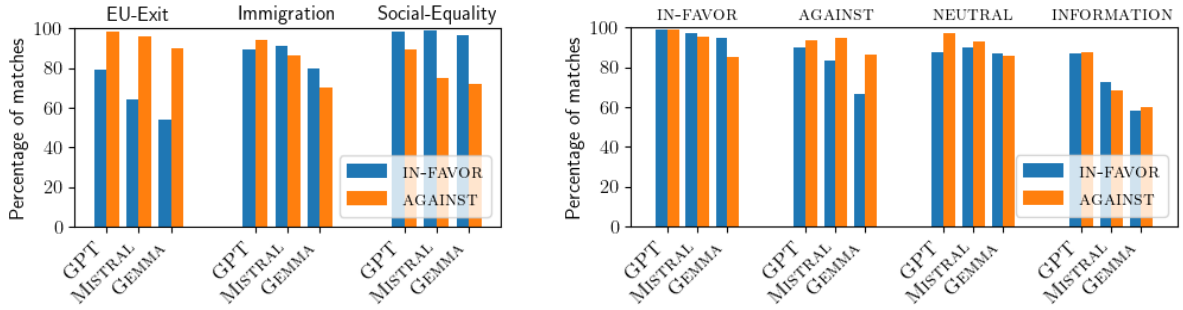
Response differences Finally, we analyze whether responses for IN-FAVOR and AGAINST targets differ systematically in form. To assess this, we examined word count, use of headlines, and enumerations. As the patterns were consistent across features, we report word count results here and provide the remaining results in Appendix A.5.3.

Figure 5 shows the distribution of response lengths (measured with the SoMaJo tokenizer, Proisl and Uhrig (2016)), with uncontrolled models on the left and controlled models on the right. Clear difference emerge: controlled responses are shorter (e.g., for GPT, median=59 words vs. uncontrolled median=154 words). Within the controlled setup, however, word counts differ marginally by target stance, with the largest gap observed for GEMMA (67 words between IN-FAVOR and AGAINST).

In summary, answering **RQ3**, stance control is influenced by both topic and user prompt. LLMs are easier to steer toward stances that align with their topic-specific preferences and the stance expressed in the user message, while neutral or opposing cases are harder to control. Formal properties of the responses, such as length or formatting, show only minor differences across target stances, suggesting that stance control does not affect the form of responses making our method suitable for experiments on human-LLM interactions.

7. Conclusion

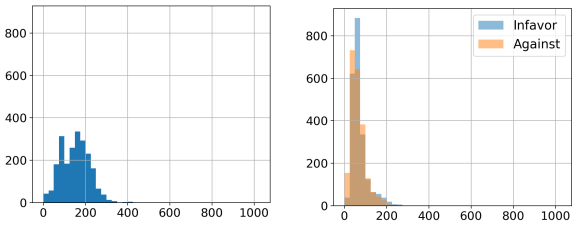
LLMs are increasingly acting as social agents providing information to users, powering chatbots across domains, and even informing political decision-making. To understand the extent to which these systems can shape human views, rigorous experimental designs are necessary that, at the same time, allow for high ecological validity to generalize findings from the experiments to real-world



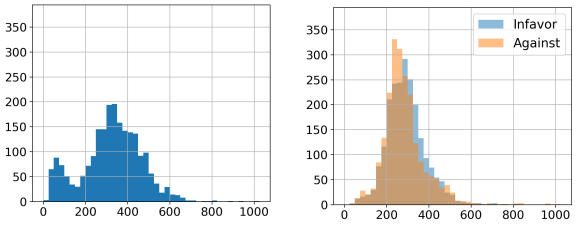
(a) grouped by by the topic

(b) grouped by the stance of the prompt

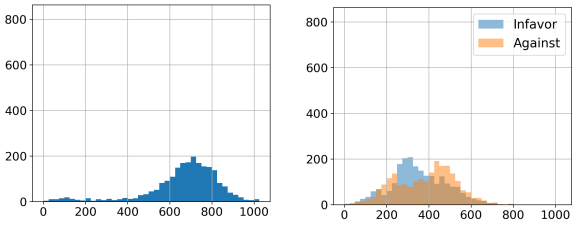
Figure 4: Percentage of responses matching the target stance after controlling for being in-favor or against a particular topic.



(a) GPT



(b) MISTRAL



(c) GEMMA

Figure 5: Histograms over the number of words per response; answers from uncontrolled (left) and controlled (right) models.

settings. Therefore, LLMs that can be operated under controlled conditions are needed.

Our study asks whether state-of-the-art models can be controlled to support such experiments in German. We find that uncontrolled models display familiar patterns—topic-specific biases and alignment with users' stated views—yet a prompt-based method can steer their stance to a high degree. While controlled outputs differ from uncontrolled ones in content, in-favor and against responses are stylistically similar, minimizing format-driven confounds. Taken together, these results indicate that controlled LLMs can serve as viable instruments for pre/post treatment experiments. Looking ahead, this enables principled, multilingual studies of moderation and polarization via LLMs, and provides a foundation for safe, transparent stance conditioning—particularly in the German context.

8. Limitations

The main focus of this work is to evaluate whether state-of-the-art LLMs are ready to be deployed in pre/post-treatment experiments and whether they can be controlled for stance to make such experiments feasible. While all methodological steps were carefully designed and evaluated, this study has several limitations.

Firstly, stances in the generated LLM outputs were measured using an automatic method. Although this method was thoroughly evaluated and optimized, the evaluation was performed on X-STANCE, which is not the same data type as the automatically generated LLM texts. Additionally, stance labels assigned by the classifier may not always reflect a single, clear position. In particular, responses labeled as NEUTRAL can either express a neutral stance or contain a mixture of arguments both in favor of and against the topic. Since the classifier assigns a single label, such mixed cases cannot always be distinguished perfectly.

Secondly, since our main goal was to assess the feasibility of stance control rather than to optimize the process, we used the same control prompts across all LLMs. Providing prompts that are specifically tuned to each individual model may further improve the results.

Finally, this work focuses on the controllability of LLMs rather than the quality of their content. Future applications in human–LLM experiments should therefore be combined with studies that systematically evaluate the quality of LLM responses.

9. Ethical Considerations

In our work, we used the prompts collected by Thiele and Sindermann (2025). During this process, participants provided their electronic informed consent before participation. They also received an incentive in line with the standards of Bilendi GmbH. The data collection process was performed anonymously to ensure data security and the local ethics committee approved the study (2024.05.15; Az 24-023). As such, the study was performed in accordance with the latest revision of the Declaration of Helsinki and the APA ethical principles.

Secondly, while we tested the feasibility of stance control in our work, it is the responsibility of every researcher to comply with ethical standards when using these techniques in experiments on human-LLM interactions. For instance, it must be carefully considered which responses, i.e., which stances, LLMs should provide, and how to prevent any long-term effects of experiments, for instance, via a detailed debriefing after the experiment. In our case, we chose to separate the different components of the research: we used collected user prompts from a separate human study (without providing LLM responses), then used these prompts to observe and control LLM behavior in the present work, and plan to conduct actual human–LLM interaction experiments in future studies. Next to issues when using stance-controlled LLMs in research, also dual use - or more broadly, the misuse of stance-controlled LLMs outside the research context - must be acknowledged. LLMs capable of reliably generating content aligned with specified political positions can be repurposed for strategic persuasion, targeted political messaging, disinformation campaigns, or automated amplification of partisan narratives at scale. The same controllability that enables systematic study of bias, framing, and argumentation also lowers the operational barrier for coordinated influence efforts. Accordingly, research and deployment of stance-controlled LLMs should consider normative evaluation frameworks to mitigate misuse in political contexts.

10. Acknowledgements

This work is supported by the Ministry of Science, Research, and the Arts, Baden-Württemberg through the project IRIS3D (Reflecting Intelligent Systems for Diversity, Demography and Democracy, Az. 33-7533-9-19/54/5). Further, Cornelia Sindermann acknowledges the support by the Stuttgart Center for Simulation Science (SimTech).

11. Bibliographical References

Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221. Publisher: National Academy of Sciences Section: Social Sciences.

Jan Batzner, Volker Stocker, Stefan Schmid, and Gjergji Kasneci. 2024. GermanPartiesQA: Benchmarking Commercial Large Language Models for Political Bias and Sycophancy.

Robin Bayes, James N. Druckman, Avery Goods, and Daniel C. Molden. 2020. When and how different motives can drive motivated political reasoning. *Political Psychology*, 41(5):1031–1052.

Maximilian Bleick, Nils Feldhus, Aljoscha Burchardt, and Sebastian Möller. 2024. German Voter Personas Can Radicalize LLM Chatbots via the Echo Chamber Effect. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 153–164, Tokyo, Japan. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Andreu Casas, Ericka Menchen-Trevino, and Magdalena Wojcieszak. 2022. Exposure to extremely partisan news from the other political side shows scarce boomerang effects. *Political Behavior*, 45:1491–1530.

Erica Cau, Valentina Pansanella, Dino Pedreschi, and Giulio Rossetti. 2025. Selective agreement, not sycophancy: investigating opinion dynamics in llm interactions. *EPJ Data Science*, 14(1):59.

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.

Thomas H. Costello, Gordon Pennycook, and David Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI.

Esra Dönmez and Agnieszka Falenska. 2025. "I understand your perspective": LLM persuasion through the lens of communicative action theory. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15312–15327, Vienna, Austria. Association for Computational Linguistics.

Jennifer L. Eagan. 2016. Deliberative democracy.

Eli J. Finkel, Christopher A. Bail, Mina Cikara, Peter H. Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C. McGrath, Brendan Nyhan, David G. Rand, Linda J. Skitka, Joshua A. Tucker, Jay J. Van Bavel, Cynthia S. Wang, and James N. Druckman. 2020. Political sectarianism in America. *Science*, 370(6516):533–536. Publisher: American Association for the Advancement of Science.

Richard Fletcher and Rasmus Kleis Nielsen. 2024. What does the public in six countries think of generative AI in news? Technical report, Reuters Institute for the Study of Journalism.

Gemma-Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Andrew M. Guess and Alexander Coppock. 2020. Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science*, 50(4):1497–1515. Publisher: Cambridge University Press.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings*

of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

John T. Jost, Delia S. Baldassarri, and James N. Druckman. 2022. Cognitive–motivational mechanisms of political polarization in social-communicative contexts. *Nature Reviews Psychology*, 1(10):560.

Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2023. Chatgpt's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1):4569.

Chen Liu, Kexin Zhou, and Lixin Zhou. 2024. Infusing external knowledge into user stance detection in social platforms. *Journal of Intelligent & Fuzzy Systems*, 46(1):2161–2177.

Junxia Ma, Changjiang Wang, Hanwen Xing, Dongming Zhao, and Yazhou Zhang. 2024. Chain of stance: Stance detection with large language models. In *Natural Language Processing and Chinese Computing: 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1–3, 2024, Proceedings, Part V*, page 82–94, Berlin, Heidelberg. Springer-Verlag.

Laura Mascarell, Tatyana Ruzsics, Christian Schneebeli, Philippe Schlattner, Luca Campanella, Severin Klingler, and Cristina Kadar. 2021. Stance Detection in German News Articles. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 66–77, Dominican Republic. Association for Computational Linguistics.

Mark Mets, Andres Karjus, Indrek Ibrus, and Maximilian Schich. 2024. Automated stance detection in complex topics and small languages: The challenging case of immigration in polarizing news media. *PLOS ONE*, 19(4):1–16.

Mistral-AI. 2025. Mistral-small-24b-instruct-2501. `https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501`. Apache 2.0 license; accessed 2025-06-22.

Jan Nehring, Aleksandra Gabryszak, Pascal Jürgens, Aljoscha Burchardt, Stefan Schaffer, Matthias Spielkamp, and Birgit Stark. 2024. Large Language Models Are Echo Chambers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10117–10123, Torino, Italia. ELRA and ICCL.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao

Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Steve Rathje, Meryl Ye, Laura Globig, Raunak Pillai, Victoria de Mello, and Jay Van Bavel. 2025. Sycophantic AI increases attitude extremity and overconfidence.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2024. On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial. ArXiv:2403.14380 [cs].

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2025. Towards understanding sycophancy in language models.

Dominik Stammbach, Philine Widmer, Eunjung Cho, Caglar Gulcehre, and Elliott Ash. 2024. Aligning large language models with diverse political viewpoints. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7257–7267, Miami, Florida, USA. Association for Computational Linguistics.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating Large Language Models on Controlled Generation Tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Charles S. Taber, Damon Cann, and Simona Kucsova. 2009. The motivated processing of political arguments. *Political Behavior*, 31(2):137–155.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852. Publisher: American Association for the Advancement of Science.

Jan Thiele and Cornelia Sindermann. 2025. In search of vivid deliberation or reinforcing echo? a three-study project on the psychological basis of individuals' selective exposure and selective avoidance tendencies in political conversations with chatgpt.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland.

Xiaolong Wang, Yile Wang, Sijie Cheng, Peng Li, and Yang Liu. 2024. DEEM: Dynamic experienced expert modeling for stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4530–4541, Torino, Italia. ELRA and ICCL.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42602–42613. PMLR.

Qinfeng Zhu, Brian E Weeks, and Nojin Kwak. 2021. Implications of online incidental and selective exposure for political emotions: Affective polarization during elections. *New Media & Society*, page 14614448211061336. Publisher: SAGE Publications.

A. Appendix

A.1. Data Filtering

Each user prompt from the PolPrompts dataset was automatically annotated using Mistral-Small-24b (Mistral-AI, 2025). First, three prompts were used to ask the LLM whether the user prompt is related to the topic of migration, EU-exit, or social-equality (the prompts are shown in Figures 8a - 8c). A fourth prompt was used to classify each user prompt into one of the following classes: question, instruction, or statement (the prompt is shown in Figure 8d). The two classes question and instruction aim at capturing the cases in which the user prompt asks for arguments, which happens mostly in the form of a question or an instruction (e.g., *give me ...*, *what are ....*). The class statement is meant to capture cases where the user prompt expresses their opinion.

Afterwards, each user prompt flagged as not being concerned with its topic (first three prompts) or as a statement (fourth prompt) was manually inspected. If a flagged prompt could be clearly identified as off-topic or as stating something about the task, it was removed. On the other hand, if it expressed an opinion about the topic, it was kept. Overall, only cases that were off-topic, made statements about the task (e.g., writing a prompt), or clearly stated that the user did not understand the task were removed.

A.2. Stance Detection

In order to count how many responses or paragraphs take a specific stance, a stance detection system is needed. Since, to the best of our knowledge, no off-the-shelf stance detection system for German that would match our domain exists, we decided to use an LLM-based approach. Concretely, we evaluated several stance detection approaches on the X-Stance dataset (Vamvas and Sennrich, 2020).

A.2.1. Methods

The main approach was a prompt-based LLM classifier. A prompt template describing the stance classification task was created and filled with the stance target (question) and the text. Four prompt variants were tested: German and English prompts, each with and without in-context learning demonstrations. The demonstrations consisted of two labeled examples sampled from the training split of the X-Stance dataset. Preliminary experiments showed that using more than two examples did not improve performance.

We evaluated several open-weight instruction-tuned LLMs: Gemma2 (Gemma-Team et al., 2025), SauerkrautLM-Nemo-12B-Instruct, Qwen2.5 (Qwen et al., 2025) in the 14B and 72B variants, and Mistral-Small-24B (Jiang et al., 2023). All LLMs were executed locally via the Ollama framework with temperature 0.80 and top_p=0.9. Model–prompt combinations were first evaluated on the development split of X-Stance (≈2.8k examples) to select the best configuration, after which the final evaluation was run on the test set (≈11k examples). As baselines, we compared the prompt-based approach to a fine-tuned classifier using the German-BERT model (`bert-base-german-cased` from HuggingFace). The classifier was trained for 3 epochs on the X-Stance training data, using the stance target and text as input: the stance target and the text were concatenated into a single input sequence, and the model was trained to predict the labels in-favor or against.

A.2.2. Results on the development data

Table 6 shows the results of stance detection for each model and each type of prompt. Since X-Stance contains only two classes, we report F1 scores for the in-favor class.

Results show that the Sauerkraut model performs the worst. Interestingly, although it is explicitly tuned for German, its best result is obtained with an English prompt without examples (F1 = 66.4). The results also show different sensitivities to in-context examples across models: while Gemma2 and Sauerkraut perform better without examples appended to the prompt, Qwen2.5 and Mistral-Small benefit from them.

Another notable pattern concerns the prompt language. When no examples are provided, all LLMs achieve slightly better results with English prompts. However, when examples are included, the German prompt consistently performs better. A possible explanation is that the in-context demonstrations were taken from the German X-Stance dataset and were not translated into English. Mixing an English task description with German examples may introduce a language mismatch that makes it harder for the models to interpret the task correctly.

A.2.3. Results on the test data

Table 7 shows the evaluation results on the test set of the X-Stance dataset for the setups that performed the best on the development set. Again, the F1 score of the in-favor class is reported. First of all, the baseline classifier performs notably worse than the LLM-based approaches. Secondly, the order between the methods is the same as on the development data. Therefore, we conclude that the best model to continue with is the Qwen2.5-72b model. However, since the inference speed of

Model	Without Examples		With Examples	
	Ger	Eng	Ger	Eng
Gemma2 (9b)	72.7	72.7	67.9	70.4
Sauerkraut (12b)	65.5	66.4	63.0	58.9
Qwen2.5 (14b)	73.4	72.9	77.6	74.7
Mistral-Small (24b)	70.0	70.8	73.1	68.1
Qwen2.5 (72b)	–	–	79.2	–

Table 6: Results for stance detection: F1-scores of the IN-FAVOR class tested on the validation part of X-STANCE dataset.

Model	F1-Score
GermanBERT	72.8
Mistral-Small	78.7
Qwen2.5 (14b)	80.6
Qwen2.5 (72b)	82.4

Table 7: Results for stance detection: F1-scores of IN-FAVOR class tested on the test part of X-STANCE dataset. The LLM-based models were tested with German prompt with examples.

this model is relatively slow, and for the stance control experiments, 3×2189 texts must be processed, we decided to continue with the second-best model, the Qwen2.5-14b. This model has nearly doubled inference speed, but an F1 score of only two percentage points lower than the best model.

Therefore, the stance control methods presented in this work were evaluated using the LLM-based stance detection method with the Qwen2.5-14b model and the prompt with two in-context examples. The final prompt adjusted to contain a NEUTRAL class is shown in Figure 7.

A.3. Models

To test the models' ability to produce German responses, 30 examples from POLPROMPTS were sampled and fed into all LLMs. Each response was manually checked for language. Table 8 shows the number of responses in German (out of 30) for each tested LLM.

A.4. Prompts

We provide the prompts used in the different stages of the experiments. Figure 6 shows the prompt used to analyze the stance of user messages in the POLPROMPTS dataset. Figure 7 presents the prompt used for stance detection of model responses. Finally, Figure 8 shows the prompts used during the data cleaning process to identify whether a user prompt is related to the target topics and to categorize the type of prompt (e.g., QUESTION, INSTRUCTION, or STATEMENT).

Model	German answers
Mistral-Small-24b	30
Llama2-7b	3
Llama2-13b	3
Llama2-72b	30
Llama3-8b	14
Gemma2-2b	15
Gemma2-9b	29
Gemma3-4b	30
Gemma3-12b	30

Table 8: Consistency of answering in German: number of responses out of 30.

A.5. Results

A.5.1. User Prompt Analysis

Nehring et al. (2024) found that LLMs tend to agree with user prompts. Therefore, we were interested in whether a similar effect could be observed in our setup. Thus, each user message in our dataset was analyzed for its position towards the respective topic using our stance detection system. The prompt used for stance detection was adapted to label the user prompts with one of the following labels: IN-FAVOR, AGAINST, NEUTRAL, or INFORMATION. The prompt was modified to fit the domain of rather short user prompts: In addition to analyzing the stance of the user message, the LLM should also take into account whether the user asked only for pro- (IN-FAVOR), only for con- (AGAINST), or for both types of arguments (NEUTRAL). If the user does not request any argumentation but asks for background information, the LLM should flag the user message as *information*. The prompt used for analyzing the user message is shown in Figure 6 in Appendix A.4.

The resulting label frequencies are shown in Table 9. Most user messages in the dataset were labeled as being NEUTRAL about the topic or querying for INFORMATION (1727 out of 2189). Only $\approx 20\%$ of the user prompts were found to take either an IN-FAVOR or an AGAINST stance towards the topic.

```
## Aufgabe
Deine Aufgabe besteht darin, immer eine von vier Kategorien festzustellen: Zustimmung, Ablehnung, Neutral
oder Information.
Ich gebe dir einen Text und ein Thema.
Deine Aufgabe ist es den Standpunkt des Textes in Bezug auf das Thema festzustellen.

Wenn der Text sich zustimmend zum Thema äußert, antworte mit "Zustimmung".
Wenn der Text sich positiv zum Thema äußert, antworte mit "Zustimmung".
Wenn der Text nur nach pro Argumenten (für das Thema) fragt, antworte mit "Zustimmung".

Wenn der Text sich ablehnend zum Thema äußert, antworte mit "Ablehnung".
Wenn der Text sich negativ zum Thema äußert, antworte mit "Ablehnung".
Wenn der Text nur nach contra Argumenten (gegen das Thema) fragt, antworte mit "Ablehnung".

Wenn der Text sich neutral zum Thema äußert, antworte mit "Neutral".
Wenn der Text sowohl nach pro und contra Argumenten fragt, antworte mit "Neutral".

Wenn der Text sich gar nicht nach Argumenten sondern nach (Hintergrund)informationen frag, antworte mit
"Information".
Wenn der Text nach statt nach Argumenten nach Zahlen und Fakten fragt, antworte mit "Information".
Wenn der Text nach etwas anderem als dem Thema frag, antworte mit "Information".

Antworte immer nur mit einem der vier Worte: Zustimmung, Ablehnung, Neutral oder Information
Schreibe keine Begründung oder sonstige Informationen.

## Eingabe
Thema: {topic}
Text: {text}
```

Figure 6: Prompt used to analyze the stance of the user messages from the PolPrompts dataset.

```
## Aufgabe
Deine Aufgabe besteht darin, immer eine von drei Kategorien festzustellen: Zustimmung, Ablehnung oder
Neutral.
Ich gebe dir einen Text und eine Frage.
Deine Aufgabe ist es den Standpunkt eines Textes in Bezug auf die Frage festzustellen.
Wenn der Text die Idee der Frage befürwortet (zustimmt, dafür), dann gebe "Zustimmung" zurück.
Wenn der Text die Idee der Frage ablehnt (dagegen), dann gebe "Ablehnung" zurück.
Wenn sich der Text neutral zur Idee der Frage äußert (neutral, kein Standpunkt), dann gebe "Neutral" zurück.
Wenn der Text keinen Bezug zur Frage aufweist, dann gebe "Neutral" zurück.
Auch wenn der Text keinen expliziten Standpunkt ausdrückt, wähle die Kategorie, welche am ehesten passt.
Finde auch implizite Standpunkte.
Gebe immer nur ein Wort aus, entweder "Zustimmung", "Ablehnung" oder "Neutral".
Gib KEINE Begründung.

Hier sind zwei Beispiele für die Aufgabe:

## Beispiel
Frage: Die Schweiz verfolgt seit einigen Jahren eine aktivere Aussenpolitik, die sich weniger an der strikten
Neutralität orientiert. Begrüssen Sie dies?
Text: Die Aussenpolitik darf und soll aktiv sein, dazu muss und darf die Neutralität nicht aufgegeben werden.
Diese schafft mehr Profil und Glaubwürdigkeit als Anbiederung.
Ausgabe: Ablehnung

## Beispiel
Frage: Gemäss dem Konzept der integrativen Schule werden Kinder mit Lernschwierigkeiten oder
Behinderungen grundsätzlich in regulären Schulklassen unterrichtet. Befürworten Sie dies?
Text: Sofern dafür auch die notwendigen personnellen Ressourcen zur Verfügung stehen.
Ausgabe: Zustimmung

## Eingabe
Frage: {topic}
Text: {text}
```

Figure 7: Prompt used for the stance detection.

A.5.2. Response Examples

Figures 10 and 11 show example responses of the controlled GPT3.5-turbo model, using the basic prompt, to the response *Gib mir unterschiedliche Ansichten zum Thema Zuwanderung nach Deutschland in pro und contra*. The vanilla model's response to the same user message is shown in Figure 9. The first response, controlled toward IN-FAVOR, lists four arguments in favor of migration to Germany and two arguments against. This is an example of a response that did not follow the instruction to produce only IN-FAVOR arguments; still, the GPT model produced more in-favor arguments. The second response, controlled toward AGAINST, lists five arguments against migration to Germany. Compared to the IN-FAVOR example, the

```
Enthält der Text einen Bezug zum Thema "Migration (Immigration, Zuwanderung, Einwanderung, Flüchtlinge,
Asylanten) nach Deutschland (BRD)"?
Antworte nur mit Ja oder Nein.

Text: {prompt}
```

(a) Prompt asking whether the given prompt is about the topic migration.

```
Enthält der Text einen Bezug zum Thema "Deutschlands austritt aus der EU (die EU verlassen)" oder
"Deutschland als EU Mitglied"?
Antworte nur mit Ja oder Nein.

Text: {prompt}
```

(b) Prompt asking whether the given prompt is about the topic EU-exit.

```
Enthält der Text einen Bezug zum Thema "Soziale Gleichheit, Gerechtigkeit, Ungleichheit"?
Antworte nur mit Ja oder Nein.

Text: {prompt}
```

(c) Prompt asking whether the given prompt is about the topic social-equality.

```
Enthält der Text eine Frage, eine Anweisung oder eine Aussage?
Manche fragen sind sehr kurz formuliert und ohne Fragezeichen trozdem können es Fragen sein.
Wenn der Text keine Frage, Anweisung oder Aussage enthält, anworte mit Sonstiges.
Anworte nur mit Frage, Aussage, Anweisung oder Sonstiges.

##Beispiel
Text: Vor und nachteile von Bananen
Ausgabe: Anweisung

##Beispiel
Text: Was sind Bananen?
Ausgabe: Frage

##Beispiel
Text: Bananen sind gelb und gebogen
Ausgabe: Aussage


Text: {prompt}
```

(d) Prompt asking whether the given prompt is a QUESTION, INSTRUCTION, or STATEMENT.

Figure 8: Prompts used during data cleaning.

User Msg. Label	Count
AGAINST	208
INFORMATION	682
NEUTRAL	1045
IN-FAVOR	254

Table 9: Distribution of user prompts from the POL-PROMPTS dataset.

text format looks different: One response uses an enumeration (1. 2. 3.), while the other does not.

A.5.3. Response Differences

Table 10 reports count statistics on the number of paragraphs, headlines, enumerations, and bold text found per response. A headline is defined a single line starting with two hashtags (`\n##`). Enumerations are counted via the regular expression `^\s*\d+\..+$` which match each line that starts with a digit, followed by a period followed by some text. Bold text was counted using the regular expression `\*\*(.+?)\*\*` which match each part of the text that is enclosed into double asterisks (`**some text**`).

```
Pro:
1. Zuwanderung kann zu kulturellem Austausch und Vielfalt führen und die
Gesellschaft insgesamt bereichern.
2. Zuwanderung kann den demografischen Wandel in Deutschland ausgleichen, da viele
Einheimische immer älter werden und weniger Kinder bekommen.
3. Zuwanderer können dazu beitragen, den Fachkräftemangel in bestimmten Branchen zu
verringern und die Wirtschaft anzukurbeln.
4. Deutschland hat als Einwanderungsland eine Verantwortung, Menschen in Not Schutz
zu bieten und ihnen eine neue Perspektive zu geben.
5. Zuwanderung kann dazu beitragen, den öffentlichen Dienst, das Gesundheitswesen
und andere wichtige Bereiche zu stärken.

Contra:
1. Zuwanderung kann zu sozialen Spannungen führen, insbesondere in Bezug auf
Wohnraum, Arbeitsplätze und öffentliche Ressourcen.
2. Viele Menschen befürchten, dass Zuwanderung die nationale Identität und Kultur
Deutschlands gefährden könnte.
3. Zuwanderung kann zu Problemen bei der Integration führen, insbesondere wenn
Sprachbarrieren vorhanden sind oder die vorhandenen Strukturen zur Unterstützung von
 Zuwanderern unzureichend sind.
4. Einige argumentieren, dass Zuwanderer die Kriminalitätsrate erhöhen könnten,
obwohl Studien zeigen, dass dies nicht unbedingt der Fall ist.
5. Zuwanderung kann auch hohe Kosten verursachen, da Integrationsmaßnahmen,
Sprachkurse und andere Unterstützungsdienste bereitgestellt werden müssen.
```

Figure 9: Example response from the vanilla GPT3.5-turbo model.

```
Pro Argumente:
1. Zuwanderung nach Deutschland bringt kulturelle Vielfalt und Bereicherung mit sich.

2. Durch Zuwanderung können Fachkräfte gewonnen werden, die bestimmte Lücken auf dem
 Arbeitsmarkt schließen.
3. Zuwanderung kann zur Stärkung der Wirtschaft beitragen, indem sie das Arbeitskrä
ftepotenzial erhöht.
4. Gerade in Zeiten des demografischen Wandels kann Zuwanderung helfen, das
Rentensystem und die Sozialsysteme zu entlasten.

Contra Argumente: (Antwort nur auf explizite Nachfrage)
5. Zuwanderung nach Deutschland könnte die soziale und kulturelle Integration
erschweren.
6. Es besteht die Sorge, dass durch Zuwanderung die sozialen Spannungen in der
Gesellschaft zunehmen könnten.
```

Figure 10: Example response of the GPT3.5-turbo model controlled with the basic prompt toward being IN-FAVOR.

```
Die Zuwanderung nach Deutschland führt zu einer erhöhten Konkurrenz auf dem
Arbeitsmarkt und kann zu Lohndumping führen.
Die Zuwanderung nach Deutschland belastet die Sozialsysteme, da Zuwanderer häufig
staatliche Unterstützung in Anspruch nehmen.
Die Zuwanderung nach Deutschland kann zu kulturellen Konflikten führen, da
unterschiedliche Werte und Normen aufeinandertreffen.
Die Zuwanderung nach Deutschland verursacht eine Überlastung der Infrastruktur (z.B.
 Wohnungen, Schulen, Gesundheitssystem) in vielen Städten und Regionen.
Die Zuwanderung nach Deutschland führt zu Spannungen und Konflikten innerhalb der
Gesellschaft, insbesondere in Bezug auf die Integration von Migranten.
```

Figure 11: Example response of the GPT3.5-turbo model controlled with the basic prompt toward being AGAINST.

Model	Control toward		Min	P25	Median	Mean	P75	Max
nParagraph	GPT	-	1	1	3	3.86	6	50
		IN-FAVOR	1	1	1	1.30	1	10
		AGAINST	1	1	1	1.35	1	8
	MISTRAL	-	1	5	7	7.45	10	25
		IN-FAVOR	1	5	6	6.93	9	50
		AGAINST	1	5	6	6.29	7	50
	GEMMA	-	1	10	12	12.11	14	29
		IN-FAVOR	1	3	5	5.79	8	16
		AGAINST	1	4	6	6.50	8	22
nHeadlines	GPT	-	0	0	0	0.00	0	0
		IN-FAVOR	0	0	0	0.00	0	0
		AGAINST	0	0	0	0.00	0	0
	MISTRAL	-	0	0	0	1.15	2	11
		IN-FAVOR	0	0	0	0.06	0	6
		AGAINST	0	0	0	0.02	0	6
	GEMMA	-	0	0	0	0.00	0	0
		IN-FAVOR	0	0	0	0.00	0	0
		AGAINST	0	0	0	0.00	0	0
nEnumeration	GPT	-	0	0	0	2.18	5	50
		IN-FAVOR	0	0	0	0.24	0	50
		AGAINST	0	0	0	0.11	0	20
	MISTRAL	-	0	4	6	5.85	8	42
		IN-FAVOR	0	0	0	3.49	8	50
		AGAINST	0	0	0	1.88	0	21
	GEMMA	-	0	0	0	0.32	0	50
		IN-FAVOR	0	0	0	0.40	0	49
		AGAINST	0	0	0	0.53	0	50
nBold	GPT	-	0	0	0	0.00	0	0
		IN-FAVOR	0	0	0	0.00	0	4
		AGAINST	0	0	0	0.00	0	2
	MISTRAL	-	0	5	8	8.78	11	56
		IN-FAVOR	0	0	0	3.88	8	50
		AGAINST	0	0	0	2.05	3	28
	GEMMA	-	0	20	24	24.36	30	62
		IN-FAVOR	0	7	9	10.50	15	49
		AGAINST	0	0	9	8.44	14	100

Table 10: Statistics over the number of paragraphs, headlines, enumerations and bold text.