Using Generative AI to Calculate Party Positions

A Comparison of Human Experts and Large Language Models

Cantay Caliskan, Junhua Huang, Yiyang Huang, Ruoxuan Lin, and Wanting Shan

University of Rochester

July 26, 2024

Abstract

The spatial theory of voting, first introduced by Downs (1957), posits that voters choose political parties based on their positions on the political spectrum, from left to right. Traditionally, left-leaning voters favor progressive policies and government intervention, while right-leaning voters support conservative values and free markets. Centrists seek a balance. This framework helps voters find parties that align with their beliefs and preferences. One important tool that voters may use to decide spatially is the Manifesto Project, which analyzes political parties' election manifestos to study their policy preferences and estimate their positions. This paper investigates the feasibility of replacing human intervention in estimating party positions with intelligent tools, such as LLMs. By examining the party positions estimated by five different LLMs (ChatGPT 3.5 Turbo, Cohere Command, Gemini 1, Llama 2, and Llama 3), our work provides a renewed look at party positions in three major democracies with two-party and multi-party systems (Germany, the United Kingdom, and the United States) and discusses the feasibility of employing AI tools to assist human experts. Results show that, on average, LLMs perceive leftist and rightist manifestos as 73.64% less extreme than they truly are. Given that CMP-trained human experts are likely performing their tasks accurately, this significant bias implies that replacing human experts with LLMs for the calculation of party positions may not be feasible in the near future.

Keywords: Generative AI, Ideological Bias, Party Politics, Party Systems, LLMs in Politics, Political Methodology

"Every election is determined by the people who show up."

– Larry Sabato

"The most important political office is that of the private citizen."

– Louis D. Brandeis

1 Introduction

Citizens of the democratic world are at another important juncture in history. On one hand, there is the global erosion of democratic values, as seen in countries like Hungary and Turkey, and voter misorientation by misinformation, such as during the Brexit referendum and the 2016 US elections. On the other hand, LLMs are progressing swiftly, with breakthroughs in natural language processing allowing them to comprehend and produce text, potentially offering an opportunity for better-informed citizens and more open democracies. This paper aims to test the capabilities of LLM models in making democracies more transparent by providing a reproducible evaluation of party lines and helping voters make more informed decisions in elections. Specifically, the paper quantifies and compares the RILE (Right-Left) Index[1] values produced by human experts and LLMs (i) and provides a time and cost comparison (ii) between human coders and LLMs to understand the effectiveness of generative AI in evaluating party positions.

Understanding the political position of a party is crucial for voters as it empowers them to make informed decisions that align with their values and interests. Knowledge of party positions helps voters predict policy outcomes and identify which parties best represent their views on issues like the economy, healthcare, and education [15]. The Manifesto Project aids in this by analyzing election manifestos to estimate party positions, thus enhancing voter awareness. This awareness fosters a more engaged and educated electorate, essential for strengthening democracies. Well-informed voters hold parties accountable, requiring clear

[1]Political scientists have designed different approaches to measure ideology, such as spatial model for roll call voting DW-Nominate scores, campaign finance-based ideology CFscores, CMP analysis Manifesto Research Project and Bayesian Ideal Points Bridge Ideal Points [9, 25, 36, 5, 44]. The RILE Index is constructed by analyzing the content of party manifestos and coding specific policy statements into predefined categories that represent left-wing or right-wing stances.

articulation and fulfillment of policy promises [29]. By understanding party positions through resources like the Manifesto Project, voters contribute to a transparent and responsive political system.

In addition, collecting data in political science is a labor-intensive process involving extensive fieldwork, meticulous archival research, and large-scale surveys. Projects like the Manifesto Project require recruiting and training coders from multiple countries to ensure reliable data [49]. Similarly, creating datasets for voting behavior and public opinion surveys involves designing instruments, training interviewers, and ensuring representative samples [7]. Qualitative data collection, such as conducting interviews and focus groups, demands substantial time for establishing rapport, gaining access, and transcribing responses [34]. These tasks highlight the significant labor investment required for rigorous data collection in political science. Additionally, the field often faces a lack of funding, which can limit the scope and quality of research efforts, making efficient data collection even more challenging [28].

Previous studies have shown that transformer-based machine learning models, including LLMs, can mimic human-like responses and behaviors, making them adaptable to social science research [19]. Notably, LLMs have produced authentic-looking survey answers, suggesting their potential to replace human respondents in data collection. Although there are concerns about LLMs' accuracy in analyzing complex languages, recent studies show that with proper training and fine-tuning, LLMs can accurately mimic human behavior in social science research [3]. Some studies have shown that GPT-4 outperforms both experts and crowd workers in annotating political Twitter messages, showcasing superior efficiency and accuracy in political content analysis [47].

In previous literature, researchers have used pre-trained LLMs, such as ConfliBERT and GPT-3, to analyze political issues quantitatively, including political conflict and violence against specific groups [21, 1]. In text analysis and interaction model construction, LLMs serve as effective proxies [50, 3, 10]. Though their application is limited by political bias [4, 40, 39], LLMs with sufficient algorithmic fidelity are still regarded as powerful tools in political science research.

We expect that our calculations will reveal some degree of political bias. LLMs are not free

from biases, which stem from the data they are trained on, embedding societal prejudices and skewed representations. Consequently, LLM outputs can perpetuate or amplify these biases, affecting their use in political science research [8]. While LLMs can efficiently generate human-like text and analyze large datasets, they may reflect unobserved biases from their training data. Despite their potential to address challenges in survey research and field studies, such as question wording and response biases, their inherent biases must be carefully managed to ensure valid and fair applications [22].

Literature suggests that LLMs tend to avoid producing politically extremist sentences due to several key factors. Developers implement advanced safety protocols and content moderation filters to recognize and mitigate harmful content, such as OpenAI's filters [41]. LLMs are also fine-tuned using reinforcement learning from human feedback (RLHF), guiding models towards neutral outputs [13]. This process helps avoid sensitive topics. These models adhere to user instructions and contextual cues, promoting moderate responses [11]. Interestingly, LLMs sometimes perceive extreme sentences as more centrist due to biases in their training data and their tendency to generalize content. This occurs because the training data averages out extreme views, making them appear more typical [6]. Additionally, the diverse nature of internet data can dilute the extremity of certain views, making them seem less radical [43]. Consequently, LLMs are cautious in generating politically charged content, aligning with ethical standards and maintaining user trust [6].

In addition, two-party systems are characterized by limited political competition and fewer ideological choices, whereas multi-party systems generally exhibit a greater diversity of political ideologies. In such systems, political competition extends beyond two dominant parties, allowing multiple parties to vie for power [16, 27]. This structure facilitates the representation of a wider range of ideological perspectives, including those outside the mainstream (Norris, 2004). Consequently, voters in multi-party systems have more choices that reflect their specific political beliefs and values [42]. This diversity can lead to more nuanced policymaking but often results in coalition governments and complex political negotiations [17].

Based on the theoretical background and the current literature, this paper aims to investigate the following two sets of hypotheses:

H1: ***Political ideology scores (RILE) calculated by human experts and LLMs will have measurable differences:***

H1a: Manifestos authored by political parties with extremist views will be interpreted as more centrist than their actual positions.

H1b: Manifestos authored by political parties with centrist views will be classified more accurately by LLMs.

H2: ***The accuracy of classifications by LLMs will exhibit measurable differences between two-party and multi-party systems:***

H2a: LLMs will provide a more accurate understanding of party positions in two-party democracies.

H2b: LLMs will introduce stronger biases in the understanding of party manifestos in multi-party democracies.

2 Data and Methodology

Our paper uses the coded party manifestos compiled by the Comparative Manifesto Project (CMP).[2] A select group of manifestos obtained from the CMP Project were then used to calculate the RILE Index. Eventually, we compared the ground truth (party positions computed by human experts) to predicted party positions. More information about our data selection, the key variable (RILE Index), and our empirical strategy is presented in the later sections. Below is a concise demonstration of our data and methods pipeline.

[2]The Comparative Manifesto Project (CMP) is a widely utilized research source that systematically collects, analyzes, and compares parties' positions on the left-right spectrum and other ideological and policy dimensions [18]. Due to its comprehensive and standardized dataset, CMP facilitates comparative political research and is widely used in studies involving electoral dynamics and the evolution of political ideologies over time [49]. The entire CMP Project covers 67 countries, includes data from 849 elections, and encompasses 1,373 political parties.

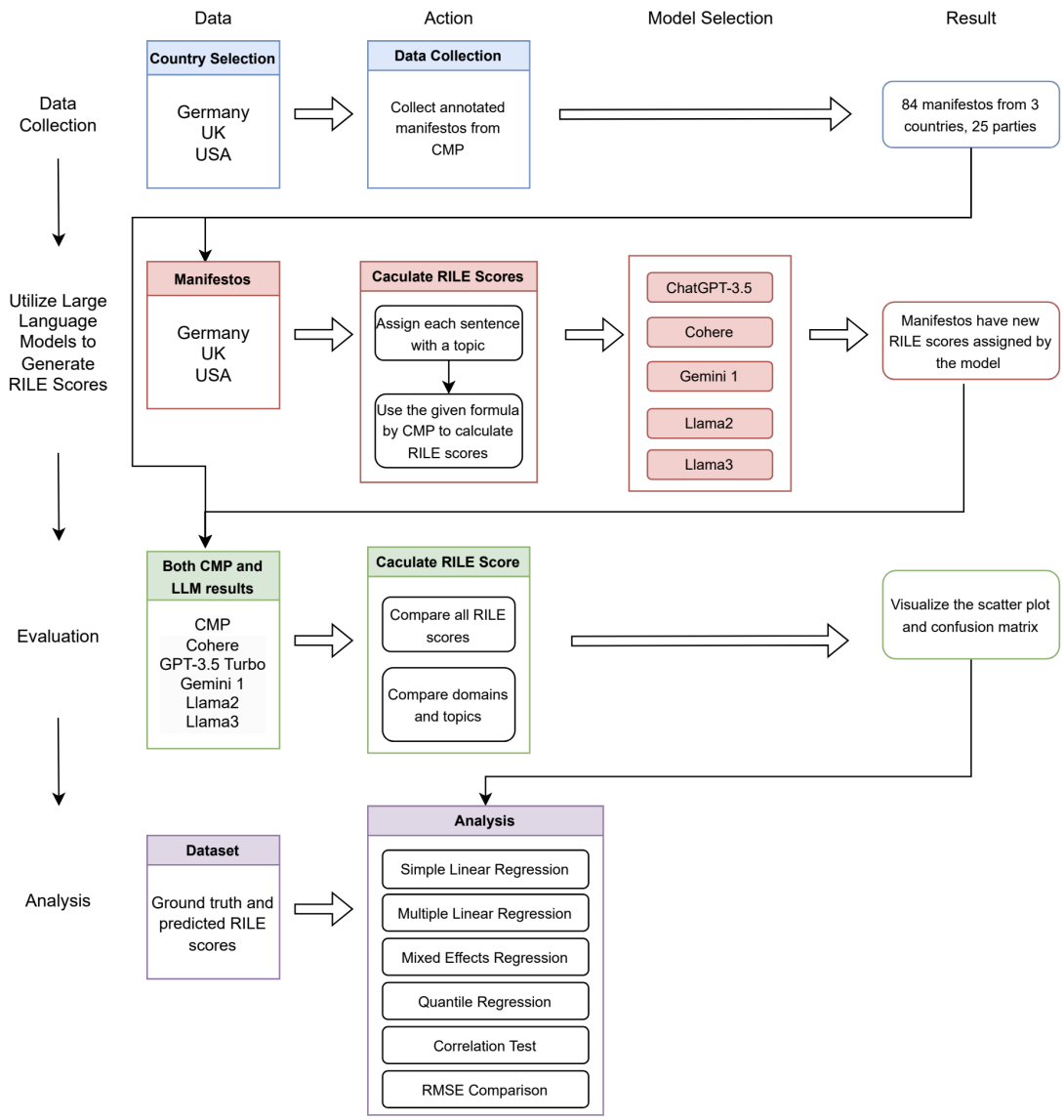


Figure 1: Data and Methods Pipeline

2.1 RILE Index

The RILE (Right-Left) index is a quantitative measure used to determine the ideological position of political parties based on their manifestos. Developed as part of the Comparative Manifesto Project (CMP), the RILE index is constructed by coding policy statements within party manifestos into predefined categories that represent left-wing or right-wing stances [12, 24]. The index is calculated using the following steps:

1. **Coding Policy Statements**: Each sentence or quasi-sentence in a party's manifesto

is coded into one of 56 predefined categories. These categories cover a wide range of policy areas, such as the economy, social policies, foreign affairs, and more [12].

2. **Aggregation of Codes**: The categories are grouped into left-wing and right-wing clusters. For example, statements supporting social welfare and government intervention are coded as left-wing, while statements favoring free markets and military strength are coded as right-wing [26].

3. **Index Calculation**: The RILE score is calculated by subtracting the sum of the percentages of left-wing categories from the sum of the percentages of right-wing categories. The formula can be expressed as:

RILE = (Sum of Right-Wing Percentages) - (Sum of Left-Wing Percentages)

A positive RILE score indicates a right-wing orientation, while a negative score indicates a left-wing orientation. A score around zero suggests a centrist position [12].[3]

The RILE index provides a standardized method for comparing party positions across different countries and over time, facilitating the analysis of ideological shifts and trends in political landscapes [24]. This index is widely used in political science research to study party competition, voter behavior, and policy changes [12].

2.2 Case Selection

Our sample data includes the party manifestos from the United States, Germany, and the United Kingdom. We followed several criteria for this selection: 1) These countries are

[3]The RILE index, part of the Comparative Manifesto Project (CMP), categorizes policy statements into left-wing and right-wing clusters. The left-wing categories include: Per101 - Foreign Special Relationships: Positive, Per103 - Anti-Imperialism, Per105 - Military: Negative, Per106 - Peace, Per107 - Internationalism: Positive, Per201 - Freedom and Human Rights, Per202 - Democracy, Per403 - Market Regulation, Per404 - Economic Planning, Per406 - Protectionism: Positive, Per412 - Controlled Economy, Per413 - Nationalisation, Per503 - Equality: Positive, Per504 - Welfare State Expansion, Per506 - Education Expansion, and Per701 - Labour Groups: Positive. The right-wing categories include: Per104 - Military: Positive, Per201 - Freedom and Human Rights, Per305 - Political Authority, Per401 - Free Market Economy, Per402 - Incentives, Per407 - Protectionism: Negative, Per414 - Economic Orthodoxy, Per505 - Welfare State Limitation, Per601 - National Way of Life: Positive, Per603 - Traditional Morality: Positive, Per605 - Law and Order, Per606 - Social Harmony, and Per608 - Multiculturalism: Negative. Left-wing categories are associated with social welfare, government intervention, international cooperation, and equality, while right-wing categories emphasize free market principles, national strength, traditional values, and limited government intervention.

widely studied, which helps avoid extreme cases or outliers and ensures the generalizability of our findings. 2) Each has a long history of democracy, as evidenced by their Freedom House Scores, being continuously labeled as *free* from 1973 to 2023. 3) These countries have a clearly defined right-left political spectrum, which is crucial for our analysis. Additionally, the political parties in these countries have established their positions firmly, particularly over the last 50 years, providing a stable and clear ideological landscape for our study. This solidification can be attributed to significant political and social changes in the post-1970s era, including economic shifts, societal movements, and the evolution of party systems that have reinforced distinct ideological identities.

The sample we have captures different continents, representing countries with varied geopolitical situations and political ecologies, such as a two-party system in the United States versus multi-party systems in Germany and the United Kingdom. This diversity enhances the robustness of our study by incorporating different democratic structures and electoral systems. Established research [16, 27] provides substantial insights into the functioning and implications of two-party and multi-party systems, respectively, illustrating the differences in political dynamics and voter representation in these contexts.

Furthermore, there is extensive literature analyzing democracy and political systems in these countries, providing a rich context for our study. For example, works by Almond and Verba (1963) on civic culture [2], by Ljiphart (1999) on the conditions of democracy [27], and more recent studies by Merkel (2014) on democratic quality [32] have significantly shaped our understanding of democratic practices in these countries. The period since the 1970s has seen the entrenchment of neoliberal economic policies in the United States and the United Kingdom, the consolidation of social market economy principles in Germany, and the overall stabilization of party systems that reflect these ideological commitments [20, 23].

Importantly, the selected countries are considered mainstream subjects in political science research. There is a substantial body of literature using these three countries as case studies for various fields, including text-based measurements of populism in party manifestos, voter behavior, and analysis of party campaign messages [14, 35, 30, 33]. This extensive existing research provides a solid foundation and context for our analysis, ensuring that our findings are well-grounded and relevant to ongoing scholarly discussions.

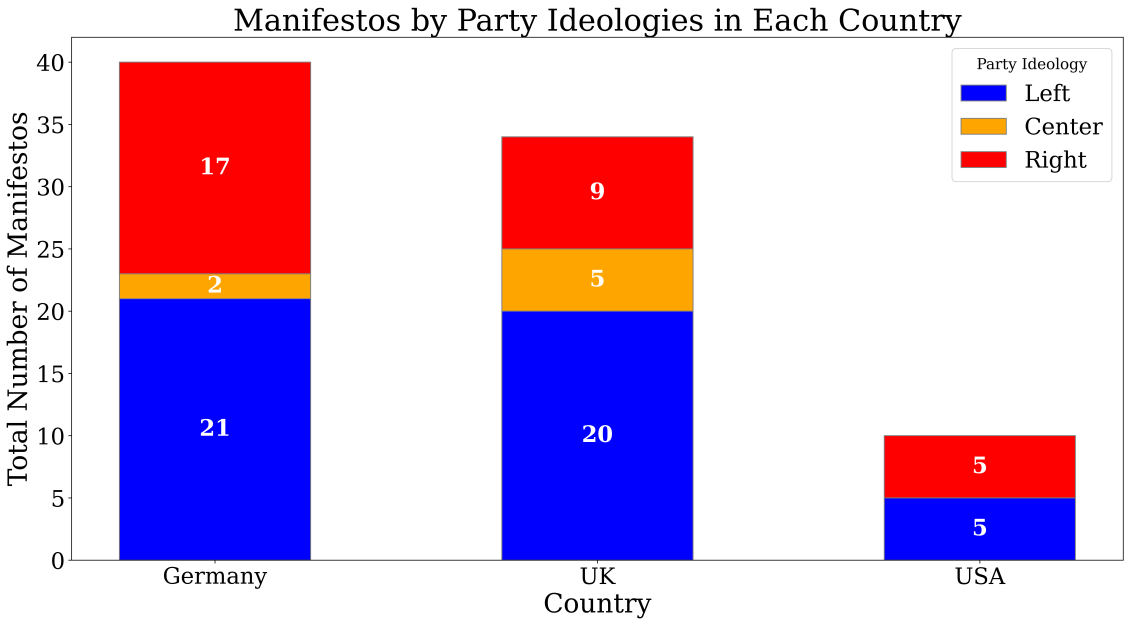


Figure 2: Number of Manifestos per Political Position

2.3 Data Collection

To calculate the RILE indices, we ran each coded segment through one of the LLMs. The Comparative Manifesto Project (CMP) structures party manifestos by dividing them into quasi-sentences, which are the smallest units of meaning. Each quasi-sentence is assigned a specific code from the CMP's detailed coding scheme, categorizing statements into various policy domains, such as economic, social, and foreign policy. The coding allows for the calculation of the RILE (Right-Left Index) score, which aggregates coded statements to indicate the overall left-right position of a party. This standardized coding process ensures consistency and allows for systematic comparison of manifestos across different parties, countries, and time periods. The data collection was completed in April 2024. In total, 84 manifestos containing a total of 140,432 text segments were collected.

To assign a code to each quasi-sentence, the entire coding scheme provided by CMP was fed into the LLMs. This prompt engineering approach differs greatly from another valuable work in the literature [31], which only uses a single-issue-based classification approach or forces LLMs to take a position on politically salient issues. We believe that, despite using more financial and computational resources, our approach leads to a more genuine and com-

plete understanding of how some popular LLMs work. We further want to emphasize that our goal in this article is not necessarily to perform prompt engineering techniques to increase the accuracy of LLMs—we are interested in comparing how human experts and LLMs react to the same set of guidelines (which, as noted, is not investigated in previous research).

To obtain left-wing or right-wing categories associated with each quasi-sentence, we submitted a prompt that contained a 'user message' instructing the LLM to return such a category.[4] An example is illustrated here:

> *You are an expert in political science. You are asked to match sentences from a political party's manifesto with a topic. Here is a sentence from the manifesto:* ***[Quasi-sentence from the party manifesto]****. All topics and their explanations are provided below. Only provide one topic that best fits the sentence from the manifesto. Return only one topic, and nothing else.* ***[Description of all topics and domains found in the CMP documentation.]***[5]

For all the models, we set the temperature parameter to 0 to ensure the LLM generated responses by selecting the most likely next token, making the outputs as deterministic as possible, which may aid in replication studies. All other parameters were left at their default settings, with no additional modifications.

[4] In an overwhelming majority of the calls, the LLMs returned a CMP category that was provided or an 'NA' value ('NA' is the most common category assigned by human experts as well). In some rare cases, a non-existing category was assigned, which we eventually converted into 'NA'.

[5] A table containing all descriptions of topics and domains can be found in the Appendix.

2.4 Potential Misalignment of Human Classification

Table 1: Examples of Three types of misalignment

Text Segment	Topic Assignment (LLM)	Topic Assignment (Human Expert)
Scenario 1		
und die viel zu hohe Arbeitslosigkeit vor allem in Süd-und Westeuropa bekämpfen.	Economic Goals	Labour Groups: Positive
Democrats will reverse this rulemaking and restore nondiscrimination protections for LGBTQ+ people and people living with HIV/AIDS in health insurance	Freedom and Human Rights	Equality: Positive
We reject Republican proposals that, in the name of simplification, would make students pay billions of dollars more on their student loans.	Education Expansion	Non-economic Demographic Groups
Scenario 2		
Fair and Simple Taxes for Growth	Economic Growth: Positive	NA
Zukunftsfähiger Güterverkehr	Technology and Infrastructure: Positive	NA
We have cut Income Tax for over 26 million people	Economic Goals	NA
Scenario 3		
As a result, our action has not been enough to cut annual net migration to the tens of thousands.	NA	European Community/Union: Negative
Plaid Cymru's answer: We will make it our target to save 10,000 lives over ten years, through a range of measures from public health actions and promoting individual lifestyle.	NA	Welfare State Expansion
Die Karenzzeit soll in Fällen besonders schwerer Interessenskonflikte auf bis zu drei Jahre ausgeweitet werden können.	NA	Political Corruption

There is one critical aspect we must acknowledge in any research: humans make mistakes. In our study, we discovered that the human-generated codes provided by the CMP team are not always the correct answers in several instances. Therefore, if the answer provided by the LLM differs from the human classification by CMP, this does not necessarily indicate that the LLM is incorrect. We identify three potential scenarios in this context.

2.4.1 LLMs vs. Humans

In certain instances, both the CMP-coded results and the LLM-generated outputs can be valid despite showing different classifications. For example, a party's manifesto might emphasize various ideological points that both CMP and LLM interpret correctly but differently due to their respective methodologies. This dual validity highlights the complexity of political manifestos, which often contain multifaceted messages catering to diverse audience segments.

Conversely, there are scenarios where the CMP code successfully classifies a manifesto, but the LLM yields no clear result (NA). This outcome might stem from the LLM's handling of ambiguous or highly nuanced texts, where its confidence level does not meet the threshold for classification.

Additionally, it should be noted that some sentences are strongly correlated to previous sentences and thus are meaningless when read alone. In this case, human researchers can categorize them based on the preceding context. However, since we feed the large language model sentence by sentence, it is unlikely to predict the previous context. This is why LLMs are likely to provide 'NA' in such cases.

There are also cases where the CMP documentation may not suggest a classification (NA), while the LLM generates a definitive result. This discrepancy can occur if the CMP coding guidelines lack sufficient data or context to classify a particular manifesto. For instance, emerging political parties or unique manifesto formats might not fit neatly into CMP's existing categories, whereas LLMs, leveraging vast and diverse training data, can offer insightful classifications.

2.5 Running the Large Language Models

In our attempt to explore the feasibility of replacing human experts with LLMs we tested the following five models: Gemini 1, LLAMA2 (13B), LLAMA3 (7B), Cohere (Command), and ChatGPT 3.5 (Turbo). These models were selected for their popularity, ease of access, and relatively low cost for deployment.

ChatGPT-3.5, developed by OpenAI, features 175 billion parameters and excels in text generation, translation, summarization, and question answering [11, 38]. Its extensive parameter set allows for maintaining context over extended conversations and can be fine-tuned for specific tasks [11]. OpenAI has implemented robust safety measures to mitigate harmful outputs [45].

Gemini 1, developed by Google, also has 175 billion parameters and performs well in similar tasks [11]. It maintains context over conversations up to 2048 tokens and includes safety measures to reduce harmful content [48]. Gemini 1 offers fine-tuning capabilities for improved task-specific performance [11].

Cohere, with 52 billion parameters, is known for its robust performance in natural language processing tasks [11, 38]. It handles long-context windows and can be fine-tuned for specific applications, enhancing performance and ensuring ethical use with advanced moderation filters [45].

LLAMA2, developed by Meta, has 70 billion parameters and supports context windows of up to 4096 tokens [48]. It offers advanced fine-tuning capabilities and balances parameter efficiency with high performance, including safety protocols to reduce harmful content [45].

LLAMA3, the advanced iteration of LLAMA2, features 100 billion parameters and exhibits superior performance in NLP tasks [11]. It supports context windows of up to 4096 tokens and offers enhanced fine-tuning capabilities, ensuring higher accuracy and ethical use with advanced safety protocols [45].

Table 2: Budget Overview for Model Deployment and Usage

Model Name	Time (hours//minutes)	API Cost Status	Deployment Details
Gemini 1	Cloud-based (varies)	No cost (API)	Online API
LLAMA2 13B	23h 28min	No cost (local)	dolphin-2.9-llama3-8b.Q8_0.gguf (Local)
LLAMA3 7B	4h 28min	No cost (local)	llama-2-13b-chat.Q8_0.gguf (Local)
Cohere Command	Cloud-based (varies)	168.52 USD (API)	Online API
ChatGPT 3.5 Turbo	Cloud-based (varies)	179.87 USD (API)	Online API

As previously mentioned, three of the five LLMs—specifically, Gemini 1, Cohere (Command), and ChatGPT 3.5 (Turbo)—were accessed via online APIs to accelerate the data collection. However, using online APIs usually requires a budget, and in our case, not a small one (as highlighted in Table 2). For that reason, we also took advantage of two locally deployed models, LLAMA 2 (13B) and LLAMA3 (7B). These models were locally run on a Windows 11 platform with an NVIDIA RTX 4090 graphics card, managed through LM Studio [37, 46]. Our machine featured an Intel i5-12600K processor, NVIDIA RTX 4090, Z690 Aorus Ultra motherboard, 32GB DDR5 5200 MHz RAM, and a Samsung 980 Pro SSD. This setup reduced external API costs but was constrained by the VRAM limitations of a single RTX 4090, preventing larger model deployments. For both cloud-based and locally-deployed models, data collection was sequential, with one sentence classification added at a time, and did not employ parallel processing.

3 Empirical Strategy

The goal of the empirical analysis in this paper is to explore the feasibility of replacing human experts trained in calculating party positions with LLMs. To understand the nature of this relationship, several different regression models, as well as traditional statistical estimates (such as correlation tests and RMSE), have been used. The empirical approach for these different models and our brief reasoning is introduced below:

1. **Simple Linear Regression**: A simple model that takes each RILE-Index associated with a country i and year j pair calculated using an LLM ($\text{RILE}_{\text{LLM}_{ij}}$) as the input and investigates the potential of $\text{RILE}_{\text{LLM}_{ij}}$ to predict the ground truth (RILE Index calculated by CMP - $\text{RILE}_{\text{CMP}_{ij}}$).

2. **Multiple Linear Regression**: A slightly more complex model that also uses $\text{RILE}_{\text{LLM}_{ij}}$ to predict $\text{RILE}_{\text{CMP}_{ij}}$. In this model, we also add *country*, *party*, and *year* dummy variables with the idea that the CMP's assignment of experts to manifestos may be different based on country, party, and year. This assignment may result in non-random differences in the calculation of RILE-Indices and therefore needs to be controlled.

3. **Mixed Effects Regression**: In this third approach, we examine the same relationship between $\text{RILE}_{\text{LLM}_{ij}}$ and $\text{RILE}_{\text{CMP}_{ij}}$. We believe that there might be a nested structure in the CMP data, such as variations within parties, countries, and years. The mixed effects models are designed to handle unobserved heterogeneity, to improve generalizability, and to address the non-independence of observations within these nested groups.

4. **Quantile Regression**: In our final regression approach, we examine the same relationship using the quantile model to better understand the distributional effects. The quantile method allows us to see how the relationship varies across different points of the distribution of the RILE indices, revealing heterogeneity that ordinary least squares regression might miss. In fact, one of our hypotheses suggests that extreme political positions will be interpreted as less extreme by the LLMs. The

quantile regression model can show if the LLMs perform differently in predicting lower versus higher RILE indices, thereby offering deeper insights into the accuracy and reliability of LLM predictions across the entire range of the data.

5. **Correlation Test**: Another simple test is to measure the strength and direction of the linear relationship between $\text{RILE}_{\text{LLM}_{ij}}$ and $\text{RILE}_{\text{CMP}_{ij}}$. By calculating the correlation coefficient, we can determine whether the LLMs produce results that are consistently aligned with the human experts' calculations.

6. **RMSE Comparison**: RMSE quantifies the difference between the values predicted by the LLMs and the ground truth values provided by human experts. By calculating the RMSE, we can evaluate how accurately the LLMs replicate the expert-calculated RILE indices, with lower RMSE values indicating better model performance. This metric is particularly useful because it penalizes larger errors more heavily.

4 Results

This section provides the main findings extracted from the dataset produced by the output of the LLMs and the methods outlined in the Empirical Strategy section. Correlation-based statistical methods introduced above indicate that there is a strong relationship between the RILE indices produced by human experts and the RILE scores predicted by LLMs.

The scatterplot below provides a comparison between the RILE scores computed by human experts and the predicted RILE scores. This is possibly the most important result of our analysis. The graph reveals an interesting finding: RILE scores computed by the LLMs are less extreme than those calculated by human experts. As a result, parties at both ends of the political spectrum are pushed towards the center. This effect is pronounced for each LLM; however, the impact on parties differs. Both Llama2 and Gemini 1 push left parties slightly more strongly towards the center than they push right parties. In contrast, ChatGPT 3.5, Cohere, and Llama3 push right parties towards the center, with Llama3 pushing almost all political parties (including right parties) towards the left side of the political spectrum. Therefore, Llama3 is the least convincing model of all five. A more

detailed breakdown by country and related explanation have been provided in the Appendix. The group of scatterplots in the Appendix shows that German parties are generally more left-wing compared to other countries, while the parties from the USA are generally more right-wing compared to the rest.

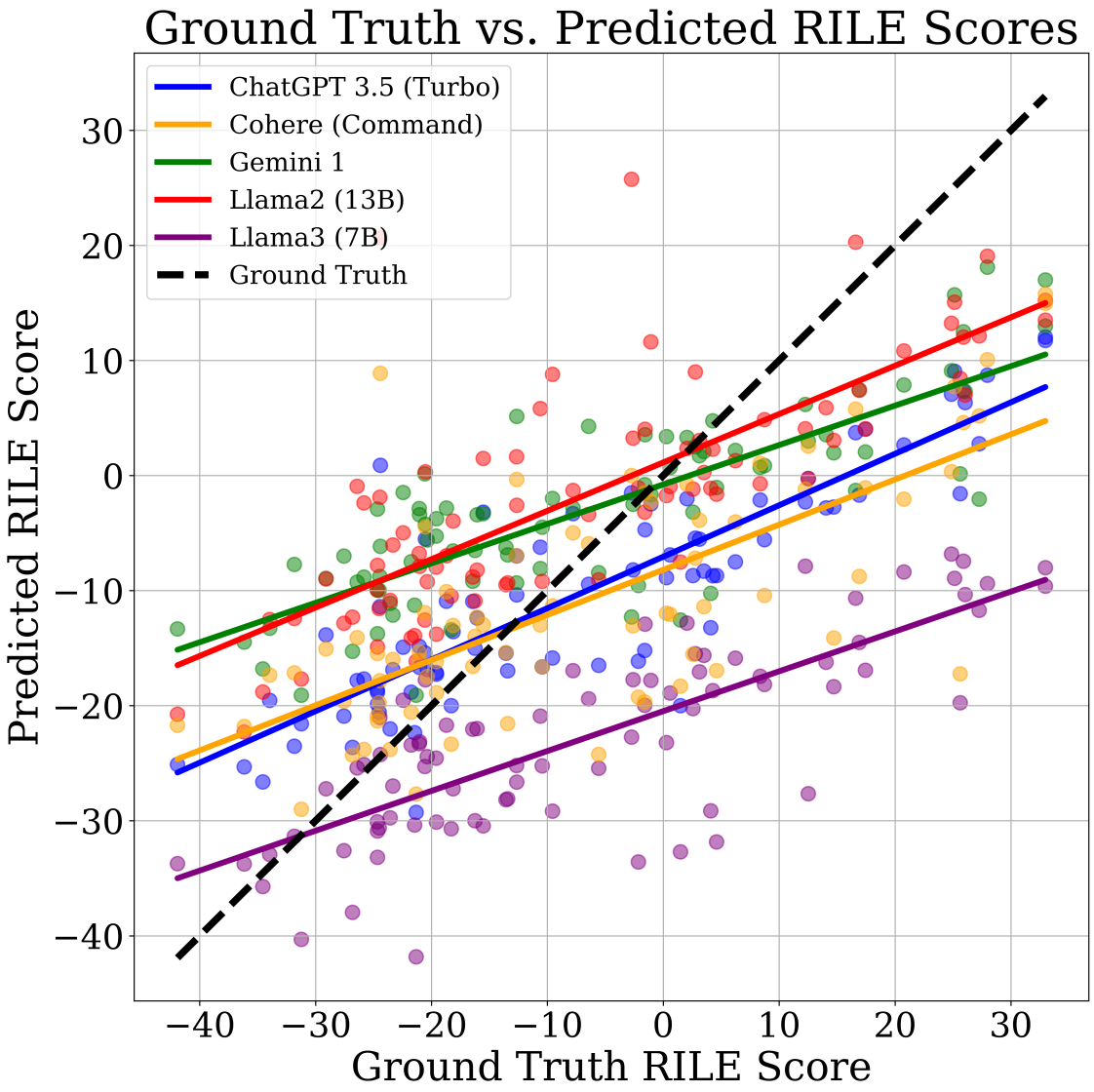


Figure 3: Party Positions: Ground Truth vs. Estimations

The RMSE and correlation plots show the pairwise quantified RMSE and correlation values between all LLMs and the RILE scores provided by the CMP. The most important finding is in the first row and the first column: ChatGPT 3.5 provides the most similar results to those obtained by human experts (RMSE = 11.83), with Llama2 closely trailing

behind (RMSE = 13.57). Another notable finding is the close similarity between the scores predicted by Cohere and ChatGPT 3.5. ChatGPT also provides the highest correlation with the RILE Index provided by CMP, statistically significantly. In this case, the correlation values are provided by ChatGPT (0.86), with Gemini 1 trailing behind (0.80). These two facts, along with the scatterplot shared above, suggest that ChatGPT 3.5 may be the most feasible candidate for replacing human experts. The breakdown by party position and country has been provided in the Appendix. The breakdown of RMSE values does not suggest that LLMs predict the RILE scores for a certain subset of manifestos more successfully than other manifestos. The breakdown of correlation values by country and party, on the other hand, indicates that LLMs may more successfully predict RILE scores for right parties, with ChatGPT 3.5 showing the highest correlation (all correlation tests between predicted values and ground truth are significant).

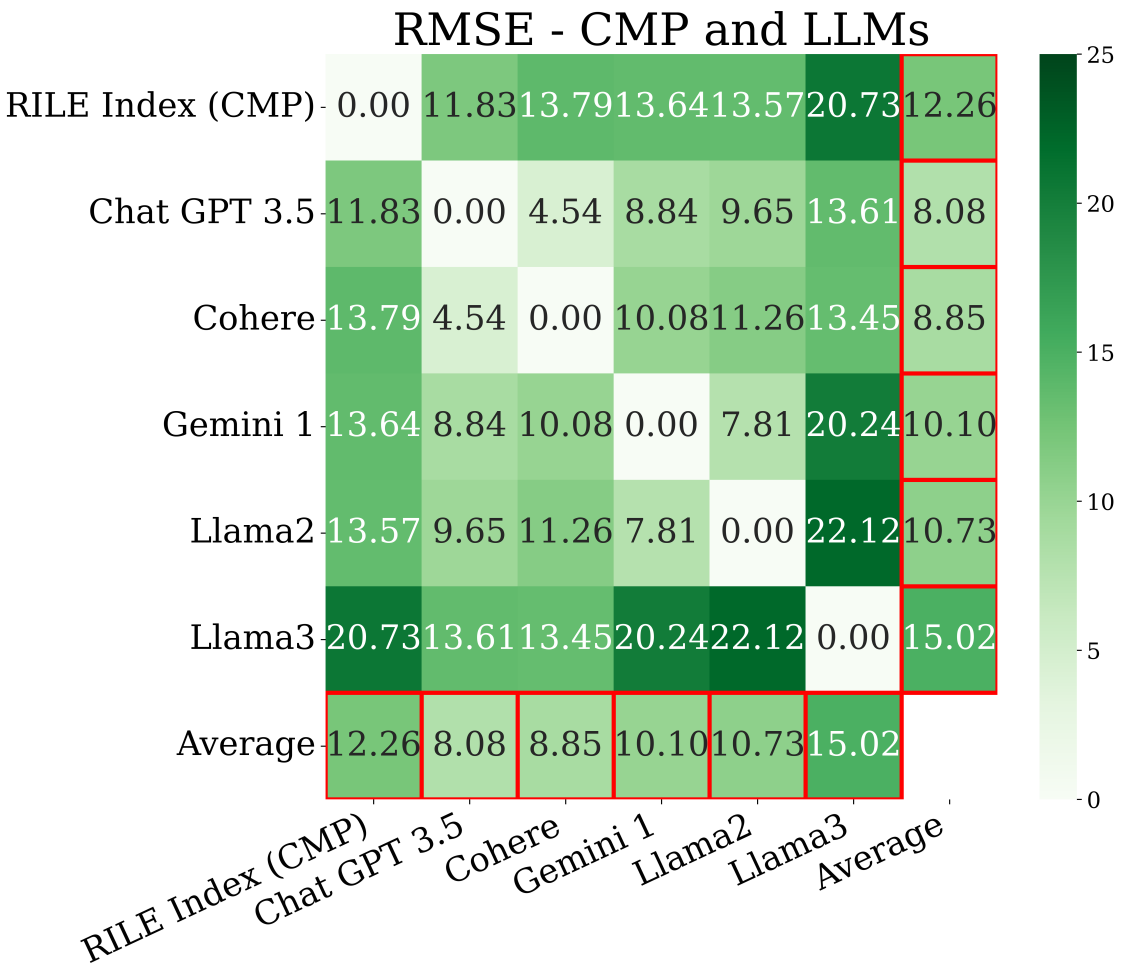


Figure 4: RMSE Between Ground Truth and Predictions

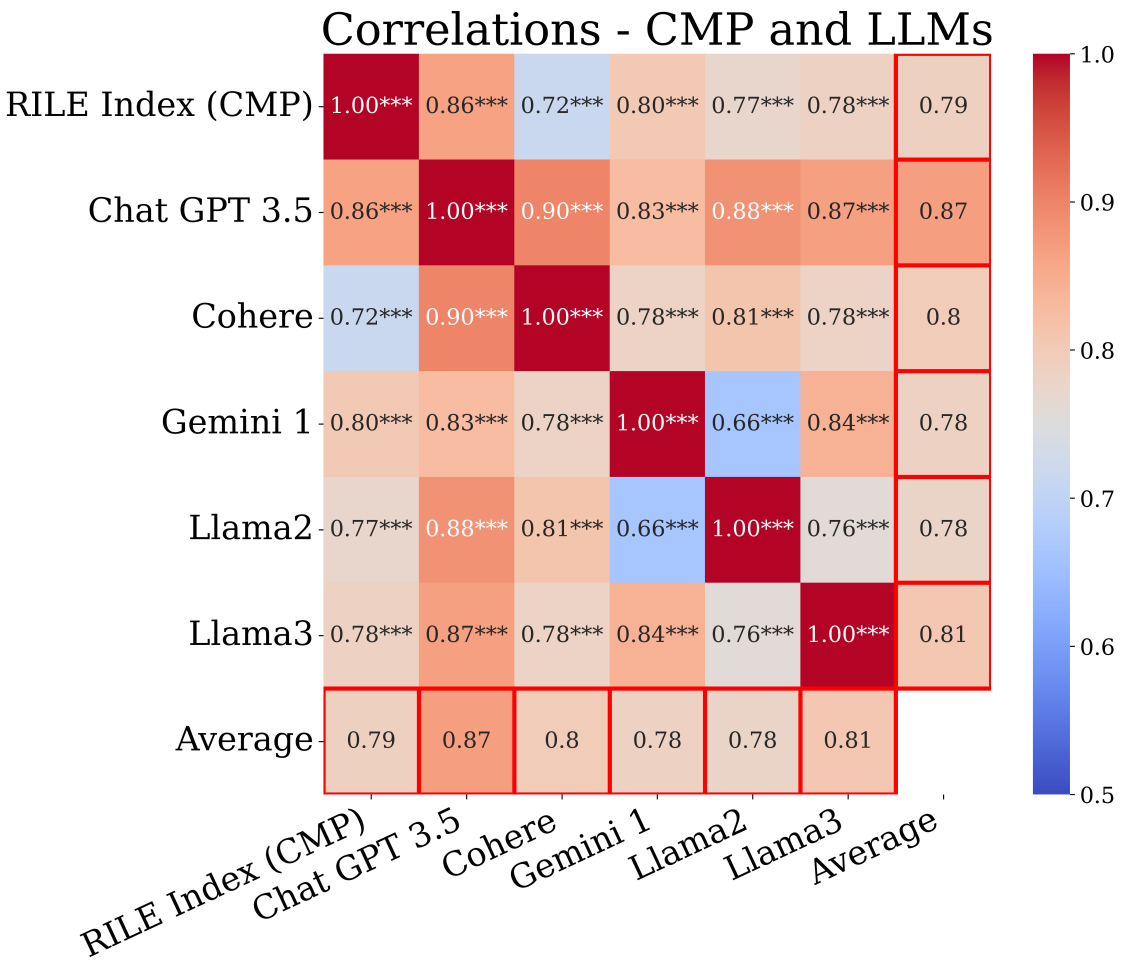


Figure 5: Correlations Between Ground Truth and Predictions

In the last section of the results, we present the findings for linear regression in the following order: (1) Simple and multiple linear regression, (2) Mixed effects regression, and (3) Quantile regression. The coefficients reported in the model tables indicate the amount of increase in RILE scores provided by CMP when the RILE score produced by LLMs is increased by a single unit. Therefore, in general, models that produce coefficients close to one (1) are considered more closely aligned with reality, as they produce results closer to those generated by human experts. Additionally, R^2 values show the amount of variance explained by the LLM output; as usual, a higher R^2 value is preferred. Focusing on these two criteria, the simple and multiple linear regression models indicate that ChatGPT 3.5 and Llama 2 provide the most convincing results, since they provide coefficient values closest to 1, and also have relatively high R^2 scores. For the mixed effect regression results, the winner

is less clear: ChatGPT 3.5 and Cohere seem to be performing equally well.

Table 3: Simple and Multiple Linear Regression Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
ChatGPT 3.5	1.671					1.675***				
Cohere		1.305					1.497***			
Gemini 1			1.884					2.131***		
Llama 2				1.419					1.346***	
Llama 3					1.776					1.889***
Country						X	X	X	X	X
Year						X	X	X	X	X
Observations	84	77	84	84	84	77	77	77	77	77
R^2	0.75	0.51	0.65	0.6	0.62	0.82	0.68	0.76	0.72	0.81
Adjusted R^2	0.74	0.51	0.64	0.59	0.61	0.77	0.59	0.69	0.63	0.74

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4: Mixed Effects Regression Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
ChatGPT 3.5	1.690***					1.679***				
Cohere		1.380***					1.389***			
Gemini 1			1.901***					1.920***		
Llama 2				1.388***					1.422***	
Llama 3					1.776***					1.878***
Group Variance	0.219	0.165	0.03	0.473	0.313	0.236	0.254	0.145	0.184	0.575
Country						X	X	X	X	X
Year						X	X	X	X	X
Observations	84	77	84	84	84	77	77	77	77	77
R^2	0.75	0.51	0.65	0.6	0.62	0.82	0.68	0.76	0.72	0.81
Adjusted R^2	0.74	0.5	0.64	0.59	0.61	0.77	0.59	0.69	0.63	0.74

Note: *p<0.1; **p<0.05; ***p<0.01

Lastly, we focus on quantile regression to understand whether the predictive power of LLMs differs for political parties in different parts of the political spectrum. Specifically, the dependent variable (RILE Index provided by the CMP) is divided into quantiles, and separate simple linear regression models are created for different subsets of the data. Here, ChatGPT seems to perform the best, providing the highest R^2 values (0.68 on average versus 0.54 produced by Gemini for the simple linear regression and 0.76 on average versus 0.70 produced by Llama 3 for the multiple linear regression). Coefficient-wise, the only model that is relatively close to 1 is Llama 2 for Q=0.25.

Overall, the coefficients obtained from the regression models show that LLMs always introduce a bias in the calculation of the RILE scores. This bias results in both left-wing and

right-wing parties being treated as centrist. The amount of push towards the center ranges from 8.6% (Llama 2 in the quantile regression model) to 113.1% (Gemini 1 in the multiple linear regression model). The average push towards the center is 73.64% (with a median of 79%). This means that, on average, leftist and rightist manifestos are being interpreted as 73.64% less extreme than they actually are. Assuming that human experts trained by CMP can do their jobs successfully, this large bias suggests that human experts may not be easily replaceable in the near future.

Table 5: Quantile Regression Results

Model	Quantile	Coefficient	Variance	Pseudo R^2	Coefficient	Variance	Pseudo R^2
ChatGPT 3.5	0.25	1.747***	0.012	**0.67**	1.853***	0.011	**0.76**
ChatGPT 3.5	0.5	1.838***	0.016	**0.74**	1.867***	0.012	**0.82**
ChatGPT 3.5	0.75	1.681***	0.015	**0.65**	1.892***	0.010	**0.72**
Cohere	0.25	1.410***	0.025	**0.33**	1.752***	0.042	**0.52**
Cohere	0.5	1.513***	0.036	**0.50**	1.835***	0.026	**0.64**
Cohere	0.75	1.312***	0.047	**0.34**	1.832***	0.024	**0.48**
Gemini 1	0.25	1.997***	0.030	**0.48**	1.956***	0.034	**0.60**
Gemini 1	0.5	1.998***	0.042	**0.64**	2.121***	0.038	**0.75**
Gemini 1	0.75	1.980***	0.064	**0.49**	2.121***	0.036	**0.65**
Llama 2	0.25	1.086***	0.039	**0.34**	1.502***	0.022	**0.61**
Llama 2	0.5	1.722***	0.021	**0.54**	1.646***	0.013	**0.69**
Llama 2	0.75	1.853***	0.016	**0.36**	1.783***	0.010	**0.49**
Llama 3	0.25	1.840***	0.031	**0.37**	1.836***	0.021	**0.65**
Llama 3	0.5	2.019***	0.044	**0.60**	1.947***	0.027	**0.79**
Llama 3	0.75	1.797***	0.073	**0.48**	2.097***	0.027	**0.66**
Country					X	X	X
Year					X	X	X

Note: *p<0.1; **p<0.05; ***p<0.01

Pseudo R^2: Pseudo R-squared in quantile regression measures how well the model explains the variability in the data, similar to R-squared in ordinary least squares regression but adapted for quantile-based models. It provides an indication of model fit, with higher values suggesting a better fit, although it does not directly represent the proportion of variance explained.

5 Discussion

This study aimed to investigate the feasibility of using large language models (LLMs) to calculate party positions from political manifestos and compare these results to those produced by human experts. By analyzing the outputs of five LLMs—ChatGPT 3.5 Turbo, Cohere Command, Gemini 1, Llama 2, and Llama 3—we evaluated their performance in estimating the RILE (Right-Left) Index for political parties in Germany, the United Kingdom,

and the United States. Our findings reveal several critical insights.

Firstly, the LLMs consistently introduced a bias towards the center, interpreting leftist and rightist manifestos as 73.64% less extreme on average than they truly are. This bias is significant, given that CMP-trained human experts are likely performing their tasks accurately. The degree of bias varied across models, with Llama 2 in the quantile regression model showing the least bias (8.6%) and Gemini 1 in the multiple linear regression model showing the most (113.1%). This suggests that LLMs may not yet be suitable replacements for human experts in the calculation of party positions due to their tendency to neutralize the extremity of political views.

Secondly, the regression models highlighted differences in the predictive power of each LLM. Simple and multiple linear regression results indicated that ChatGPT 3.5 and Llama 2 produced the most convincing results, with coefficients closest to 1 and relatively high R^2 scores. In mixed effects regression, both ChatGPT 3.5 and Cohere performed well. Quantile regression results showed that ChatGPT 3.5 provided the highest R^2 values across different quantiles, reinforcing its potential as a leading candidate among the LLMs evaluated.

Additionally, while deploying these models incurred some computational costs, these were not significant when compared to the financial costs that would be required if human experts were to perform the same tasks. For instance, the cost of training and employing human coders to analyze 84 manifestos, containing a total of 140,432 text segments, would be substantially higher. Cloud-based models such as Gemini 1, Cohere Command, and ChatGPT 3.5 involved API costs, but these were still far less than the costs associated with human labor. Locally deployed models like Llama 2 and Llama 3 provided a cost-effective alternative, although they were constrained by hardware limitations.

The bias introduced by LLMs, which tends to portray leftist and rightist manifestos as more centrist than they are, can have significant implications for democracy and voter perception. If voters rely on these biased outputs to understand party positions, they may be misled about the true ideological stances of political parties. This misrepresentation can distort voter expectations and influence their voting decisions, potentially leading to a misalignment between voter preferences and elected representatives.

Moreover, the perceived centrist bias of LLM outputs could undermine trust in AI tools

used in political analysis, especially if voters become aware of these biases. Maintaining voter trust is crucial for the adoption of AI in democratic processes. As AI ethicist Timnit Gebru points out, "AI systems are not neutral; they reflect the biases of their creators and the data they are trained on." This underscores the importance of transparency and bias mitigation in AI deployment.

In summary, while LLMs offer promising capabilities in text analysis and could potentially aid in the evaluation of political manifestos, their current biases and limitations suggest that they cannot yet replace human experts. Future advancements in LLM technology, combined with refined prompt engineering and bias mitigation strategies, may enhance their accuracy and reliability. Until then, human expertise remains crucial for accurately interpreting and positioning political party manifestos. This study underscores the need for continued research and development in AI tools to better support democratic processes and voter education.

References

[1] Abubakar Abid, Maheen Farooqi, and James Zou. "Large language models associate Muslims with violence". In: *Nature Machine Intelligence* 3.6 (2021), pp. 461–463.

[2] Gabriel A. Almond and Sidney Verba. *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Princeton University Press, 1963. ISBN: 978-0-691-07503-7. URL: https://www.jstor.org/stable/j.ctt183pnr2 (visited on 07/25/2024).

[3] Lisa P Argyle et al. "Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale". In: *Proceedings of the National Academy of Sciences* 120.41 (2023), e2311627120.

[4] Lisa P Argyle et al. "Out of one, many: Using language models to simulate human samples". In: *Political Analysis* 31.3 (2023), pp. 337–351.

[5] Pablo Barberá. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data". In: *Political analysis* 23.1 (2015). Publisher: Cambridge University Press, pp. 76–91. URL: https://www.cambridge.org/core/journals/political-analysis/article/birds-of-the-same-feather-tweet-together-bayesian-ideal-point-estimation-using-twitter-data/91E37205F69AEA32EF27F12563DC2A0A (visited on 07/25/2024).

[6] Emily M. Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? " In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 610–623. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922 (visited on 07/25/2024).

[7] Adam J. Berinsky. "Measuring Public Opinion with Surveys". eng. In: *Annual review of political science* 20.1 (2017). Publisher: Annual Reviews, pp. 309–329. ISSN: 1094-2939. DOI: 10.1146/annurev-polisci-101513-113724.

[8] Su Lin Blodgett et al. "Language (Technology) is Power: A Critical Survey of "Bias" in NLP". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 5454–5476. DOI: 10.18653/v1/2020.acl-main.485. URL: https://aclanthology.org/2020.acl-main.485 (visited on 07/25/2024).

[9] Adam Bonica. "Ideology and Interests in the Political Marketplace". en. In: *American Journal of Political Science* 57.2 (2013). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12014, pp. 294–311. ISSN: 1540-5907. DOI: 10.1111/ajps.12014. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12014 (visited on 07/25/2024).

[10] Thomas Brambor, William Roberts Clark, and Matt Golder. "Understanding interaction models: Improving empirical analyses". In: *Political analysis* 14.1 (2006), pp. 63–82.

[11] Tom B. Brown et al. *Language Models are Few-Shot Learners*. arXiv:2005.14165 [cs]. July 2020. DOI: 10.48550/arXiv.2005.14165. URL: http://arxiv.org/abs/2005.14165 (visited on 07/25/2024).

[12] Ian Budge. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments, 1945-1998*. en. Google-Books-ID: bwkjzbsDwAsC. Oxford University Press, 2001. ISBN: 978-0-19-924400-3.

[13] Paul Christiano et al. *Deep reinforcement learning from human preferences*. arXiv:1706.03741 [cs, stat]. Feb. 2023. DOI: 10.48550/arXiv.1706.03741. URL: http://arxiv.org/abs/1706.03741 (visited on 07/25/2024).

[14] Jessica Di Cocco and Bernardo Monechi. "How populist are parties? Measuring degrees of populism in party manifestos using supervised machine learning". In: *Political Analysis* 30.3 (2022), pp. 311–327.

[15] Anthony Downs. *An Economic Theory of Democracy*. en. Google-Books-ID: kLEGAAAAMAAJ. Harper, 1957. ISBN: 978-0-06-041750-5.

[16] Maurice Duverger. *Political parties: their organization and activity in the modern state*. eng. Trans. by Barbara North and Robert North. OCLC: 983396. London, New York: Methuen & Co. ; John Wiley & Sons, 1954.

[17] Michael Gallagher, Michael Laver, and Peter Mair. *Representative Government in Modern Europe*. en. Google-Books-ID: _odqtwAACAAJ. Mcgraw Hill Higher Education, 2011. ISBN: 978-0-07-712967-5.

[18] Kostas Gemenis. "What to do (and not to do) with the comparative manifestos project data". In: *Political Studies* 61.1_suppl (2013), pp. 3–23.

[19] Igor Grossmann et al. "AI and the transformation of social science research". In: *Science* 380.6650 (2023), pp. 1108–1109.

[20] Peter A. Hall et al., eds. *Varieties of Capitalism: The Institutional Foundations of Comparative Advantage*. Oxford, New York: Oxford University Press, Aug. 2001. ISBN: 978-0-19-924775-2.

[21] Yibo Hu et al. "Conflibert: A pre-trained language model for political conflict and violence". In: Association for Computational Linguistics. 2022.

[22] Bernard J. Jansen, Soon-gyo Jung, and Joni Salminen. "Employing large language models in survey research". In: *Natural Language Processing Journal* 4 (Sept. 2023), p. 100020. ISSN: 2949-7191. DOI: 10.1016/j.nlp.2023.100020. URL: https://www.sciencedirect.com/science/article/pii/S2949719123000171 (visited on 07/25/2024).

[23] Herbert Kitschelt. *The Transformation of European Social Democracy*. Cambridge Studies in Comparative Politics. Cambridge: Cambridge University Press, 1994. ISBN: 978-0-521-45106-2. DOI: 10.1017/CBO9780511622014. URL: https://www.cambridge.org/core/books/transformation-of-european-social-democracy/C92F284FC17302253C3B5B14123BBA80 (visited on 07/25/2024).

[24] Hans-Dieter Klingemann et al. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union, and OECD 1990-2003*. Oxford, New York: Oxford University Press, Nov. 2006. ISBN: 978-0-19-929631-6.

[25] Michael Laver and John Garry. "Estimating policy positions from political texts". In: *American Journal of Political Science* (2000), pp. 619–634.

[26] Michael Laver and John Garry. "Estimating Policy Positions from Political Texts". In: *American Journal of Political Science* 44.3 (2000). Publisher: [Midwest Political Science Association, Wiley], pp. 619–634. ISSN: 0092-5853. DOI: 10.2307/2669268. URL: https://www.jstor.org/stable/2669268 (visited on 07/25/2024).

[27] Arend Lijphart. *Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries*. English. New Haven: Yale University Press, July 1999. ISBN: 978-0-300-07893-0.

[28] Vivien Lowndes, David Marsh, and Gerry Stoker, eds. *Theory and Methods in Political Science*. English. 4th edition. Basingstoke, Hampshire: Bloomsbury Academic, Oct. 2017. ISBN: 978-1-137-60351-7.

[29] Arthur Lupia and Mathew D. McCubbins. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* en. Google-Books-ID: 2Vv6BhLC6HUC. Cambridge University Press, Mar. 1998. ISBN: 978-0-521-58593-4.

[30] Ian McAllister. "Housing tenure and party choice in Australia, Britain and the United States". In: *British Journal of Political Science* 14.4 (1984), pp. 509–522.

[31] Gaël Le Mens and Aina Gallego. *Scaling Political Texts with Large Language Models: Asking a Chatbot Might Be All You Need*. en. Nov. 2023. URL: https://arxiv.org/abs/2311.16639v2 (visited on 07/25/2024).

[32] Wolfgang Merkel. "Is there a crisis of democracy?" In: *Democratic Theory* 1.2 (2014). Publisher: Berghahn Journals, pp. 11–25. URL: https://www.berghahnjournals.com/view/journals/democratic-theory/1/2/dt010202.xml (visited on 07/25/2024).

[33] Thomas M Meyer, Martin Haselmayer, and Markus Wagner. "Who gets into the papers? Party campaign messages and the media". In: *British Journal of Political Science* 50.1 (2020), pp. 281–302.

[34] Maria Elayna Mosley. *Interview Research in Political Science*. eng. 1st ed. Pages: x–x. Ithaca: Cornell University Press, 2013. ISBN: 978-0-8014-5194-2. DOI: 10.7591/9780801467974.

[35] Stefan Müller and Sven-Oliver Proksch. "Nostalgia in European Party Politics: A Text-Based Measurement Approach". In: *British Journal of Political Science* (2023), pp. 1–13.

[36] Keith T. Poole and Howard Rosenthal. "Patterns of congressional voting". In: *American journal of political science* (1991). Publisher: JSTOR, pp. 228–278. URL: https://www.jstor.org/stable/2111445?casa_token=2hHMqtatYWMAAAAA:IwOWV_GQD-Z6AA8KiSza-3eh3bmfCGUHbzySi2S3WJSAfeIQAPUlRDNLnMj99KstLkcTcmoccdh_3iRq9j9N9CyPMfE6GabhvLjoqUuB5YbJIK9o7cQ (visited on 07/25/2024).

[37] *QuantFactory/dolphin-2.9-llama3-8b-GGUF · Hugging Face — huggingface.co*. https://huggingface.co/QuantFactory/dolphin-2.9-llama3-8b-GGUF. [Accessed 28-05-2024].

[38] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". en. In: ().

[39] David Rozado. "The political biases of chatgpt". In: *Social Sciences* 12.3 (2023), p. 148.

[40] Jérôme Rutinowski et al. "The self-perception and political biases of ChatGPT. arXiv". In: *arXiv preprint arXiv:2304.07333* (2023).

[41] *Safety & responsibility*. en-US. URL: https://openai.com/safety/ (visited on 07/25/2024).

[42] Giovanni Sartori. *Parties and Party Systems: A Framework for Analysis*. English. Colchester: ECPR Press, Apr. 2005. ISBN: 978-0-9547966-1-7.

[43] Anna Schmidt and Michael Wiegand. "A Survey on Hate Speech Detection using Natural Language Processing". In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Ed. by Lun-Wei Ku and Cheng-Te Li. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1–10. DOI: 10.18653/v1/W17-1101. URL: https://aclanthology.org/W17-1101 (visited on 07/25/2024).

[44] Boris Shor, Nolan McCarty, and Christopher R Berry. "Methodological Issues in Bridging Ideal Points in Disparate Institutions in a Data Sparse Environment". In: *Available at SSRN 1746582* (2011).

[45] Irene Solaiman et al. *Release Strategies and the Social Impacts of Language Models*. arXiv:1908.09203 [cs]. Nov. 2019. DOI: 10.48550/arXiv.1908.09203. URL: http://arxiv.org/abs/1908.09203 (visited on 07/25/2024).

[46] *TheBloke/Llama-2-13B-chat-GGUF · Hugging Face — huggingface.co*. https://huggingface.co/TheBloke/Llama-2-13B-chat-GGUF. [Accessed 28-05-2024].

[47] Petter Törnberg. "Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning". In: *arXiv preprint arXiv:2304.06588* (2023).

[48] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (visited on 07/25/2024).

[49] Andrea Volkens et al. *Manifesto Project Dataset*. eng. 2020. DOI: 10.25522/manifesto.mpds.2020a. URL: https://commons.datacite.org/doi.org/10.25522/manifesto.mpds.2020a (visited on 07/25/2024).

[50] Yu Wang. "On Finetuning Large Language Models". In: *Political Analysis* (2023), pp. 1–5.

6 Appendix

6.1 Additional Statistical Results

The group of scatterplots below shows the predictions made by the LLMs versus the human experts. The graphs indicate that the differences in party positions introduced by the LLMs are similar to each other except for a few minor discrepancies. A push towards the left end of the political spectrum appears to be stronger in the case of the United States. The strength of the shift towards the left is more pronounced for some LLMs and ranked as follows: (1) Llama2, (2) Gemini 1, (3) ChatGPT 3.5, (4) Cohere, and (5) Llama3.

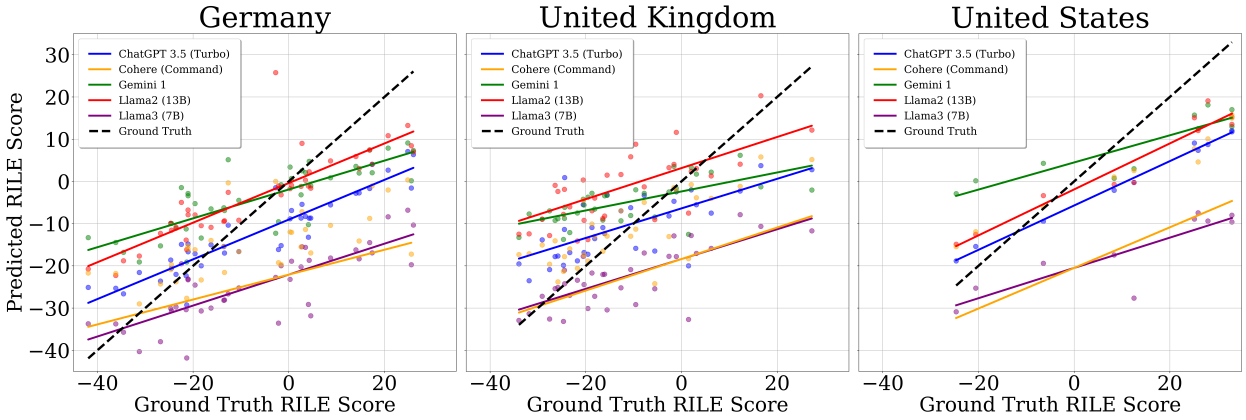


Figure 6: Party Positions: Ground Truth vs. Estimations for Different Countries

The groups of correlation matrices below show the strength of correlations between LLMs and RILE scores provided by CMP for political parties from different ideologies and different countries used in the analysis. The results indicate that the predictions made by the LLMs are most closely correlated with the RILE scores associated with manifestos from the USA and those compiled by right-wing parties. This is expected, as the political parties from the USA are more right-wing compared to other parties in the dataset (as seen in the group of scatterplots above).

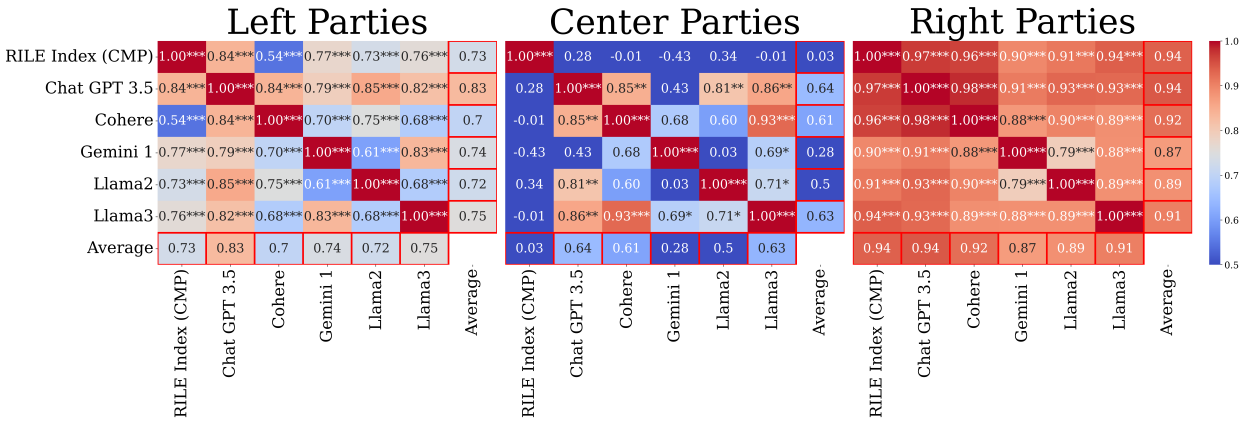


Figure 7: Correlations Between Ground Truth and Predictions for Different Party Positions

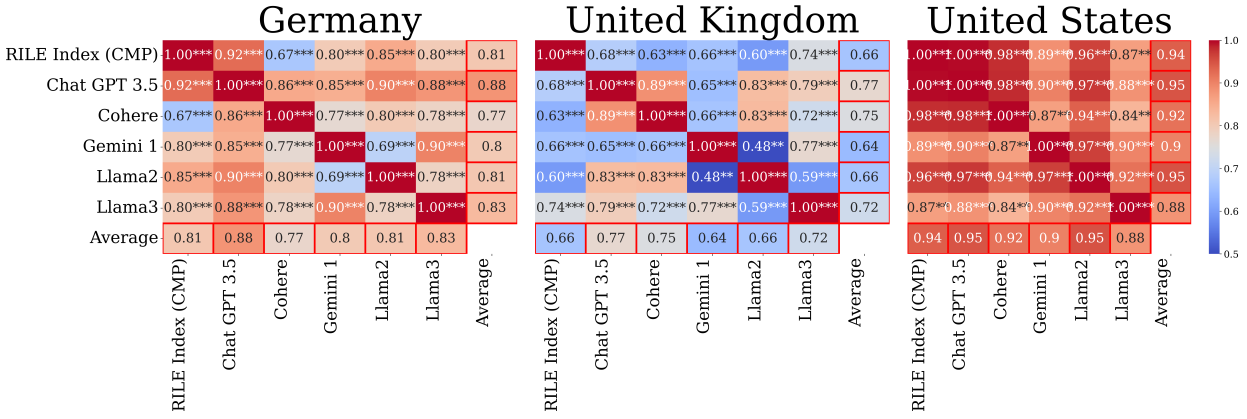


Figure 8: Correlations Between Ground Truth and Predictions for Different Countries

The RMSE scores below provide a comparative evaluation of the LLM scores for different party positions and countries. In most cases, ChatGPT 3.5 and Cohere seem to be the best-performing LLMs. The variation between the different country and party subsets is not significant enough to claim that the performance of LLMs for certain sub-groups is significantly higher or lower compared to other sub-groups.

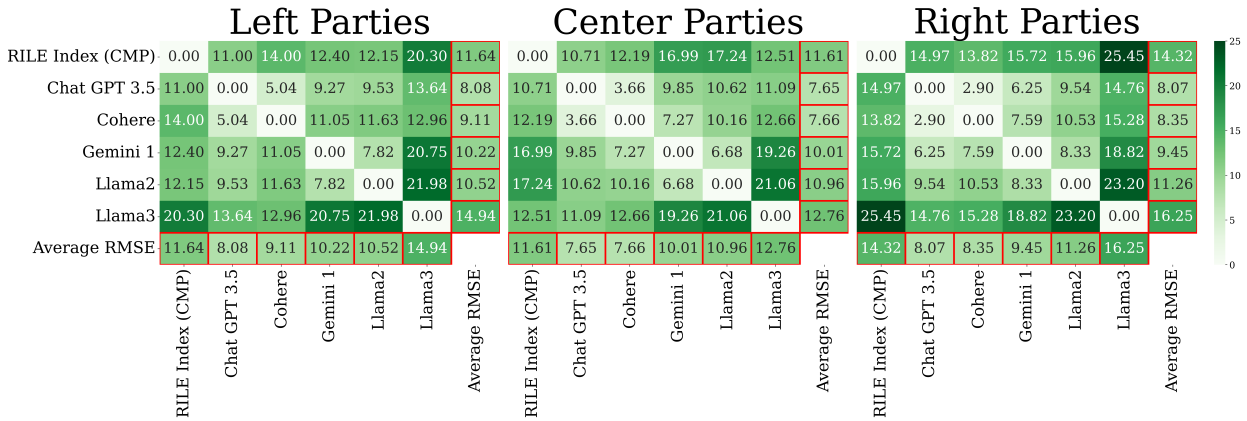


Figure 9: RMSE Between Ground Truth and Predictions for Different Party Positions

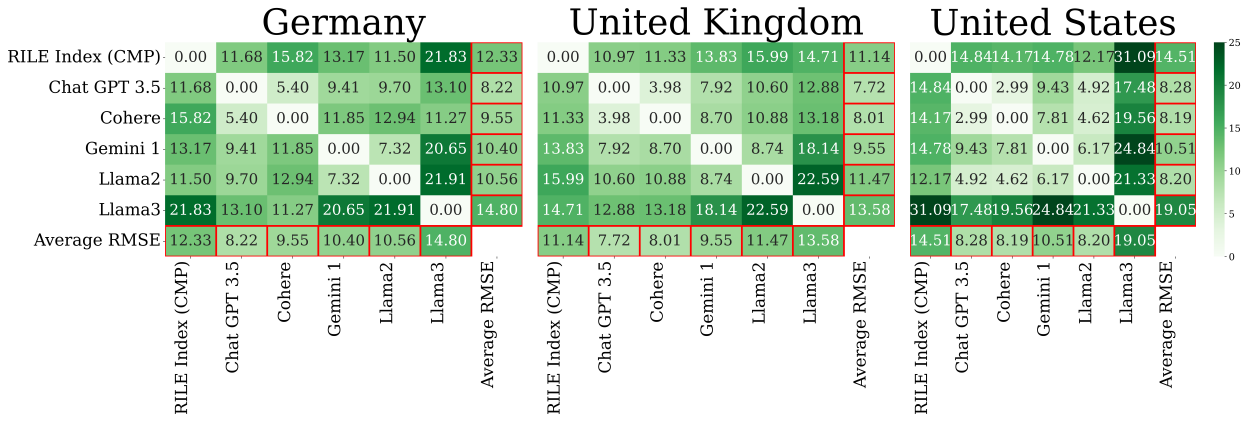
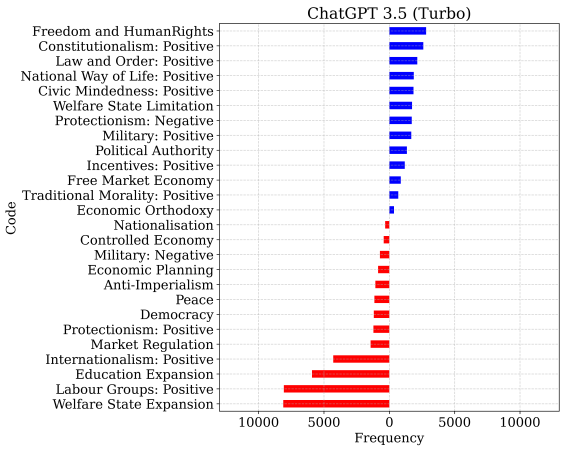
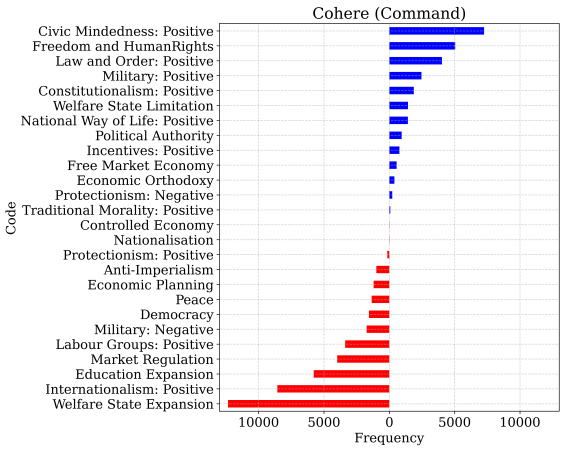


Figure 10: RMSE Between Ground Truth and Predictions for Different Countries

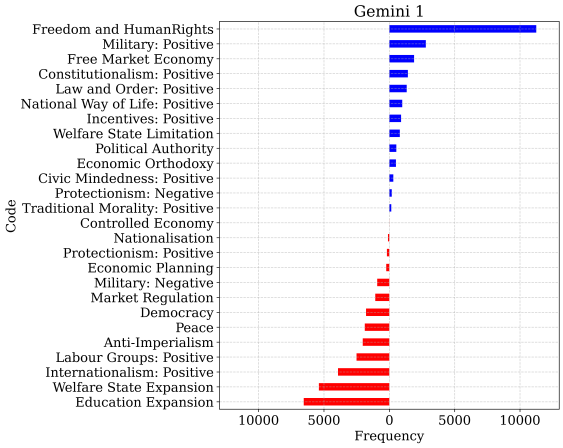
The bar charts below show the distribution of domains predicted by different LLMs. Blue represents domains categorized as left-wing policies, while red represents domains associated with right-wing policies.



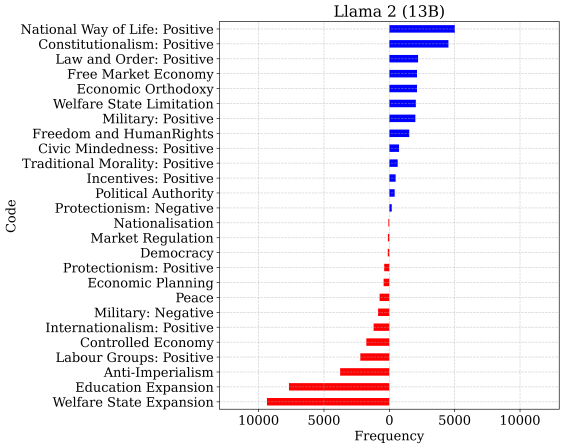
(a) ChatGPT 3.5 (Turbo)



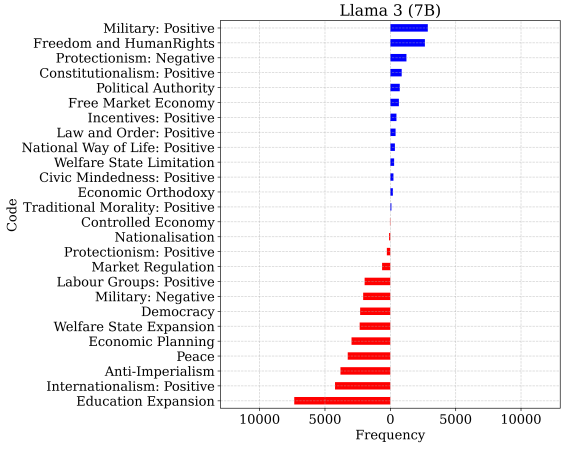
(b) Cohere (Command)



(c) Gemini 1



(d) Llama 2 (13B)



(e) Llama 3 (7B)

Figure 11: Topic Frequencies By LLMs

6.2 Accuracy Values for Topics and Domains

The confusion matrices below show the percentage and count values for each domain and topic. The results indicate that the accuracies obtained by LLMs are not very high - however, this is understandable due to the broad range of domains and topics. Additionally, two text segments may be associated with more than one topic (or domains), in which case the human usually needs to make a subjective decision. A set of examples are illustrated below:

1. Germany: CDU/CSU (Christian Democratic Union/Christian Social Union)
Manifesto: CDU/CSU Election Manifesto 2021
Text Segment:

> "We are committed to ensuring that all citizens have access to quality education and training throughout their lives. This not only strengthens our economy by providing a skilled workforce but also promotes social cohesion and equal opportunities for all."

Potential Topics:

(1) Education

(2) Economy

(3) Social Policy

2. United Kingdom: Labour Party
Manifesto: Labour Party Manifesto 2019
Text Segment:

> "Our plan to invest in green technologies will create new jobs and help combat climate change, ensuring a sustainable future for our children. This initiative will also reduce energy costs for households, making life more affordable for everyone."

Potential Topics:

(1) Environment/Climate Change

(2) Economy/Employment

(3) Social Policy/Energy

3. United States: Democratic Party

Manifesto: Democratic Party Platform 2020

Text Segment:

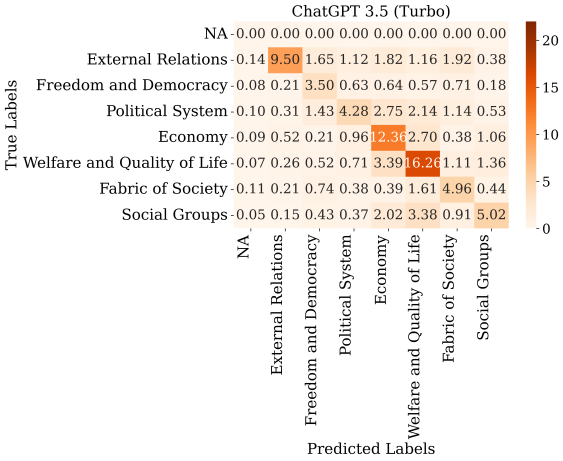
> "We believe that healthcare is a fundamental human right. By expanding access to affordable healthcare, we can improve public health outcomes, reduce financial strain on families, and promote economic stability."

Potential Topics:

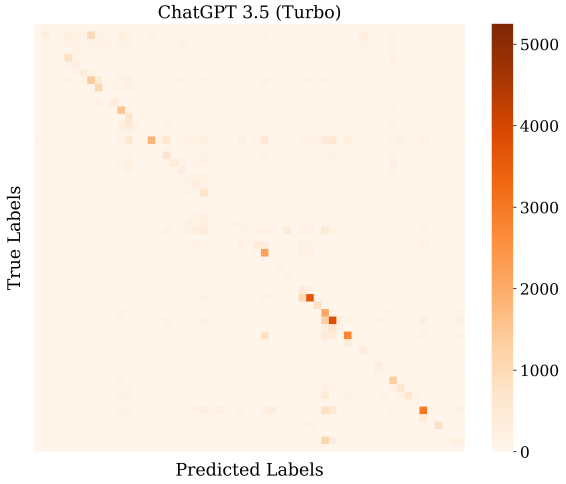
(1) Healthcare

(2) Social Policy

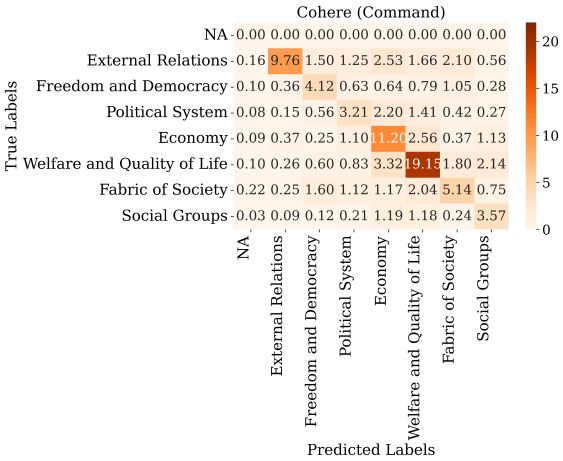
(3) Economy



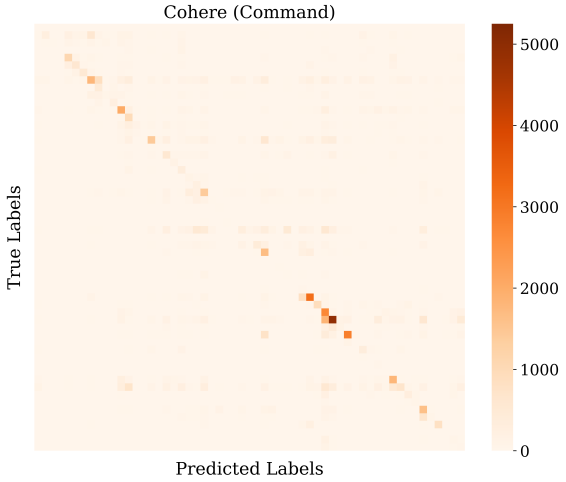
(a) ChatGPT 3.5 (Turbo) - Domains



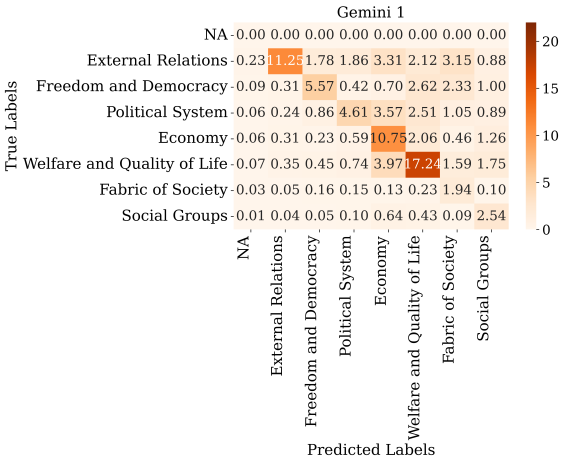
(b) ChatGPT 3.5 (Turbo) - Topics



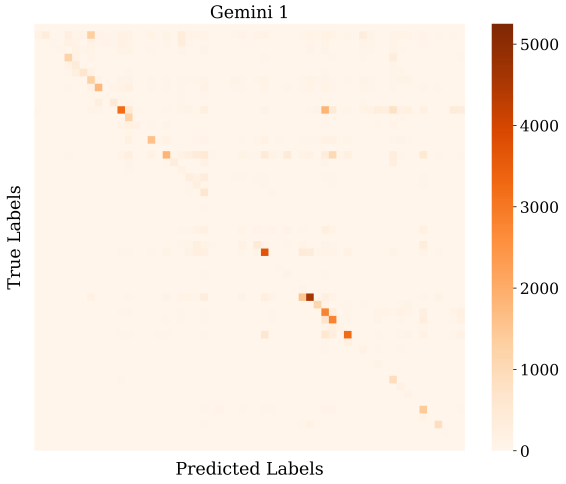
(c) Cohere (Command) - Domains



(d) Cohere (Command) - Topics

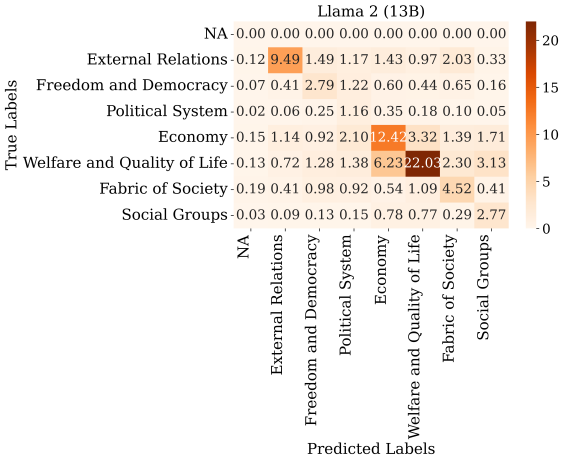


(e) Gemini 1 - Domains

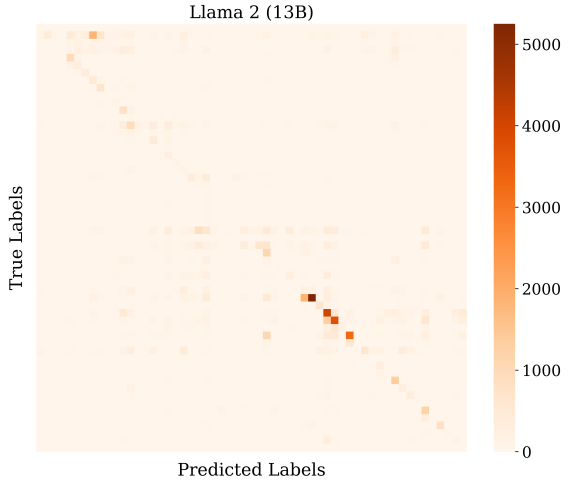


(f) Gemini 1 - Topics

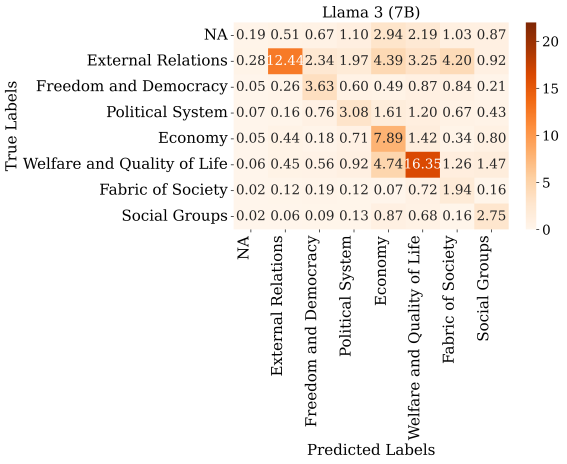
Figure 12: Confusion Matrices for Topics and Domains



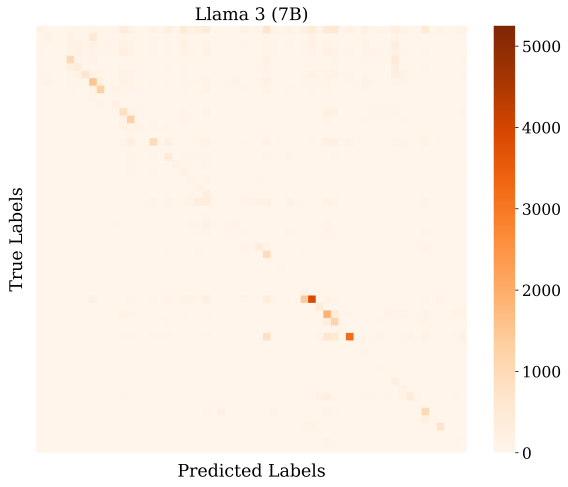
(a) Llama 2 - Domains



(b) Llama 2 - Topics



(c) Llama 3 - Domains



(d) Llama 3 - Topics

Figure 13: Confusion Matrices for Topics and Domains (Cont.)

6.3 Completeness of Data

The LLM models we ran to classify the quasi-sentences from the manifestos produced a considerable number of NA values. Thus, the models cannot match the given text with any categories or domains provided in the description. This may seem to be a problem with accuracy; however, we believe the cause is the subjectivity associated with the classification process. In fact, not all text segments found in the manifestos can clearly be associated with the given domains and categories. This is also an issue faced by human experts. As shown in the table below, completeness varies by the model used with Cohere (Command)

providing the least amount of 'NA' values and Llama3 providing the most. For examples on text segments that can potentially be associated with more than one topic, please check the Appendix.

Table 6: Level of Completeness

	ChatGPT 3.5 (Turbo)	Cohere (Command)	Gemini 1	Llama2	Llama3
Germany	83.91%	95.22%	79.4%	80.24%	62.93%
	(3.37%)	(1.28%)	(3.02%)	(13.12%)	(10.45%)
UK	79.75%	95.03%	86.26%	79.63%	63.25%
	(14.11%)	(1.78%)	(5.7%)	(14.22%)	(12.15%)
USA	83.51%	94.5%	82.06%	73.89%	62.34%
	(2.53%)	(1.11%)	(4.08%)	(24.7%)	(21.09%)

6.4 Descriptive Statistics

The next section provides information on the countries examined and the manifestos extracted. The table shows the names of the parties, party ideology established through public consensus (and extracted from Wikipedia), finalized ideology categories estimated based on the previous category, and finally the number of manifestos per decade as well as the total number of text segments extracted from the manifestos (shown in parentheses).

Table 7: Descriptive Statistics

	CMP ID	Party Ideology	Official Annotation	(1990-2000)	(2000-2010)	(2010-2020)	(2020-2023)
USA							
Democratic Party	61320	Left	Left	1	1	3	0
				(444)	(1000)	(5750)	.
Republican Party	61620	Right	Right	0	2	3	0
				.	(3143)	(6372)	.
UK							
Liberal Democrats	51421	Center	Centre to centre-left	1	0	3	0
				(972)	.	(4797)	.
Scottish National Party	51902	Left	Centre-left	1	1	3	0
				(824)	(868)	(3251)	.
Labour Party	51320	Left	Centre-left	0	1	3	0
				.	(1786)	(4273)	.
Green Party of England and Wales	51110	Left	Left	0	0	3	0
				.	.	(4013)	.
We Ourselves	51210	Left	Centre-left to left-wing	0	0	3	0
				.	.	(685)	.
Social Democratic and Labour Party	51340	Left	Centre-left	0	0	2	0
				.	.	(963)	.
Conservative Party	51620	Right	Centre-right to right-wing	0	0	3	0
				.	.	(4778)	.
Ulster Unionist Party	51621	Right	Centre-right	0	0	1	0
				.	.	(475)	.
The Party of Wales	51901	Left	Centre-left to left-wing	0	0	3	0
				.	.	(2378)	.
Democratic Unionist Party	51903	Right	Centre-right to right-wing	0	0	3	0
				.	.	(1344)	.
United Kingdom Independence Party	51951	Right	Right-wing to far-right	0	0	2	0
				.	.	(2942)	.
Alliance Party of Northern Ireland	51430	Center	Centre to centre-left	0	0	1	0
				.	.	(680)	.

Table 8: Descriptive Statistics [continued]

	CMP ID	Party Ideology	Official Annotation	(1990-2000)	(2000-2010)	(2010-2020)	(2020-2023)
Germany							
Alliance‘90/Greens	41113	Left	Centre-left	1	3	2	1
				(2354)	(7478)	(9777)	(4321)
Party of Democratic Socialism	41221	Left	Left-wing	1	1	0	0
				(1006)	(892)	.	.
The Left. Party of Democratic Socialism	41222	Left	Left-wing	0	1	0	0
				.	(615)	.	.
The Left	41223	Left	Left-wing	0	1	2	1
				.	(1701)	(6598)	(5209)
Social Democratic Party of Germany	41320	Left	Centre-left	1	3	2	1
				(1143)	(4992)	(5736)	(1776)
Free Democratic Party	41420	Right	Centre-right	1	3	2	1
				(1745)	(5858)	(5380)	(2681)
Christian Democratic Union/Christian Social Union	41521	Right	Centre-right	1	3	2	1
				(596)	(4276)	(4398)	(3121)
South Schleswig Voters’ Union	41912	Center	N/A	0	0	0	1
				.	.	.	(1533)
Pirates	41952	Center	N/A	0	0	1	0
				.	.	(2204)	.
Alternative for Germany	41953	Right	Far-right	0	0	2	1
				.	.	(1231)	(2073)

6.5 Description of Topics and Domains

The next section provides descriptions of the topics and domains shared in the documentation of the Comparative Manifesto Project. The descriptions provided by the project were also used in the prompts to associate each segment with a topic.

Table 9: Description of Topics and Domains

CMP Code	Topic	Definition
Domain 1	*External Relations*	
101	Foreign Special Relationships: Positive	Favourable mentions of particular countries with which the manifesto country has a special relationship; the need for co-operation with and/or aid to such countries.
102	Foreign Special Relationships: Negative	Negative mentions of particular countries with which the manifesto country has a special relationship.
103	Anti-Imperialism	Negative references to imperial behaviour and/or negative references to one state exerting strong influence (political, military or commercial) over other states.
104	Military: Positive	The importance of external security and defence.
105	Military: Negative	Negative references to the military or use of military power to solve conflicts.
106	Peace	Any declaration of belief in peace and peaceful means of solving crises – absent reference to the military.
107	Internationalism: Positive	Need for international co-operation, including co-operation with specific countries.
108	uropean Community/Union: Positive E	Favourable mentions of European Community/Union in general.
109	Internationalism: Negative	Negative references to international co-operation. Favourable mentions of national independence and sovereignty with regard to the manifesto country's foreign policy, isolation and/or unilateralism as opposed to internationalism.

[**Continued**]

Table 9: Description of Topics and Domains

CMP Code	Topic	Definition
110	European Community/Union: Negative	Negative references to the European Community/Union.
Domain 2	*Freedom and Democracy*	
201	Freedom and Human Rights	Favourable mentions of importance of personal freedom and civil rights in the manifesto and other countries.
202	Democracy	Favourable mentions of democracy as the "only game in town". General support for the manifesto country's democracy.
203	Constitutionalism: Positive	Support for maintaining the status quo of the constitution. Support for specific aspects of the manifesto country's constitution. The use of constitutionalism as an argument for any policy.
204	Constitutionalism: Negative	Opposition to the entirety or specific aspects of the manifesto country's constitution. Calls for constitutional amendments or changes.
Domain 3	*Political System*	
301	Decentralization	Support for federalism or decentralisation of political and/or economic power.
302	Centralisation	General opposition to political decision-making at lower political levels. Support for unitary government and for more centralisation in political and administrative procedures.

[Continued]

Table 9: Description of Topics and Domains

CMP Code	Topic	Definition
303	Governmental and Administrative Efficiency	Need for efficiency and economy in government and administration and/or the general appeal to make the process of government and administration cheaper and more efficient.
304	Political Corruption	Need to eliminate political corruption and associated abuses of political and/or bureaucratic power. Need to abolish clientelist structures and practices.
305	Political Authority	References to the manifesto party's competence to govern and/or other party's lack of such competence. Also includes favourable mentions of the desirability of a strong and/or stable government in general.
Domain 4	*Economy*	
401	Free Market Economy	Favourable mentions of the free market and free market capitalism as an economic model.
402	Incentives: Positive	Favourable mentions of supply side oriented economic policies (assistance to businesses rather than consumers).
403	Market Regulation	Support for policies designed to create a fair and open economic market.
404	Economic Planning	Favourable mentions of long-standing economic planning by the government.
405	Corporatism/Mixed Economy	Favourable mentions of cooperation of government, employers, and trade unions simultaneously. The collaboration of employers and employee organisations in overall economic planning supervised by the state.

[Continued]

Table 9: Description of Topics and Domains

CMP Code	Topic	Definition
406	Protectionism: Positive	Favourable mentions of extending or maintaining the protection of internal markets (by the manifesto or other countries).
407	Protectionism: Negative	Support for the concept of free trade and open markets. Call for abolishing all means of market protection (in the manifesto or any other country).
408	Economic Goals	Broad and general economic goals that are not mentioned in relation to any other category General economic statements that fail to include any specific goal.
409	Keynesian Demand Management	Favourable mentions of demand side oriented economic policies (assistance to consumers rather than businesses). Particularly includes increase private demand through • Increasin public demand; • Increasing social expenditures.
410	Economic Growth: Positive	The paradigm of economic growth.
411	Technology and Infrastructure: Positive	Importance of modernisation of industry and updated methods of transport and communication.
412	Controlled Economy	Support for direct government control of economy.
413	Nationalisation	Favourable mentions of government ownership of industries, either partial or complete; calls for keeping nationalised industries in state hand or nationalising currently private industries.
414	Economic Orthodoxy	Need for economically healthy government policy making.

[**Continued**]

Table 9: Description of Topics and Domains

CMP Code	Topic	Definition
415	Marxist Analysis	Positive references to Marxist-Leninist ideology and specific use of Marxist-Leninist terminology by the manifesto party
416	Anti-Growth Economy: Positive	Favourable mentions of anti-growth politics. Rejection of the idea that all growth is good growth. Opposition to growth that causes environmental or societal harm. Call for sustainable economic development.
Domain 5	*Welfare and Quality of Life*	
501	Environmental Protection	General policies in favour of protecting the environment, fighting climate change, and other "green" policies.
502	Culture: Positive	Need for state funding of cultural and leisure facilities including arts and sport.
503	Equality: Positive	Concept of social justice and the need for fair treatment of all people.
504	Welfare State Expansion	Favourable mentions of need to introduce, maintain or expand any public social service or social security scheme.
505	Welfare State Limitation	Limiting state expenditures on social services or social security. Favourable mentions of the social subsidiary principle
506	Education Expansion	Need to expand and/or improve educational provision at all levels.
507	Education Limitation	Limiting state expenditure on education.
Domain 6	*Fabric of Society*	
601	National Way of Life: Positive	Favourable mentions of the manifesto country's nation, history, and general appeals.

[Continued]

Table 9: Description of Topics and Domains

CMP Code	Topic	Definition
602	National Way of Life: Negative	Unfavourable mentions of the manifesto country's nation and history.
603	Traditional Morality: Positive	Favourable mentions of traditional and/or religious moral values.
604	Traditional Morality: Negative	Opposition to traditional and/or religious moral values.
605	Law and Order: Positive	Favourable mentions of strict law enforcement, and tougher actions against domestic crime Only refers to the enforcement of the status quo of the manifesto country's law code.
606	Civic Mindedness: Positive	Appeals for national solidarity and the need for society to see itself as united. Calls for solidarity with and help for fellow people, familiar and unfamiliar.
607	Multiculturalism: Positive	Favourable mentions of cultural diversity and cultural plurality within domestic societies. May include the preservation of autonomy of religious, linguistic heritages within the country including special educational provisions.
608	Multiculturalism: Negative	The enforcement or encouragement of cultural integration. Appeals for cultural homogeneity in society.
Domain 7	*Social Groups*	
701	Labour Groups: Positive	Favourable references to all labour groups, the working class, and unemployed workers in general. Support for trade unions and calls for the good treatment of all employees.
702	Labour Groups: Negative	Negative references to labour groups and trade unions. May focus specifically on the danger of unions 'abusing power'.

[Continued]

Table 9: Description of Topics and Domains

CMP Code	Topic	Definition
703	Agriculture and Farmers: Positive	Specific policies in favour of agriculture and farmers. Includes all types of agriculture and farming practises. Only statements that have agriculture as the key goal should be included in this category.
704	Middle Class and Professional Groups	General favourable references to the middle class.
705	Underprivileged Minority Groups	Very general favourable references to underprivileged minorities who are defined neither in economic nor in demographic terms.
706	Non-economic Demographic Groups	General favourable mentions of demographically defined special inter-General favourable mentions of demographically defined special interest groups of all kinds.

6.6 Full Prompt

To obtain the category classifications we wanted, we ran each text segment found in the manifestos using the prompt below.

> ***You are an expert in political science. You are asked to match sentences from a political party's manifesto with a topic. Here is a sentence from the manifesto: [Quasi-sentence from the party manifesto]. All topics and their explanations are provided below. Only provide one topic that best fits the sentence from the manifesto. Return only one topic, and nothing else.*** Topic: Foreign Special Relationships: Positive, Explanation: Favourable mentions of particular countries with which the manifesto country has a special relationship; the need for co-operation with and/or aid to such countries. Topic: Foreign Special Relationships: Negative, Explanation: Negative mentions of particular countries with which the manifesto country has a special relationship. Topic: Anti-Imperialism, Explanation: Negative references to imperial behaviour and/or negative references to one state exerting strong influence (political, military or commercial) over other states. Topic: Military: Positive, Explanation: The importance of external security and defence. Topic: Military: Negative, Explanation: Negative references to the military or use of military power to solve conflicts. References to the 'evils of war'. Topic: Peace, Explanation: Any declaration of belief in peace and peaceful means of solving crises – absent reference to the military. Topic: Internationalism: Positive, Explanation: Need for international co-operation, including co-operation with specific countries Topic: European Community/Union: Positive, Explanation: Favourable mentions of European Community/Union in general. Topic: Internationalism: Negative, Explanation: Negative references to international cooperation. Favourable mentions of national independence and sovereignty with regard to the manifesto country's foreign policy, isolation and/or unilateralism as opposed to internationalism. Topic: European Community/Union: Negative, Explanation: Negative references to the European Community/Union. Topic:

Freedom and Human Rights, Explanation: Favourable mentions of importance of personal freedom and civil rights in the manifesto and other countries. Topic: Democracy, Explanation: Favourable mentions of democracy as the "only game in town". General support for the manifesto country's democracy. Topic: Constitutionalism: Positive, Explanation: Support for maintaining the status quo of the constitution. Support for specific aspects of the manifesto country's constitution. The use of constitutionalism as an argument for any policy. Topic: Constitutionalism: Negative, Explanation: Opposition to the entirety or specific aspects of the manifesto country's constitution. Calls for constitutional amendments or changes. Topic: Decentralization, Explanation: Support for federalism or decentralisation of political and/or economic power. Topic: Centralisation, Explanation: General opposition to political decision-making at lower political levels. Support for unitary government and for more centralisation in political and administrative procedures. Topic: Governmental and Administrative Efficiency, Explanation: Need for efficiency and economy in government and administration and/or the general appeal to make the process of government and administration cheaper and more efficient. Topic: Political Corruption, Explanation: Need to eliminate political corruption and associated abuses of political and/or bureaucratic power. Need to abolish clientelist structures and practices. Topic: Political Authority, Explanation: References to the manifesto party's competence to govern and/or other party's lack of such competence. Also includes favourable mentions of the desirability of a strong and/or stable government in general. Topic: Free Market Economy, Explanation: Favourable mentions of the free market and free market capitalism as an economic model. Topic: Incentives: Positive, Explanation: Favourable mentions of supply side oriented economic policies (assistance to businesses rather than consumers). Topic: Market Regulation, Explanation: Support for policies designed to create a fair and open economic market. Topic: Economic Planning, Explanation: Favourable mentions of long-standing economic planning by the government. Topic: Corporatism/Mixed Economy, Explanation: Favourable mentions of cooperation of government, employers, and trade

unions simultaneously. The collaboration of employers and employee organisations in overall economic planning supervised by the state. Topic: Protectionism: Positive, Explanation: Favourable mentions of extending or maintaining the protection of internal markets (by the manifesto or other countries). Topic: Protectionism: Negative, Explanation: Support for the concept of free trade and open markets. Call for abolishing all means of market protection (in the manifesto or any other country). Topic: Economic Goals, Explanation: Broad and general economic goals that are not mentioned in relation to any other category. General economic statements that fail to include any specific goal. Topic: Keynesian Demand Management, Explanation: Favourable mentions of demand side oriented economic policies (assistance to consumers rather than businesses). Particularly includes increase private demand through • Increasing public demand; • Increasing social expenditures. Topic: Economic Growth: Positive, Explanation: The paradigm of economic growth. Topic: Technology and Infrastructure: Positive, Explanation: Importance of modernisation of industry and updated methods of transport and communication. Topic: Controlled Economy, Explanation: Support for direct government control of economy. Topic: Nationalisation, Explanation: Favourable mentions of government ownership of industries, either partial or complete; calls for keeping nationalised industries in state hand or nationalising currently private industries. Topic: Economic Orthodoxy, Explanation: Need for economically healthy government policy making. Topic: Marxist Analysis, Explanation: Positive references to Marxist-Leninist ideology and specific use of Marxist-Leninist terminology by the manifesto party Topic: Anti-Growth Economy: Positive, Explanation: Favourable mentions of anti-growth politics. Rejection of the idea that all growth is good growth. Opposition to growth that causes environmental or societal harm. Call for sustainable economic development. Topic: Environmental Protection, Explanation: General policies in favour of protecting the environment, fighting climate change, and other "green" policies. Topic: Culture: Positive, Explanation: Need for state funding of cultural and leisure facilities including arts and sport. Topic: Equality: Positive,

Explanation: Concept of social justice and the need for fair treatment of all people. Topic: Welfare State Expansion, Explanation: Favourable mentions of need to introduce, maintain or expand any public social service or social security scheme. Topic: Welfare State Limitation, Explanation: Limiting state expenditures on social services or social security. Favourable mentions of the social subsidiary principle Topic: Education Expansion, Explanation: Need to expand and/or improve educational provision at all levels. Topic: Education Limitation, Explanation: Limiting state expenditure on education. Topic: National Way of Life: Positive, Explanation: Favourable mentions of the manifesto country's nation, history, and general appeals. Topic: National Way of Life: Negative, Explanation: Unfavourable mentions of the manifesto country's nation and history. Topic: Traditional Morality: Positive, Explanation: Favourable mentions of traditional and/or religious moral values Topic: Traditional Morality: Negative, Explanation: Opposition to traditional and/or religious moral values. Topic: Law and Order: Positive, Explanation: Favourable mentions of strict law enforcement, and tougher actions against domestic crime. Only refers to the enforcement of the status quo of the manifesto country's law code. Topic: Civic Mindedness: Positive, Explanation: Appeals for national solidarity and the need for society to see itself as united. Calls for solidarity with and help for fellow people, familiar and unfamiliar. Topic: Multiculturalism: Positive, Explanation: Favourable mentions of cultural diversity and cultural plurality within domestic societies. May include the preservation of autonomy of religious, linguistic heritages within the country including special educational provisions. Topic: Multiculturalism: Negative, Explanation: The enforcement or encouragement of cultural integration. Appeals for cultural homogeneity in society. Topic: Labour Groups: Positive, Explanation: Favourable references to all labour groups, the working class, and unemployed workers in general. Support for trade unions and calls for the good treatment of all employees Topic: Labour Groups: Negative, Explanation: Negative references to labour groups and trade unions. May focus specifically on the danger of unions 'abusing power'. Topic: Agriculture and Farmers: Positive,

Explanation: Specific policies in favour of agriculture and farmers. Includes all types of agriculture and farming practises. Only statements that have agriculture as the key goal should be included in this category Topic: Middle Class and Professional Groups, Explanation: General favourable references to the middle class. Topic: Underprivileged Minority Groups, Explanation: Very general favourable references to underprivileged minorities who are defined neither in economic nor in demographic terms (e.g. the handicapped, homosexuals, immigrants, indigenous). Topic: Non-economic Demographic Groups, Explanation: General favourable mentions of demographically defined special interGeneral favourable mentions of demographically defined special interest groups of all kinds.