

Appeal, Align, Divide? Stance Detection for Group-Directed Messages in German Parliamentary Debates

Ines Rehbein♠, Maris Buttmann♣, Julian Schlenker♣, Simone Paolo Ponzetto♣

University of Münster♠, University of Mannheim♣
irehbein@uni-muenster.de, maris@buttmann.com
{julian.schlenker, ponzetto}@uni-mannheim.de

Abstract

This paper presents a new benchmark for detecting *group-based appeals*, i.e., positive or negative references towards social groups, in German parliamentary debates. In the first step, group mentions are identified as targets for stance detection. In the next step, three human annotators assign stance labels to the group mentions, coding the speaker’s perspective towards the specific group. The created benchmark data is then used to investigate the capability of Large Language Models (LLMs) for detecting politicians’ stances towards social groups. We explore the potential of different prompting strategies (zero-shot prompting, few-shot prompting, Chain-of-Thought) for this task and compare the results to a supervised BERT baseline, showing that in low-resource scenarios LLMs can outperform smaller fine-tuned models without the need for annotating large datasets.

Keywords: Group Appeals, Stance Detection, Parliamentary Debates, Political Text Analysis

1. Introduction

Recent work has highlighted the importance of social identity and group attitudes for political behavior. References to social groups often encode social variables like race, ethnicity, class, gender and religion, and it has been shown that such group-based references can impact voters’ political behaviour (White, 2007; Jackson, 2011; Weber and Thornton, 2012; Holman et al., 2015; Robison et al., 2021). Several studies have focused on how political actors or parties make strategic use of *group appeals*, i.e., positive or negative references towards social groups, when communicating with voters in order to gain new voters’ support or mobilise their core supporter base (Thau, 2019b; Huber, 2022; Thau, 2024; Huber and Haselmayer, 2024).

However, there may also be demand-side effects that impact the political behaviour of the supply side, influencing how politicians use group-based appeals. Particularly before elections, parties may be willing to change their policies toward certain target groups because they believe they could otherwise lose votes. A recent example from Germany is the more repressive policy toward refugees and asylum seekers shortly before the 2024 elections, triggered by a public debate after a rejected asylum seeker killed several people with a knife. The effects on politics and voter behaviour described above underscore the importance of observing and understanding group-based appeals in political communication, i.e., references to groups in combination with a positive or negative attitude toward those groups.

In the paper, we present a new benchmark for identifying group-based appeals in German parliamentary debates and evaluate the potential of

Large Language Models (LLMs) for this task. We decompose the problem into two subtasks, (i) identifying group references and (ii) predicting positive, neutral or negative stance towards each respective group. For the first task, we employ the out-of-the-box classifier of Klamm et al. (2023) that identifies mentions of *the people* and *the elite* in political text and classifies them into more fine-grained categories, such as people by nation, ethnicity, functionality, social variables and generic mentions. In addition, the model identifies mentions of *the elite*, defined as “individuals, groups and organisations that have a disproportionate influence on society” (Klamm et al., 2023). The detected group mentions are then used as targets for stance detection, where we evaluate four LLMs as well as different prompting strategies for this task. Figure 1 illustrates the overall pipeline. All data, baseline models and prompt templates are available from our GitHub repository.¹

2. Related Work

In this section, we review work on detecting group mentions and group-based appeals in political text. Unlike traditional stance detection in NLP, the targets for group appeal detection are not limited to an often small set of known entities or statements, but can include any mention of a social group.

2.1. Social Group Mentions

There is no uniform, generally accepted definition of the term “social groups” and different operationalisations have been used in the literature.

¹<https://github.com/umanlp/GroupAppeals>.

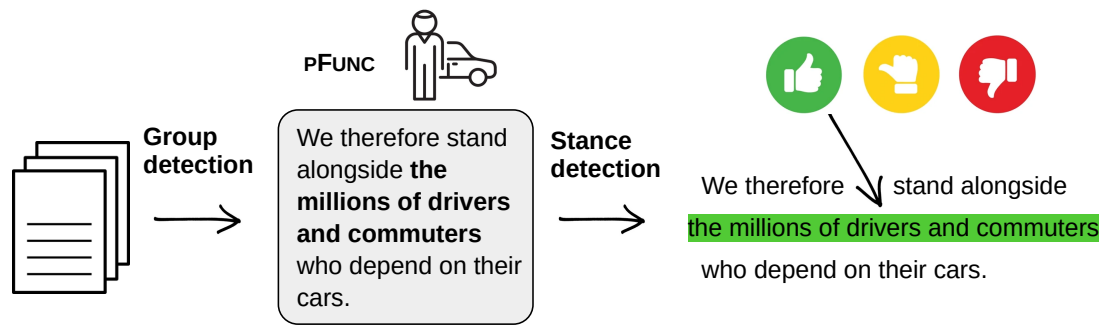


Figure 1: Pipeline for group-appeal detection in parliamentary debates (pFunc: People by Function, e.g., profession, function). Source: Spaniel, AfD, 27.09.2018 (translated to English by the authors).

Wlezien and Miller (1997) describe social groups as “categories that provide a source of identification for group members or information cues for non-members”. Building on this concept, follow-up studies have defined social groups as a collection of individuals who share common socio-demographic traits, such as gender, heritage, profession, religion, place of residence, or age (Rhodes and Johnson, 2015; Huber and Dolinsky, 2023; Stuckelberger and Tresch, 2024; Horne et al., 2025).

Some studies have limited the list of social groups they consider in their work to a predefined set of group terms while others do not provide any information on what they consider a social group. For instance, Thau (2019a) broadly refer to groups as a “category of people”. While Huber and Dolinsky (2023) explicitly exclude groups based on shared opinions, ideology, or party affiliation in order to maintain a narrow definition of social groups, more recent research reflects a shift towards broader frameworks, for example by incorporating political ideology such as *liberals* (Gevers et al., 2024; Licht and Sczepanski, 2025).

In our work, we adopt a broad definition of social groups where we also consider party affiliation, nationality, ethnicity, and political beliefs.

2.2. Group Appeals

Many works have shown how politicians use *group appeals* to leverage social group mentions as a strategic tool (Hersh and Schaffner, 2013; Holman et al., 2015; Thau, 2019a; Robison et al., 2021). However, as stated by Huber and Dolinsky (2023), research on group appeals often lacks a clear definition and consistent terminology, which hinders meaningful comparison across studies. Huber and Dolinsky (2023) describe group-based appeals as intentional acts that associate a political actor with a social group. Importantly, the authors argue that such appeals cannot be reduced to group mentions but also need to consider the stance associated with the group. Consequently, Huber (2022, p.294) define *group appeals* as positive or nega-

tive references to social groups, made by political actors in order to “mobilize the party’s core voters and broaden their support base among the general electorate”. Therefore, to grasp the full concept, we also need to identify the speaker’s stance towards a particular social group mentioned in the discourse.

Our work aims to support research in this area by providing both a method and a dataset for the automatic detection and evaluation of group appeals in German parliamentary debates.

2.3. Automatic Detection of Group Mentions and Group-based Appeals

Group mentions Extracting group mentions from text is far from trivial, as a single group can be referenced in multiple ways and each mention can include multiple words (e.g., “the working people”).

Approaches to extracting group mentions can be broadly categorised into four types. The first approach is based on **manual content analysis**, a labour-intensive process where human coders examine the text to identify each group reference in the data (Rhodes and Johnson, 2015; Thau, 2019a; Horn et al., 2021; Huber, 2022; Dolinsky, 2023). The second approach leverages dictionaries for a **rule-based extraction** of group mentions, looking for a set of predefined group terms (Dolinsky, 2023; Dausgaard and Hjorth, 2024). Another approach is based on **supervised token classification**, using fine-tuned, transformer-based models to identify group mentions (Klamm et al., 2023; Licht and Sczepanski, 2025; Horne et al., 2025). Finally, Jalali Farahani et al. (2024) employ **large language models** for token classification in a **few-shot setting** without fine-tuning. Among these approaches, supervised token classification based on small pre-trained language models has emerged as an effective method, demonstrating strong performance in recent studies (Klamm et al., 2023; Licht and Sczepanski, 2025).

In their work, Klamm et al. (2023) consider references to instances of *the People* and *the Elite* as proxies for *thin populism*, a “communication style

of political actors that refers to the people” (Jagers and Walgrave, 2007). To detect populist rhetoric, the authors released a dataset with manual annotations of references to *the People* and to *the Elite* and a token classification model that automatically predicts those references. The Elite is defined as all groups of individuals, organisations, institutions or others that hold a disproportionate amount of wealth, privilege, political power, or skill in society.

Their hierarchical annotation scheme encodes the distinction between the two groups, *the Elite* and *the People*, at the highest level. On the next level, elite mentions are further subdivided into persons and organisations, and level three of the taxonomy further divides the classes into 26 fine-grained categories, including elite groups from the political, military, religious domain or the media, while *the People* are split up into people by nation, by ethnicity, social variables, etc. They report results for their transformer-based classifier in the range of 78% F1 (micro) for the coarse-grained labels (top level) and around 74% for the fine-grained labels. Licht and Sczepanski (2025) present work similar in spirit, coding group mentions in British manifestos and applying a transformer-based classifier to detect group mentions.

In our work, we use the classifier of Klamm et al. (2023), fine-tuned on German parliamentary debates, to predict group mentions in our data.

Group-based appeals While the importance of group-based appeals has been established in the literature, not much work has been done to automatically extract group appeals. A notable exception is work in progress by Horne et al. (2025) who present a pipeline to analyse group-based appeals in German and British party manifestos. First, a multilingual DeBERTa model (He et al., 2020) is fine-tuned to predict whether a sentence contains at least one group mention. The model achieves a high accuracy for the task and is mostly used to extract sentences that potentially include group mentions. In the next step, a token classification model is trained on a set of 3,800 English and German sentences with manually coded group mentions.

After identifying the group mentions, the authors model stance detection as a sentence pair classification problem. They fine-tune a DeBERTaV3 model (He et al., 2021) for the stance detection task.² Finally, a third DeBERTa model classifies the identified group mention into one of 43 pre-defined categories for social groups.

Our work is similar in spirit to the one of Horne et al. (2025), with the following key differences. First, we focus on German parliamentary debates,

²Before fine-tuning, the model has additionally been fine-tuned on a Natural Language Inference (NLI) dataset, following the approach of Laurer et al. (2024).

a more argumentative text type compared to the manifestos, with a higher ratio of group mentions. Second, in addition to benchmarking a smaller pre-trained BERT model (Devlin et al., 2019) for this task, we also investigate the potential of large language models (LLMs) for the identification of group-based appeals, to find out whether the labour-intensive process of data annotation is still necessary and whether we can apply those models to new domains where no annotated data for fine-tuning is available.

3. Data and Annotation

We now introduce the data and annotation process used to create the benchmark for detecting group appeals in German parliamentary debates.

3.1. Data

The data has been sampled from the set of all German parliamentary debates held from September 1949 to June 2025. For newer legislative terms starting from 2017, we collected the transcripts provided by the open data service of the German Bundestag.³ For older legislative terms where the original data is not in an easily parsable format, the pre-structured GERMAPARLTEL corpus from the PolMine project has been used (Blaette, 2017).

Preprocessing We extracted all speeches by members of the Bundestag but excluded speeches held by the Bundestag presidents, which mainly contain moderating content. From the data, we removed all interjections and extracted 3,242,437 individual paragraphs that we processed, using the group mention classifier of Klamm et al. (2023).⁴ The classifier assigns token level annotations, using the BIO scheme (Ramshaw and Marcus, 1995). We filter out noise in the predicted annotations in a postprocessing step by only including paragraphs where the tagged group mentions start with a “B” (Beginning) tag. We also discarded all spans tagged as geo-political units (GPE) as they do not refer to social groups. As a result, we end up with a large dataset in which 6,970,801 group mentions have been identified by the classifier.

3.2. Annotation Process

Sampling We sample 1,000 instances from the preannotated debates to create the benchmark dataset, excluding very long paragraphs with more

³The transcripts are made available from the open data service of the German Bundestag: <https://www.bundestag.de/services/opendata>.

⁴We used model 3, available from <https://github.com/umanlp/mope>, to predict group mentions.

than 3,000 characters. As the class distribution of the group mentions is heavily imbalanced, we combine random undersampling with a fixed extraction number per group label to create a more balanced dataset.

This method ensures that we extract a sufficient number of instances for the lower-frequency classes while being able to preserve the class distribution in the original data in a moderated fashion. From each of the 25 label classes, at least 10 mentions will be extracted. The extraction of the remaining 750 labels will be guided by their respective group sizes. However, the frequent EPOL and EOPOL labels (political elite, person and political elite, organisation) will be penalised by extracting only one-fortieth of their proportional share, as they occur in over 66% of the preprocessed paragraphs. We also downsample the PEOPLE BY FUNCTION (PFUN) class to one-fifth of its original size. This results in a total sample size of 999 instances due to the application of a floor function.⁵ These samples were drawn using reservoir sampling (Vitter, 1985) with a fixed seed, to ensure reproducibility.

Annotation Procedure Each of the 999 paragraphs was labelled independently by three student annotators who were provided with detailed guidelines, including definitions of the different classes and instructions for handling edge cases. In the training phase, the three coders labelled a batch of 150 instances each. Then we identified disagreements in the annotations and provided feedback to assure alignment with the guidelines. After this training phase, the coders annotated the remaining instances in the benchmark dataset.

Annotators were instructed to assign stance to paragraphs extracted from longer speeches. The coders could choose from four labels: {FAVOUR, AGAINST, NEITHER, NOT A GROUP}. The “not a group” label was used exclusively to discard paragraphs where the group mention classifier had incorrectly identified a text span as a group mention. Following Küçük and Can (2020), we adopt the three-way classification of {FAVOUR, AGAINST, NEITHER} for identifying group-based appeals.

Incorrect group mentions To resolve disagreements concerning the correct group spans, we inspected the 36 instances marked as “not a group” by at least one annotator.⁶ Out of these, the annotators agreed on 9 instances. The remaining 27

instances were manually reviewed by the annotators to reach a consensus. After resolving the disagreements, the incorrectly predicted group spans were removed from the data⁷ and the annotators assigned stance labels to the remaining instances.

Group mention spans In some cases, the classifier correctly identified a group mention but struggled to extract the complete span. In a manual post-processing step, we corrected the span boundaries in the benchmark data. 160 out of the remaining 974 paragraphs in the final dataset required such span correction.

3.3. Data Curation and Disagreements

We acknowledge the subjectivity of the task and therefore do not consider disagreements as annotation noise. Most disagreements concern cases where one or two coders selected the “neither” label while the remaining coder(s) assigned a positive or negative stance. We observed seven instances with opposing labels (positive – negative) and asked the coders to discuss those in order to correct straightforward annotation errors. Four of the seven cases could be identified as errors and have been corrected. For the remaining three instances, a consensus of “neither” was reached for one, while two disagreements remained. Our final benchmark dataset includes 973 labelled group mentions.

Keeping all the individual labels assigned by the three coders enables us to identify group-based appeals with a strong, unambiguous stance and distinguish them from the ones that are open to interpretation. For some applications, we might be mostly interested in inspecting clear cases where the speaker expresses a strong and unambiguous stance towards the target group to send a strong signal to their voter base. For other applications, we might also be interested in group-based appeals with a less explicit stance, which can serve to send a message to certain, more extreme voter groups while at the same time trying not to alienate the core electorate.

Inter-annotator agreement (IAA) We report three metrics to assess IAA for stance towards group mentions: (i) overall percentage agreement, (ii) Fleiss’ Kappa (κ), and (iii) results for the Dawid-Skene model, a Bayesian model of annotation, to provide a comprehensive understanding of disagreements.

We observe a percentage agreement of 67.5%, with a Fleiss’ κ of .63 for the three coders, which indicates a moderate level of agreement (Landis and Koch, 1977). Pairwise percentage agreement

⁷25 incorrect group spans have been removed.

⁵The final counts for the group labels selected for manual annotation are shown in Table 6 in the appendix.

⁶As this is not an inherently subjective task, the correction of the group mentions has been done by two of the coders while the third coder was presented with the correct group mentions.

Setting	Favour	Against	Neither
MV	410	95	468
Neither	286	64	623

Table 1: Distribution of labels for each evaluation setting (MV: majority vote; Neither: all disagreements considered as 'Neither').

between the coders is considerably higher, with scores between 77.7% and 78.6% for each coder pair. Not surprisingly, when excluding instances where at least one annotator assigned the “neither” label, the κ score rises to a near-perfect .978.

Data curation The most straightforward way to create ground truth labels for training and evaluation, based on the annotations of multiple coders, is to take the majority vote over the individual labels assigned to an instance. In recent years, this approach has been criticised primarily in connection with subjective tasks, where several, possibly contradictory views may be valid at the same time (Plank, 2022). Simply taking the majority vote would thus result in ignoring important signals encoded in the annotations. Several approaches have been proposed to address this issue, among them the use of soft labels or confidence scores to reflect the inherent ambiguity for labelled instances (Peterson et al., 2019), or employing multi-annotator models that predict the individual annotators’ decisions in a multi-task setup with shared representations (Mostafazadeh Davani et al., 2022). Other works decided to simply accept all labels that have been predicted by at least one of the coders, in order to take *all* perspectives on board (Erhard et al., 2025).

To determine the best approach for our task and to investigate annotator bias in our data, we employ Bayesian models of annotation, specifically, the Dawid-Skene model (Dawid and Skene, 1979; Passonneau and Carpenter, 2014), to investigate the reliability of our coders and to look for inherent biases.⁸ This approach also allows us to test whether the annotators encode different perspectives on the data.

The Dawid-Skene model provides a Maximum Likelihood Estimation of the error rates in human annotation, using the EM algorithm, and determines the most probable label for each instance. In addition, the model provides us with reliability scores for each coder. We run the model on our annotations and observe high reliability scores for all three coders (0.87-0.88). We then compare the model’s predictions with the labels obtained from the majority vote. We find perfect agreement between the two variants, showing that none of the coders ex-

⁸We use the implementation of the Dawid-Skene model provided here: <https://github.com/kajjuuen/Dawid-skene>.

hibits a strong bias toward a particular label. We take this as motivation for using the majority vote as ground truth for training and evaluation.

One drawback of the majority vote, however, is that it does not allow us to distinguish clear FAVOUR/AGAINST appeals from the less clear cases where one of the coders disagreed with the label assigned by the majority of annotators. We therefore curate a second version of the dataset where we only consider instances where all three coders agreed on the FAVOUR/AGAINST label as instances of the respective class, while all other instances are labelled NEITHER. This approach aims at training a classifier that detects only clear, unambiguous group-based appeals. While this approach might be more suitable for some research questions, it also has the drawback that the resulting dataset is more imbalanced due to the significant increase in the number of NEITHER labels. The distribution of labels for both variants of our benchmark is shown in Table 1. Our benchmark gives users the flexibility to adapt the ground truth and thus the model’s predictions for their purposes. We report results for both settings.

4. Experiments

4.1. Experimental Setup

Data split We split the data into development set (10%) and test set (90%). The development data is used for prompt development and selecting few-shot examples, while the test set is reserved exclusively for the final benchmarking of the LLM prompting frameworks.

Models We compare an open-weight model, the instruction-tuned Gemma-3-27b (Gemma Team et al., 2025), with a closed-source model, Gemini 2.5 Pro (Gemini Team et al., 2024). In addition, we benchmark two other popular LLMs from the Llama family, LLAMA-3.1-8B and LLAMA-3.3-70B.

BERT Baselines As a baseline, we employ two pre-trained encoder-only models trained exclusively on German data: GBERT-large (hereafter GBERT; 335M parameters) (Chan et al., 2020) and ModernGBERT (134M parameters) (Ehrmantraut et al., 2025).

For the stance detection task, we fully fine-tune each model and train a classification head on top to perform multi-class, single-label sequence classification, where the final label corresponds to the majority vote of the three annotators. For more detailed training configurations, see Section A.2 (appendix). To indicate the target group mention within the input, we extend the vocabulary by adding spe-

cial / tokens, which are wrapped around the critical mention.

Since no dedicated training set is available, we adopt a 5-fold cross-validation strategy: in each iteration, the model is fine-tuned on the development data and four folds (80%) of the test data, and evaluated on the remaining fold (20%). This ensures that every instance of the test data serves once as validation. After completing all five iterations, we aggregate the results across folds and compute metrics.⁹

gemma-3-27b-it (GEMMA-3) is a multimodal, instruction-tuned model from Google with 27 billion parameters, featuring a 128k token context window. GEMMA-3 builds upon previous versions in the series (Gemma Team et al., 2025) and has been pre-trained on a 14T token dataset composed of a mixture of images and text. Its subsequent alignment phase employed distillation, instruction tuning, and Reinforcement Learning from Human Feedback (RLHF). As of mid-2025, it is one of the highest-ranking, publicly available language models in its parameter class (White et al., 2024; LMArena, 2025).

gemini-2.5-pro (GEMINI-2.5) is a multimodal, sparse Mixture-of-Experts (MoE) model leading multiple industry benchmarks (as of mid-2025) (White et al., 2024; LMArena, 2025). GEMINI-2.5 has a context window exceeding one million tokens, capable of processing text, images, video, and audio. The model has been pre-trained on a large, though not further specified, dataset of public web documents, code, and multimodal data (images, video, and audio). Its performance and safety are further enhanced through alignment techniques, including instruction tuning and RLHF.

While the authors do not disclose the exact mechanics of the model’s reasoning process, they describe a system fine-tuned to perform up to tens of thousands of forward passes. During this process, the model generates and evaluates multiple internal hypotheses before producing a final answer. This proprietary description could be interpreted as an advanced implementation of Chain-of-Thought (CoT) reasoning that incorporates both multiple reasoning paths and self-verification. This model was selected to compare the performance of a “ready-to-use” API pipeline against that of an open-weight approach. The term “ready-to-use” refers to how the pipeline can automatically perform techniques, such as generating multiple reasoning chains, performing self-verification, or conducting

⁹We also experimented with fine-tuning only on the development set; however, this low-resource setup yielded scores below chance level.

a web search that would normally be implemented manually alongside an open-weight model.¹⁰

Llama-3.1-8B is also a multilingual instruction-tuned model and with 8 billion parameters the smallest in our experimental setup. In contrast to the first two models, LLAMA-3.1-8B is a text-only model.

Llama-3.3-70B We include another, larger model from the Llama family with a parameter size of 70 billion in order to test the impact of model size on results. As LLAMA-3.1-8B, this model is a multilingual text-only model tuned on a variety of languages and refined, using instruction tuning and RLHF.

Inference settings A notable consideration for instruction-tuned models like gemma-3-27b-it is that they benefit from the use of chat templates, which structure the input to distinguish between a high-level “system prompt” and a task-specific “user prompt”. Since Google Gemini also supports this distinction, the preliminary benchmarks were also used to determine the optimal prompt structure. We compared two approaches: The first, a “split-prompt” approach, places the detailed task description and any few-shot examples in the system prompt. In this setting, the user prompt contains only the paragraph and group mention to be analysed. The second, a “full-prompt” approach, provides all information solely within the user prompt. As we obtained best results using the “split-prompt” approach, we apply this setting in all experiments. For both models, the temperature is set to 0.0 to ensure the results are reproducible.

4.2. Prompting Techniques

We compare a baseline zero-shot approach against several advanced prompting techniques, namely few-shot prompting with 1, 5, and 10 shots, CoT prompting and expert prompting (Xu et al., 2025; Lan et al., 2024) for our task. In expert prompting, we assign the LLM the role of a precise analyst of political language. As this approach consistently improved results, we use it in all prompt settings.

Zero-shot We adopt the zero-shot prompt template of (Zhang et al., 2023) and present the model with the following instruction:

“Perform Stance Detection. Given the text and the target, assign a stance label from [‘favour’, ‘against’, ‘neither’]. Return the label only”.

The baseline prompt is shown in Fig. 2 (without the dashed box). Next, we iteratively adapt the prompt, using the development set to select the

¹⁰To ensure a fair(er) comparison, supplemental API features like the integrated web search were disabled.

Model	prec	rec	F1
GBERT-large			
favour	67.9 \pm 1.5	70.6 \pm 3.1	69.2 \pm 2.1
against	49.8 \pm 11.8	25.6 \pm 9.5	33.3 \pm 9.9
neither	67.4 \pm 0.1	72.1 \pm 1.7	69.7 \pm 0.7
total (weighted)	65.8	66.7	65.8
ModernGBERT			
favour	61.5 \pm 2.8	65.5 \pm 3.4	63.3 \pm 0.2
against	33.3 \pm 6.0	4.8 \pm 2.3	8.3 \pm 3.8
neither	63.4 \pm 2.2	71.2 \pm 3.8	67.0 \pm 0.5
total (weighted)	59.5	62.0	59.5

Table 2: Micro F1 for individual classes and weighted F1 for two baseline BERT models, averaged over three runs with different seeds.

most promising approach. The final prompt template with improved instructions is shown in Fig. 4 (right) in the appendix. The improved prompt instructions lead to a substantial increase in results and we apply them in all prompt settings.

Few-shot For few-shot prompting, the task description is followed by k examples, each containing a text paragraph, the group mention and its corresponding label (see Figure 2 with dashed box). To ensure a controlled comparison, a fixed set of k examples was assigned to each paragraph in the test set for $k \in \{1, 5, 10\}$. This guarantees that each model sees the exact same few-shot examples for each prompt. The examples themselves are drawn from the engineering set, using unambiguous instances where the annotators agreed on the label. Sampling was conducted using the reservoir algorithm. For few-shot prompting, we also use the improved prompt instructions.

CoT For Chain-of-Thought prompting, we replace the part of the prompt that restricts the output to one of the three labels. The new instruction directs the model to first “think step-by-step” and then, after presenting its reasoning, to provide the final predicted stance in a `<stance>label</stance>` format (see Appendix, Figure 4, right). While this prompt template gave good results for GEMINI-2.5, it did not work very well for the open-weight models. After some prompt engineering, we decided to use a slightly simplified version for the three models, shown in Figure 6 in the Appendix.

4.3. Results

4.3.1. BERT Baselines

Table 2 reports results for the two BERT baselines, evaluated on the test set using 5-fold cross-validation (see Section 4.1). GBERT consistently

outperforms ModernGBERT across classes, arguably due to its larger model size. Both models achieve moderate F1 scores for the majority classes *favour* and *neither*, reaching up to 69.2 and 69.7, respectively. Performance on the minority class *against* remains low. This is not surprising given that the instances account for only around 10% of the data.

Overall, our results indicate that standard fine-tuning for sequence classification struggles to clearly separate the stance classes, attributed to the dataset’s limited size and its pronounced class imbalance.

4.3.2. LLM Results

We now report results for the four large language models (Table 3).

LLMs vs. BERT Compared to the BERT baselines, we see substantial improvements for three of the four models. GEMINI-2.5 outperforms the best BERT baseline by 12%, LLAMA-3.3-70B by around 8% and GEMMA-3 by around 5%. The smallest Llama model with 8b parameters, however, yields results that are slightly below the best fine-tuned BERT baseline (64.7% vs. 65.8% F1).

Compared to BERT, we see the largest improvements for the minority class *against* where we now obtain scores up to 76% F1 for GEMINI-2.5 and 64% F1 for LLAMA-3.3-70B while the fine-tuned GBERT model only yielded 33% for this class. However, scores for the *against* class are still substantially lower than the ones for *favour* and *neither*, probably due to the fact that the minority label is also underrepresented in the few-shot examples shown to the model.

Impact of model size Not surprisingly, size has a large impact on model performance. Best results for the largest model, GEMINI-2.5, are around 78% weighted F1 and outperform the best scores obtained by LLAMA-3.1-8B, our smallest model, by nearly 14%. Results for the other two models range between 71% (GEMMA-3) and 75% (LLAMA-3.3-70B), making them a cost-effective alternative to the closed-source model, as they require fewer parameters and less computing power, yet still achieve reasonable performance on our task.

Impact of prompting strategy For the closed-source model, results for all prompting strategies are similar with around 78% F1, indicating that there is some prompt tuning behind the scenes that is not transparent to the user. For the open-weight models, however, the prompting strategy is of crucial importance and few-shot prompting with 5 to 10 examples performs best for our task. CoT prompting,

Model	LLAMA-3.1-8B			GEMMA-3			LLAMA-3.3-70B			GEMINI-2.5		
	prec	rec	F1	prec	rec	F1	prec	rec	F1	prec	rec	F1
<i>zero-shot</i>												
favour	81.6	16.8	27.9	87.3	33.6	48.5	78.2	66.3	71.7	75.9	91.3	82.9
against	0.0	0.0	0.0	44.4	65.6	52.9	52.0	70.7	59.9	67.8	86.7	76.1
neither	51.3	97.8	67.3	60.1	86.4	70.9	71.4	76.2	73.7	86.2	65.7	74.6
weighted	58.9	53.9	43.9	69.9	62.1	59.7	72.6	71.5	71.7	80.0	78.6	<u>78.2</u>
<i>1-shot</i>												
favour	69.7	35.5	47.0	83.9	49.3	62.1	78.6	66.7	72.1	75.3	91.6	82.6
against	76.7	25.8	38.7	35.0	83.3	49.3	52.6	78.9	63.1	63.1	85.6	72.6
neither	56.2	88.3	68.7	68.8	73.3	71.0	72.6	74.5	73.5	85.7	62.9	72.5
weighted	63.9	59.8	56.6	71.6	64.3	65.0	73.0	71.6	71.9	79.0	77.2	76.8
<i>5-shot</i>												
favour	64.5	70.5	67.4	82.2	57.5	67.6	78.3	73.4	75.7	76.8	94.0	84.8
against	53.7	24.4	33.6	37.7	86.7	52.5	60.2	75.6	67.0	61.6	85.6	71.6
neither	67.8	70.2	69.0	72.7	71.7	72.2	74.3	74.3	74.3	88.4	63.6	74.0
weighted	65.0	65.6	64.7	73.1	67.2	68.3	74.5	74.0	74.1	80.8	78.6	<u>78.2</u>
<i>10-shot</i>												
favour	63.1	67.2	65.1	79.7	66.1	72.3	78.4	77.8	78.1	76.4	90.2	82.7
against	75.0	20.0	31.6	41.1	82.2	54.8	57.4	73.3	64.4	65.0	84.4	73.4
neither	64.7	71.2	67.8	76.3	71.4	73.8	77.1	73.1	75.1	84.7	65.7	74.0
weighted	65.1	64.3	63.0	74.2	70.3	71.2	75.7	75.1	75.2	79.2	77.9	77.6
<i>CoT</i>												
favour	70.6	54.0	61.2	75.3	71.8	73.5	70.5	90.5	79.2	78.0	86.7	82.2
against	24.5	86.4	38.2	36.9	92.2	52.7	57.1	88.9	69.6	67.3	82.2	74.0
neither	68.8	46.3	55.3	77.1	55.5	64.5	86.8	54.8	67.1	81.9	70.0	75.5
weighted	65.1	53.6	56.0	72.2	66.1	67.1	76.9	73.3	72.5	78.8	78.3	78.1

Table 3: Weighted F1 for LLAMA-3.1-8B, Gemma-3-27b-it, LLAMA-3.3-70B and Gemini-Pro (majority vote) for different prompting strategies. Best results for each model are bold, overall best results are underlined.

on the other hand, does not yield the expected improvements. In contrast to the larger open-weight models, the small LLAMA-3.1-8B does not benefit from additional examples for few-shot prompting but yields best results for the 5-shot setting.

While results heavily depend on model size and prompting strategy, other architectural differences between the models do not seem to have much impact at least for our task.

4.4. Error Analysis

We now provide a qualitative error analysis, focussing on the following two questions:

- (i) Why are the results for the AGAINST class substantially below the ones for the other classes?
- (ii) Why does zero-shot CoT prompting fail to improve results in our experiments, in contrast to previous results from the literature?¹¹

4.4.1. Lower accuracy for the AGAINST class

To answer our first question, we focus on the predictions of the best open-weight model, LLAMA-3.3-70B, with 10-shot prompting. The model achieves an overall weighted F1 of 75.2% with an F1 of 64.4% for the AGAINST class. For comparison, our

supervised BERT baseline scored 65.8% F1 on the same data, with an F1 of 33.3% for AGAINST instances.

An obvious explanation for the lower results for the AGAINST class in the supervised setting is the label distribution, where AGAINST instances account for less than 10% of the instances in our data. This, however, cannot explain the lower results in the LLM experiments. To further investigate this, we take a closer look at the classification errors for the AGAINST class. Out of the 90 AGAINST instances in the test set, only 2 have been predicted as FAVOUR by the model while 22 of the instances labelled as AGAINST by our human annotators have been predicted as NEITHER.

To get a better idea of the errors made by the model, we classify the erroneous instances into the following categories: (1) The correct answer requires world knowledge (this includes coreference resolution problems, such as knowing whether a certain party can be considered as coreferent with the government at a certain point in time); (2) The negative stance is highly implicit; (3) The negative stance is expressed through lexical connotation of the group term (e.g., *organised crime*); (4) The speaker expresses negative stance toward some event associated with the group (rather than toward the group itself); and (5) Ambiguous cases that are also hard for human coders.

Table 4 shows that most of the incorrect model

¹¹See, e.g., Kojima et al. (2022); Wei et al. (2022).

	Error type	No. instances
(1)	World knowledge	2
(2)	Implicit stance	6
(3)	Lexical connotation	1
(4)	Stance toward event	15
(5)	Ambiguous cases	4

Table 4: Error types for the Llama-3-70b model (10-shot setting) for misclassifications of the AGAINST class (more than one error type can apply to the same instance).

predictions for the AGAINST class are cases where the speaker has a negative stance toward an event and, by inference, also toward the group associated with or responsible for the event (Ex. 1). Other common errors concern implicit stance as we often see in political debates.

- (1) The German feuilleton ponders the liberating potential of hatred. Describing men as an evolutionary relic, women and, even more so, androgyny as the future, as the author duo Welpe and Welpe do, has become completely normal.

4.4.2. CoT prompting errors

We now focus on cases where the LLAMA-3.3-70B model made incorrect predictions when prompted in the CoT setting but provided correct labels when using few-shot prompting. There are 213 cases where the CoT predictions did not agree with the ones from the 10-shot setting. Out of those, the CoT predictions are correct in 92 cases while the 10-shot prompting strategy predicted the correct labels for 112 of those cases. For 9 instances, both prompting strategies failed to predict the correct answer.

Interestingly, only 13 of the 213 cases that were incorrectly predicted in the CoT setting but not in the 10-shot setting are instances of the AGAINST class (following the class distribution, we would expect nearly twice as many AGAINST instances). Instead, most of the errors are NEITHER instances where the CoT model either predicted AGAINST or FAVOUR. Comparing the results, it seems as if the few-shot models are better at adapting to the label distribution in the data while the same model with zero-shot CoT prompting has a strong bias toward the FAVOUR class. Ex. 2 shows a typical example where the CoT version of the model predicted FAVOUR while our human annotators and the 10-shot model voted for NEITHER (see Fig. 7 in the appendix for the step-by-step reasoning provided with the answer).

- (2) Wherever people are involved, mistakes are made; there is no question about that. However, efforts must be made to correct mistakes and avoid them in the future. You refuse to

adopt such an attitude. You do not want to avoid mistakes. On the contrary, when I look at the 2005 budget, I see it as the best proof that you have learned nothing. You continue to set revenues too high and expenditures too low in order, as the Frankfurter Allgemeine Zeitung has said, to formally cheat and deceive.

4.5. Discussion

Our results show that for detecting group appeals in low-resource settings LLMs are an attractive alternative to smaller supervised models, as they make the time-consuming process of data annotation superfluous. However, running LLMs comes at high computational and, for closed-source models, also financial costs. Processing large data with such compute-heavy models seems inefficient and, from an energy policy perspective, also not desirable. We therefore suggest using the approach to collect more training data for fine-tuning smaller models, especially for the minority class. This approach, however, still needs to be validated in order to find out whether the amount of noise in the data is tolerable for it to be useful. Here, different scenarios are possible. We can put a human in the loop to ensure the quality of the data, or we can adapt the model to give more weight to the manually coded examples. We leave this for future work.

5. Conclusions

In the paper, we presented a new benchmark for detecting group-based appeals in German parliamentary debates. We then reported baseline results for two BERT models, fine-tuned on our data, and showed that in low-resource scenarios like ours LLMs can easily outperform a transformer baseline even without fine-tuning. Our evaluation showed that, not surprisingly, results heavily depend on model size and that few-shot sampling yields best results for the detection of group-appeals while CoT prompting failed to further improve the scores.

Our findings indicate some directions for future work. As CoT prompting did not work well for our task, future work should explore whether more advanced sample selection for few-shot prompting can yield further improvements. Ensuring that the model sees enough instances of the minority class is one important issue that should be addressed. Another approach might look at efficient ways to select examples for few-shot prompting that are most informative for the model. What needs to be explored is how we can best operationalise the concept of “informativeness”, i.e., whether the similarity of topic is more important than the target group featured in the examples.

6. Limitations

Our work comes with limitations, the most obvious one being the high computational and, for closed-source models, financial cost of LLMs, raising questions about the efficiency of using very large models for tasks like stance detection, especially when smaller, fine-tuned models can be more resource-efficient. To answer this question, it would be good to know how many annotated training instances are needed to obtain competitive performance with a smaller, fine-tuned model. Another potential limitation is the size of our benchmark. Especially the minority class could benefit from more annotated examples. It is also not clear how well the different time periods in our diachronic corpus are represented in the benchmark. While we took care to sample instances across all legislative terms, it is clear that a representative sample would need more data.

Acknowledgments

The work presented in this paper is funded by the German Research Foundation (DFG) under the UNCOVER project (PO1900/7-1 and RE3536/3-1). We would like to thank our annotators, Lilly Brauner and Harriet Eser, for their dedicated work. We would also like to thank the anonymous reviewers for their constructive feedback.

7. Bibliographical References

- Andreas Blaette. 2017. [PolMine/GermaParlTEI: GermaParl: Corpus of Plenary Protocols of the German Bundestag \(TEI Format\)](#).
- Branden Chan, Stefan Schweter, and Timo Möller. 2020. [German’s next language model](#).
- Christoffer H Dausgaard and Frederik Hjorth. 2024. Elite Rhetoric and the Running Tally of Party-Group Linkages. *Working Paper*.
- Alexander Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 1(28):20–28.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alona O Dolinsky. 2023. [Parties’ group appeals across time, countries, and communication channels—examining appeals to social groups via the Parties’ Group Appeals Dataset](#). *Party Politics*, 29(6):1130–1146. Publisher: SAGE Publications Ltd.
- Anton Ehrmantraut, Julia Wunderle, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. [ModernBERT: German-only 1b encoder model trained from scratch](#).
- Lukas Erhard, Sara Hanke, Uwe Remer, Agnieszka Falenska, and Raphael Heiko Heiberger. 2025. [PopBERT. Detecting Populism and Its Host Ideologies in the German Bundestag](#). *Political Analysis*, 33(1):1–17.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, ..., and Oriol Vinyals. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). ArXiv:2403.05530 [cs].
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, ..., and Léonard Hussenot. 2025. [Gemma 3 Technical Report](#). ArXiv:2503.19786 [cs].
- Ine Gevers, August De Mulder, and Walter Daelemans. 2024. [Towards a large scale analysis of claims: developing a machine learning method for detecting and classifying politicians’ claims of representation](#). *Journal of Computational Social Science*, 7(1):917–961.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTaV3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). *CoRR*, abs/2111.09543.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.
- Eitan D. Hersh and Brian F. Schaffner. 2013. [Targeted Campaign Appeals and the Value of Ambiguity](#). *The Journal of Politics*, 75(2):520–534. Publisher: The University of Chicago Press.
- Mirya R. Holman, Monica C. Schneider, and Kristin Pondel. 2015. [Gender Targeting in Political Advertisements](#). *Political Research Quarterly*, 68(4):816–829. Publisher: SAGE Publications Inc.

- Alexander Horn, Anthony Kevins, Carsten Jensen, and Kees van Kersbergen. 2021. [Political parties and social groups: New perspectives and data on group and policy appeals](#). *Party Politics*, 27(5):983–995. Publisher: SAGE Publications Ltd.
- Will Horne, Alona Dolinsky, and Lena Huber. 2025. [Using LLMs to Detect Group Appeals in Parties' Election Manifestos](#).
- Lena M. Huber and Matthias Haselmayer. 2024. Promising links: How parties combine policy issues with group appeals. *West European Politics*, pages 1–28.
- Lena Maria Huber. 2022. [Beyond Policy: The Use of Social Group Appeals in Party Communication](#). *Political Communication*, 39(3):293–310. Publisher: Routledge _eprint: <https://doi.org/10.1080/10584609.2021.1998264>.
- Lena Maria Huber and Alona O. Dolinsky. 2023. [How parties shape their relationship with social groups: A roadmap to the study of group-based appeals](#).
- M. S. Jackson. 2011. [Priming the sleeping giant: The dynamics of latino political identity and vote choice](#). *Political Psychology*, 32(4):691–716.
- Jan Jagers and Stefaan Walgrave. 2007. Populism as Political Communication Style: An Empirical Study of Political Parties' Discourse in Belgium. *European Journal of Political Research*, 46(3):319–345.
- Farane Jalali Farahani, Sara Hanke, Corina Dima, Raphael Heiko Heiberger, and Steffen Staab. 2024. [Who is targeted? Detecting social group mentions in online political discussions](#). In *Companion Publication of the 16th ACM Web Science Conference, Websci Companion '24*, pages 24–25, New York, NY, USA. Association for Computing Machinery.
- Christopher Klamm, Ines Rehbein, and Simone Paolo Ponzetto. 2023. [Our kind of people? detecting populist references in political debates](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1227–1243, Dubrovnik, Croatia. Association for Computational Linguistics.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Dilek Küçük and Fazli Can. 2020. [Stance detection: A survey](#). *ACM Comput. Surv.*, 53(1).
- Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. [Stance Detection with Collaborative Role-Infused LLM-Based Agents](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 18:891–903.
- J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).
- Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. [Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli](#). *Political Analysis*, 32(1):84–100.
- Hauke Licht and Ronja Sczepanski. 2025. [Detecting group mentions in political rhetoric a supervised learning approach](#). *British Journal of Political Science*, 55:e119.
- LMarena. 2025. [Overview Leaderboard | LMarena](#). <https://lmarena.ai/leaderboard> (Accessed: 12/07/2025).
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. [Dealing with disagreements: Looking beyond the majority vote in subjective annotations](#). *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Rebecca J. Passonneau and Bob Carpenter. 2014. [The Benefits of a Model of Annotation](#). *Transactions of the Association for Computational Linguistics*, 2:311–326.
- Joshua Peterson, Ruairidh Battleday, Thomas Griffiths, and Olga Russakovsky. 2019. [Human uncertainty makes classification more robust](#). In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Lance Ramshaw and Mitch Marcus. 1995. [Text chunking using transformation-based learning](#). In *Third Workshop on Very Large Corpora*.
- Jesse H. Rhodes and Kaylee Johnson. 2015. [The Politics of Group Targeting in Presidential Campaign Advertisements: A Preliminary Investigation](#).

- Joshua Robison, Rune Stubager, Mads Thau, and James Tilley. 2021. [Does Class-Based Campaigning Work? How Working Class Appeals Attract and Polarize Voters.](#) *Comparative Political Studies*, 54(5):723–752. Publisher: SAGE Publications Inc.
- Simon Stuckelberger and Anke Tresch. 2024. [Group Appeals of Parties in Times of Economic and Identity Conflicts and Realignment.](#) *Political Studies*, 72(2):463–485. Publisher: SAGE Publications Ltd.
- Mads Thau. 2019a. [How political parties use group-based appeals: Evidence from Britain 1964–2015.](#) *Political Studies*, 67(1):63–82.
- Michael Thau. 2019b. [How political parties use group-based appeals: Evidence from Britain 1964–2015.](#) *Political Studies*, 67(1):63–82.
- Michael Thau. 2024. [The group appeal strategy: Beyond the policy perspective on party electoral success.](#) *Political Studies*, 72(4):1653–1674.
- Jeffrey S. Vitter. 1985. [Random sampling with a reservoir.](#) *ACM Trans. Math. Softw.*, 11(1):37–57.
- C. Weber and M. Thornton. 2012. [Courting Christians: How political candidates prime religious considerations in campaign ads.](#) *The Journal of Politics*, 74(2):400–413.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.](#) *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddhartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2024. [LiveBench: A Challenging, Contamination-Limited LLM Benchmark.](#) In *The 13th International Conference on Learning Representations*, ICLR.
- I. K. White. 2007. [When race matters and when it doesn't: Racial group differences in response to racial cues.](#) *American Political Science Review*, 101(2):339–354.
- Christopher Wlezien and Arthur H. Miller. 1997. [Social Groups and Political Judgments.](#) *Social Science Quarterly*, 78(3):625–640. Publisher: [University of Texas Press, Wiley].
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [Transformers: State-of-the-art natural language processing.](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2025. [ExpertPrompting: Instructing Large Language Models to be Distinguished Experts.](#) ArXiv:2305.14688 [cs].
- Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. [Sentiment Analysis in the Era of Large Language Models: A Reality Check.](#) ArXiv:2305.15005 [cs].

Supplementary Material

A. Appendix

A.1. Preprocessing and sampling

Table 6 shows the distribution of group tags in the sampled instances used as input for the annotation process. The full label names can be found in Table 7 below.

A.2. BERT Baseline Hyperparameters

Hyperparameter	Value
Learning rate	5×10^{-6}
Optimizer	AdamW
Weight decay	0.01
Warmup ratio	0.1
Dropout	0.2
Batch size	8
Number of epochs	10

Table 5: Unspecified hyperparameters are set to the default values as provided by the transformers library (Wolf et al., 2020). Dropout is applied to both attention and MLP layers.

Table 5 reports the hyperparameters used for fine-tuning our BERT-baseline models. We found these hyperparameters to work similarly well for GBERT and ModernGBERT. All models are evaluated using the best checkpoint according to the evaluation loss.

A.3. Prompt design

This section details the prompt templates used for generating stance predictions with large language models. Placeholders for dynamic content are shown as `{placeholder}`. Figures 2, 3 show the “split-prompt” and “full-prompt” structure and figures 4, 5 illustrate the difference between Chain-of-Thought (CoT) prompting and zero-shot prompting.

Finally, Figure 7 shows the step-by-step reasoning output for example 2 in the paper.

Label	Size	Label	Size	Label	Size
PFUNC	125	EPOL	34	EPMOV	11
PAGE	111	EOFIN	33	EPMED	11
PGEN	105	EPSCI	29	EPMIL	11
PNAT	101	EONGO	25	EPNGO	10
PSOC	69	EOMED	25	EPECON	10
EOECON	64	EOSCI	17	EPREL	10
PETH	54	EOMOV	16	EPFIN	10
EOMIL	52	EPCULT	13		
EOPOL	42	EOREL	11		

Table 6: Distribution of group labels in the data sampled for annotation.

Label	Description	Example
<i>People (P: People)</i>		
PAGE	People by Age	adolescents, the elderly
PETH	People by Ethnicity/Religion	Muslims, the Kurd
PFUNC	People by Function	farmers, tax payers, parents
PGEN	Generic mentions of the People	
PNAT	People by Nation	the Germans, the Chinese
PSOC	People by Social Variables	women, the working class
<i>Individual instances of the Elite (EP: Elite, Person)</i>		
EPCULT	Culture	Bob Marley, Selma Lagerlöf
EPECON	Economy	the CEO of Siemens
EPFIN	Finance	the chairman of Goldman Sachs
EPMED	Media	Mark Zuckerberg
EPMIL	Military	the general
EPMOV	Movements	Greta Thunberg
EPNGO	NGOs	Representatives of migrant organisations
EPOL	Politics	the chancellor, prime minister
EPREL	Religion	the Pope, the Dalai Lama
EPSCI	Science	Marie Curie
<i>Groups and organisations of the Elite (EO: Elite, Organisation)</i>		
EOCULT	Culture	the ESC, Metropolitan Opera
EOECON	Economy	Amazon, Wirecard
EOFIN	Finance	European Central Bank
EOMED	Media	Twitter, the NYT
EOMIL	Military	US Air Force
EOMOV	Movements	Fridays for Future
EONGO	NGOs	Doctors without Borders, Oxfam
EOPOL	Politics	Labour Party, Sinn Féin
EOREL	Religion	World Buddhist Forum, Jews for Jesus
EOSCI	Science	Science Europe, CERN

Table 7: Full label names for social groups included in our data.

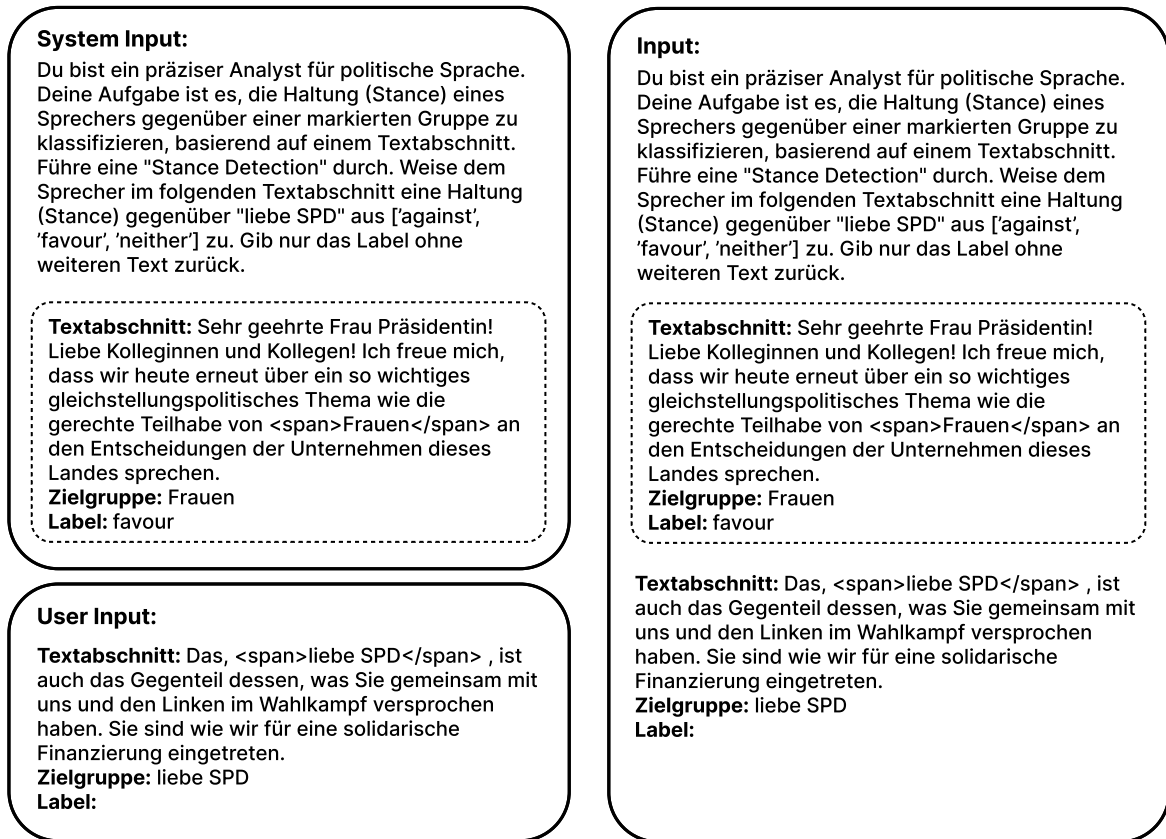


Figure 2: The left prompt shows an example for the “split-prompt” structure, based on the ‘zero-shot’ prompt template and the right prompt illustrates the “full-prompt” structure. The dashed box indicates where additional few-shot examples are placed in the few-shot setting. For the English translation, see Fig. 3 below.

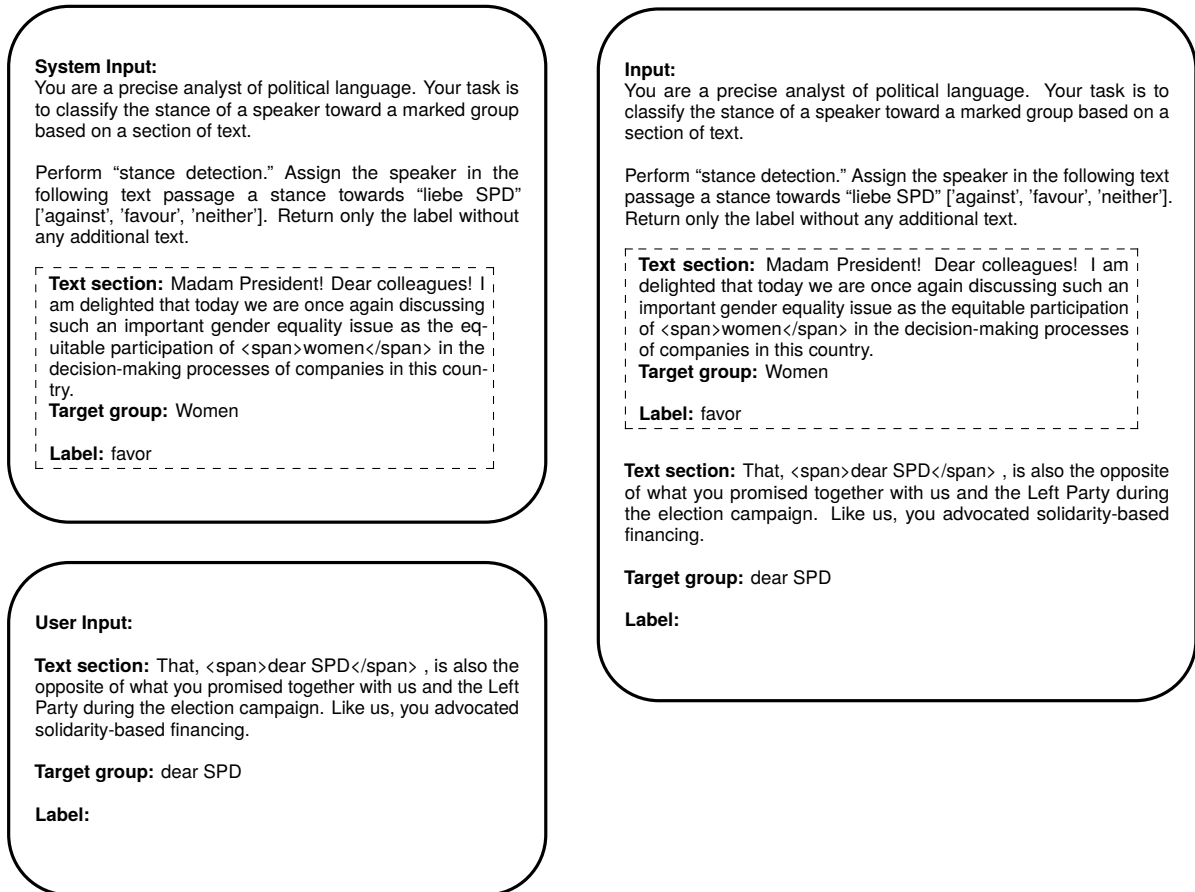


Figure 3: English translation of Fig. 2 above. The left prompt shows an example for the “split-prompt” structure, based on the ‘zero-shot’ prompt template and the right prompt illustrates the “full-prompt” structure. The dashed box indicates where additional few-shot examples are placed in the few-shot setting.

System Input:

****Rolle und Ziel:****
 Du agierst als streng unparteiischer Analyst für politische Sprache. Deine Aufgabe ist es, die Haltung (Stance) eines Sprechers gegenüber einer spezifisch markierten Gruppe zu klassifizieren.

****Grundprinzip der Analyse:****
 Deine Analyse muss sich ****ausschließlich**** auf die expliziten Aussagen im vorgelegten Textabschnitt stützen. Die Haltung muss ****direkt aus dem Wortlaut**** ableitbar sein. Interpretiere oder schlussfolgere nicht über den Text hinaus. Insbesondere bei Ambiguität oder Unklarheit ist höchste Vorsicht geboten.

****Anweisungen:****

1. Lies den bereitgestellten ****Textabschnitt**** und identifiziere die markierte ****Gruppe****.
2. Analysiere die Haltung des Sprechers ****ausschließlich**** basierend auf den Worten, die er in diesem Abschnitt verwendet.
3. Wähle eine der folgenden drei Kategorien.

****Definition der Kategorien:****

- * **``favour``**: Wähle diese Kategorie nur, wenn der Sprecher sich ****eindeutig und direkt**** positiv, unterstützend oder wohlwollend gegenüber der Gruppe äußert. Die positive Haltung muss unmissverständlich im Text formuliert sein.
- * **``against``**: Wähle diese Kategorie nur, wenn der Sprecher sich ****eindeutig und direkt**** negativ, kritisch oder ablehnend gegenüber der Gruppe äußert. Die negative Haltung muss unmissverständlich im Text formuliert sein.
- * **``neither``**: Dies ist die ****Standardkategorie****. Wähle sie, wenn die Gruppe nur neutral erwähnt wird ODER wenn die Haltung des Sprechers ambivalent, unklar oder nicht eindeutig aus dem Text bestimmbar ist. Im Zweifelsfall, wenn die Kriterien für `favour` oder `against` nicht ****zweifelsfrei**** erfüllt sind, wähle ****immer**** `neither`.

****Anforderung an die Antwort:****

Bitte denke Schritt für Schritt nach und gebe im Anschluss deine Label in dem Format <stance>label</stance> aus.

System Input:

****Rolle und Ziel:****
 Du agierst als streng unparteiischer Analyst für politische Sprache. Deine Aufgabe ist es, die Haltung (Stance) eines Sprechers gegenüber einer spezifisch markierten Gruppe zu klassifizieren.

****Grundprinzip der Analyse:****
 Deine Analyse muss sich ****ausschließlich**** auf die expliziten Aussagen im vorgelegten Textabschnitt stützen. Die Haltung muss ****direkt aus dem Wortlaut**** ableitbar sein. Interpretiere oder schlussfolgere nicht über den Text hinaus. Insbesondere bei Ambiguität oder Unklarheit ist höchste Vorsicht geboten.

****Anweisungen:****

1. Lies den bereitgestellten ****Textabschnitt**** und identifiziere die markierte ****Gruppe****.
2. Analysiere die Haltung des Sprechers ****ausschließlich**** basierend auf den Worten, die er in diesem Abschnitt verwendet.
3. Wähle eine der folgenden drei Kategorien.

****Definition der Kategorien:****

- * **``favour``**: Wähle diese Kategorie nur, wenn der Sprecher sich ****eindeutig und direkt**** positiv, unterstützend oder wohlwollend gegenüber der Gruppe äußert. Die positive Haltung muss unmissverständlich im Text formuliert sein.
- * **``against``**: Wähle diese Kategorie nur, wenn der Sprecher sich ****eindeutig und direkt**** negativ, kritisch oder ablehnend gegenüber der Gruppe äußert. Die negative Haltung muss unmissverständlich im Text formuliert sein.
- * **``neither``**: Dies ist die ****Standardkategorie****. Wähle sie, wenn die Gruppe nur neutral erwähnt wird ODER wenn die Haltung des Sprechers ambivalent, unklar oder nicht eindeutig aus dem Text bestimmbar ist. Im Zweifelsfall, wenn die Kriterien für `favour` oder `against` nicht ****zweifelsfrei**** erfüllt sind, wähle ****immer**** `neither`.

****Anforderung an die Antwort:****

Deine Antwort muss **ausschließlich eines der drei folgenden Wörter enthalten, ohne zusätzliche Erklärungen, Begrüßungen oder Satzzeichen:**

`favour`
 `against`
 `neither`

User Input:

Textabschnitt: Das, liebe SPD , ist auch das Gegenteil dessen, was Sie gemeinsam mit uns und den Linken im Wahlkampf versprochen haben. Sie sind wie wir für eine solidarische Finanzierung eingetreten.

Zielgruppe: liebe SPD

Label:

User Input:

Textabschnitt: Das, liebe SPD , ist auch das Gegenteil dessen, was Sie gemeinsam mit uns und den Linken im Wahlkampf versprochen haben. Sie sind wie wir für eine solidarische Finanzierung eingetreten.

Zielgruppe: liebe SPD

Label:

Figure 4: An example of a Chain-of-Thought (CoT) prompt (left) compared to a standard zero-shot prompt (right). The CoT version adds an instruction for the model to “think step by step”, while the standard prompt asks for a direct classification. The prompt template uses the split-prompt setup. The English translation is shown in Fig. 5 below.

System Input:
****Role and objective:****
 You act as a strictly impartial analyst of political language. Your task is to classify the stance of a speaker toward a specific, marked group.
****Basic principle of analysis:****
 Your analysis must be based ****exclusively**** on the explicit statements in the text passage provided. The stance must be ****directly derivable from the wording****. Do not interpret or draw conclusions beyond the text. Particular caution is required in cases of ambiguity or uncertainty.
****Instructions:****

1. Read the ****text passage**** provided and identify the highlighted ****group****.
2. Analyze the speaker's attitude ****exclusively**** based on the words they use in this passage.
3. Choose one of the following three categories.

****Definition of categories:****
****favour****: Select this category only if the speaker expresses ****clear and direct**** positive, supportive, or benevolent sentiments toward the group. The positive attitude must be unambiguously stated in the text.
****against****: Only select this category if the speaker expresses a ****clear and direct**** negative, critical, or dismissive attitude toward the group. The negative attitude must be unambiguously stated in the text.
****neither****: This is the ****default category****. Select it if the group is only mentioned neutrally OR if the speaker's attitude is ambivalent, unclear, or cannot be unequivocally determined from the text. In case of doubt, if the criteria for 'favour' or 'against' are not ****unequivocally**** met, ****always**** select 'neither'.
****Requirements for the answer:****

Please think step by step and then enter your label in the format <stance>label</stance>.

System Input:
****Role and objective:****
 You act as a strictly impartial analyst of political language. Your task is to classify the stance of a speaker toward a specific, marked group.
****Basic principle of analysis:****
 Your analysis must be based ****exclusively**** on the explicit statements in the text passage provided. The stance must be ****directly derivable from the wording****. Do not interpret or draw conclusions beyond the text. Particular caution is required in cases of ambiguity or uncertainty.
****Instructions:****

1. Read the ****text passage**** provided and identify the highlighted ****group****.
2. Analyze the speaker's attitude ****exclusively**** based on the words they use in this passage.
3. Choose one of the following three categories.

****Definition of categories:****
****favour****: Select this category only if the speaker expresses ****clear and direct**** positive, supportive, or benevolent sentiments toward the group. The positive attitude must be unambiguously stated in the text.
****against****: Only select this category if the speaker expresses a ****clear and direct**** negative, critical, or dismissive attitude toward the group. The negative attitude must be unambiguously stated in the text.
****neither****: This is the ****default category****. Select it if the group is only mentioned neutrally OR if the speaker's attitude is ambivalent, unclear, or cannot be unequivocally determined from the text. In case of doubt, if the criteria for 'favour' or 'against' are not ****unequivocally**** met, ****always**** select 'neither'.
****Requirements for the answer:****

Your answer must contain only one of the following three words, without any additional explanations, greetings, or punctuation marks:
 'favour'
 'against'
 'neither'

User Input:
Text section: That, dear SPD, is also the opposite of what you promised together with us and the Left Party during the election campaign. Like us, you advocated for solidarity-based financing.
Target group: dear SPD
Label:

User Input:
Text section: That, dear SPD, is also the opposite of what you promised together with us and the Left Party during the election campaign. Like us, you advocated for solidarity-based financing.
Target group: Dear SPD
Label:

Figure 5: English translation of Fig. 4 above. The figure shows an example of a Chain-of-Thought (CoT) prompt (left) compared to a standard zero-shot prompt (right). The CoT version adds an instruction for the model to "think step by step", while the standard prompt asks for a direct classification. The prompt template uses the split-prompt setup.

Prompt: CoT (open-weight models, German)

Du bist ein präziser Analyst für politische Sprache. Deine Aufgabe ist es, die Haltung (Stance) eines Sprechers gegenüber einer markierten Zielgruppe zu klassifizieren, basierend auf einem Textabschnitt.

Führe eine "Stance Detection" durch. Weise dem Sprecher im folgenden Textabschnitt eine Haltung (Stance) gegenüber "group" aus ['against', 'favour', 'neither'] zu.

Definition der Kategorien:

- favour: Die Haltung ist eindeutig und direkt positiv. Der Sprecher lobt die Gruppe, verteidigt sie, fordert etwas zu ihren Gunsten oder gibt explizit an, für ihre Interessen einzutreten (z.B. "wir kämpfen für diese Gruppe").
- against: Die Haltung ist eindeutig und direkt negativ. Der Sprecher kritisiert, verurteilt oder warnt vor der Gruppe oder macht sie für ein Problem verantwortlich.
- neither: Dies ist die Standardkategorie im Zweifelsfall. Wähle sie, wenn die Haltung neutral, ambivalent oder unklar ist, oder wenn der Sprecher die Gruppe nur sachlich erwähnt.

Bitte denke Schritt für Schritt nach und gebe im Anschluss deine Label in dem Format <stance>label</stance> aus.

Prompt: CoT (open-weight models, English translation)

You are a precise analyst of political language. Your task is to classify the stance of a speaker toward a marked target group based on a text passage.

Perform a "stance detection." Assign the speaker in the following text passage a stance toward "group" ["against", "favour", "neither"].

Definition of categories:

- favour: The stance is clearly and directly positive. The speaker praises the group, defends it, demands something in its favor, or explicitly states that they are advocating for its interests (e.g., "we are fighting for this group").
- against: The stance is clearly and directly negative. The speaker criticizes, condemns, or warns against the group, or blames it for a problem.
- neither: This is the default category when in doubt. Choose it if the stance is neutral, ambivalent, or unclear, or if the speaker only mentions the group in a factual manner.

Please think step by step and then output your labels in the format <stance>label</stance>.

Figure 6: CoT prompt used with GEMMA-3, LLAMA-3.1-8B and LLAMA-3.3-70B (German original prompt and English translation).

German (original model output):

Um die Haltung des Sprechers gegenüber "die Frankfurter Allgemeine Zeitung" zu bestimmen, analysieren wir den Textabschnitt Schritt für Schritt:

1. **Kontext**: Der Sprecher diskutiert über Fehler, die gemacht wurden, und die Notwendigkeit, diese zu korrigieren und in Zukunft zu vermeiden. Er kritisiert eine Gruppe (nicht explizit "die Frankfurter Allgemeine Zeitung", sondern eine andere Gruppe, die für den Haushaltsplan verantwortlich ist) für das Ignorieren dieser Notwendigkeit.
2. **Erwähnung der "Frankfurter Allgemeine Zeitung"**: Die Zeitung wird erwähnt, als der Sprecher sagt, dass der Haushaltsplan "der beste Beweis dafür ist, dass Sie nichts gelernt haben. Sie setzen die Einnahmen weiterhin zu hoch und die Ausgaben zu niedrig an, um so formal zu türken und zu täuschen, wie es die 'Frankfurter Allgemeine Zeitung' gesagt hat."
3. **Haltung**: Die Erwähnung der "Frankfurter Allgemeine Zeitung" erfolgt in einem Kontext, in dem der Sprecher eine kritische Aussage über eine andere Gruppe macht. Die "Frankfurter Allgemeine Zeitung" wird als Quelle zitiert, die diese Kritik unterstützt, indem sie die Praxis des "Türken und Täuschens" durch die fragliche Gruppe anspricht.
4. **Interpretation**: Der Sprecher zitiert die "Frankfurter Allgemeine Zeitung" in einem positiven Licht, da er ihre Aussage als Beweis für seine Kritik verwendet. Dies bedeutet, dass der Sprecher die Aussage der Zeitung als gültig und unterstützend für seine Argumentation ansieht.

Basierend auf dieser Analyse kann die Haltung des Sprechers gegenüber "die Frankfurter Allgemeine Zeitung" als positiv eingestuft werden, da er ihre Aussage als Beweis und Unterstützung für seine Kritik an einer anderen Gruppe verwendet.

<stance>favour</stance>

English translation:

To determine the speaker's attitude toward "die Frankfurter Allgemeine Zeitung," we will analyze the text passage step by step:

1. **Context**: The speaker discusses mistakes that have been made and the need to correct them and avoid them in the future. He criticizes a group (not explicitly "the Frankfurter Allgemeine Zeitung," but another group responsible for the budget) for ignoring this necessity.
2. **Mention of the "Frankfurter Allgemeine Zeitung"**: The newspaper is mentioned when the speaker says that the budget "is the best proof that you have learned nothing. You continue to set revenues too high and expenditures too low in order to formally deceive and mislead, as the Frankfurter Allgemeine Zeitung has said."
3. **Attitude**: The mention of the Frankfurter Allgemeine Zeitung occurs in a context in which the speaker makes a critical statement about another group. The Frankfurter Allgemeine Zeitung is cited as a source that supports this criticism by addressing the practice of "fudging and deceiving" by the group in question.
4. **Interpretation**: The speaker quotes the Frankfurter Allgemeine Zeitung in a positive light, as he uses its statement as evidence for his criticism. This means that the speaker considers the newspaper's statement to be valid and supportive of his argument.

Based on this analysis, the speaker's stance toward the Frankfurter Allgemeine Zeitung can be classified as positive, as he uses its statement as evidence and support for his criticism of another group.

Figure 7: Step-by-step reasoning output for Example 2 (model: Llama-3.3-70B with CoT prompting).