EuroCon: Benchmarking Parliament Deliberation for Political Consensus Finding

Zhaowei Zhang[1,2] ✉ **Minghua Yi**[3] **Mengmeng Wang**[2] **Fengshuo Bai**[4,5]
Zilong Zheng[2] **Yipeng Kang**[2*] ✉ **Yaodong Yang**[1*] ✉

[1]Institute for Artificial Intelligence, Peking University
[2]State Key Laboratory of General Artificial Intelligence, BIGAI
[3]Wuhan University [4]Shanghai Jiao Tong University [5]Zhongguancun Academy
zowiezhang.github.io/projects/EuroCon

Abstract

Achieving political consensus is crucial yet challenging for the effective functioning of social governance. However, although frontier AI systems represented by large language models (LLMs) have developed rapidly in recent years, their capabilities on this scope are still understudied. In this paper, we introduce *EuroCon*, a novel benchmark constructed from 2,225 high-quality deliberation records of the European Parliament over 13 years, ranging from 2009 to 2022, to evaluate the ability of LLMs to reach political consensus among divergent party positions across diverse parliament settings. Specifically, *EuroCon* incorporates four factors to build each simulated parliament setting: specific political issues, political goals, participating parties, and power structures based on seat distribution. We also develop an evaluation framework for *EuroCon* to simulate real voting outcomes in different parliament settings, assessing whether LLM-generated resolutions meet predefined political goals. Our experimental results demonstrate that even state-of-the-art models remain undersatisfied with complex tasks like passing resolutions by a two-thirds majority and addressing security issues, while revealing some common strategies LLMs use to find consensus under different power structures, such as prioritizing the stance of the dominant party, highlighting *EuroCon*'s promise as an effective platform for studying LLMs' ability to find political consensus.

1 Introduction

One of the fundamental prerequisites for effective social governance is establishing political consensus across diverse stakeholders [1–4]. From infrastructure development to welfare policies, consensus-building underpins the legitimacy [5] and implementation of collective decisions [6–8]. Yet, in pluralistic societies, conflicting values, power dynamics, and issue complexity render this process exceptionally challenging [9–12]. While large language models (LLMs) have shown promise in facilitating group discussions [13], supporting democratic deliberation [14–17], resolving regional conflicts [18], and analyzing ideological stances [19, 20], their capacity to find consensus in real and complex political scenarios remains underexplored. This gap raises a critical question: ***Can LLMs bridge divergent party divides to forge political consensus in real-world parliamentary settings?***

To study this problem, in this paper, we introduce *EuroCon*, a benchmark constructed from 2,225 real deliberation records of the European Parliament over a 13-year period ranging from 2009 to 2022, covering the full terms of the 7th and 8th Parliaments, as well as half of the 9th Parliament, which can evaluate the ability of LLMs to reach political consensus among various party positions within the rich context of parliamentary settings.

Specifically, *EuroCon* has designed four adjustable factors to construct different simulated parliaments, which are: (1) **Political issues**: the political problems to be discussed and their topic classification, (2) **Political goals**: the criteria for meeting political consensus, (3) **Participating parties**: different

*Corresponding author.

Preprint. Under review.

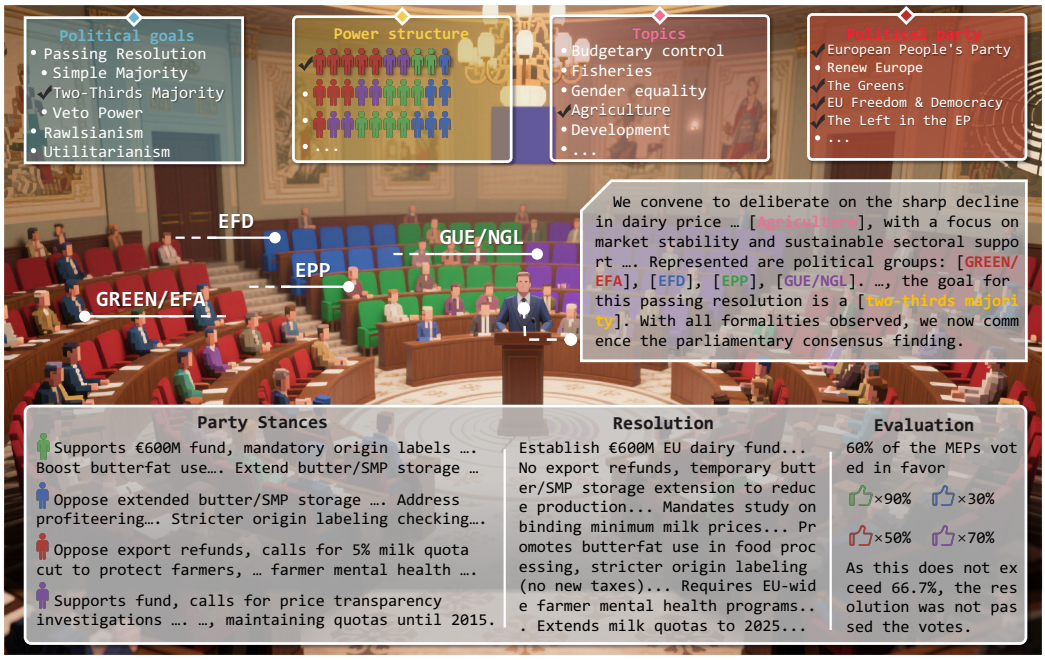


Figure 1: **An example scenario in *EuroCon***. In each task, *EuroCon* constructs a simulated parliament with varying political goals, power structures, issues, and participating parties. The tested LLM then attempts to find political consensus based on the parliament's setup and the parties' divergent positions. The outcome is evaluated via a simulated voting by *EuroCon*'s evaluation framework.

numbers of parties with varying stances involved in the parliament, and (4) **Parliamentary power structures**: differences in influence and discourse power of each party due to their number of seats. By combining these settings, we have constructed a total of 28,620 different parliamentary scenarios. To assess whether LLM-generated resolutions meet the corresponding political goals, we further developed an open-ended evaluation framework in *EuroCon* based on GPT-4o mini. Through our experiments, we have verified its strong capability to simulate the real voting results, thereby allowing the effective evaluation in *EuroCon* (subsection 4.1).

We illustrate one of the *EuroCon*'s test scenarios in Figure 1. The upper part presents the setting of the current simulated parliament. The seating colors in the figure represent the seat distribution among the four participating parties, which are GREEN/EFA (red, 50%), EPP (green, 20%), GUE/NGL (purple, 20%), and EFD (blue, 10%). The lower part demonstrates the LLM's consensus-finding process. The parliamentary president announced the need to discuss the issue of surplus dairy products and introduced the political goal is to passing the resolution with a two-thirds majority among the members of the European Parliament (MEPs).

Subsequently, each participating party expresses inconsistent positions on this issue. For example, EPP and EFD have significant disagreements on the matter of extending storage time, while GREEN/EFA and GUE/NGL have differences over export refunds. Although the resolution generated by the evaluated LLM partially considered the apportionment of seats among different parties to balance conflicting positions, it still failed to reconcile a new consensus resolution beyond the compromise. As a result, in *EuroCon*'s evaluation, 50%, 90%, 70%, and 30% of the MEPs from GREEN/EFA, EPP, GUE/NGL, and EFD vote in favor of the resolution, respectively. Considering the seat distribution, only 60% of the entire parliament voted in favor of the resolution, which does not meet the two-thirds majority standard, and thus the resolution was not passed.

We perform a comprehensive evaluation using *EuroCon* in six representative LLMs, revealing notable variations in their ability to find political consensus (subsection 4.2). While most LLMs perform well on simple majority tasks, they struggle with more difficult challenges, such as passing resolutions with a two-thirds majority or addressing security issues (subsection 4.3). Furthermore, our analysis uncovers several common strategies that LLMs employ to achieve political consensus (subsection 4.4).

In summary, this paper makes four major contributions. **Firstly**, we conduct a large-scale scraping and thorough cleaning of a vast amount of European Parliament deliberation records, compiling

2,225 high-quality complete parliamentary records. **Secondly**, we define the problem of evaluating LLMs' ability to find political consensus and construct *EuroCon* based on these records. **Thirdly**, we develop an open-ended evaluation framework that can simulate the proportion of MEPs who vote in favor in each party. **Lastly**, we demonstrate that *EuroCon* can well assess the LLMs' ability to find political consensus, highlighting its promise as an effective research platform for the realm.

2 Related Work

Political consensus finding, due to its realistic and complex scenarios, the conflict of stances and values, and the need to consider diverse power structures, differs from existing works that primarily consider conversational grounding [21–25] and game-theoretic bargaining [26–32], becoming a novel and challenging problem. To our knowledge, there are currently no studies that construct a benchmark to evaluate LLMs' ability to find political consensus, but there are some works that have explored LLMs for democratic deliberation and benchmarks in political settings. Below, we will introduce these two aspects separately.

LLMs for Democratic Deliberation. The powerful text generation and information processing capabilities of LLMs have led some studies to explore how they can accelerate the process of democratic deliberation. Konya et al. [33] design a pipeline allowing LLMs to participate in every stage of democratic elections, aiding in extracting and summarizing complex texts to improve decision-making efficiency. Fish et al. [15] utilizes LLMs' generative abilities to synthesize a set of opinions most satisfactory to the majority based on survey results about chatbot personalization. Small et al. [14] apply LLMs to the deliberation platform Polis, finding that LLMs enhance efficiency but still pose unresolved risks. Bakker et al. [34], Tessler et al. [16] fine-tune LLMs to repeatedly generate and refine statements representing a group's collective stances on social or political issues.

Benchmarks in Political Settings. LLMs have been widely applied to political science tasks [35]. However, political science covers a wide range of research questions, resulting in diverse benchmarks. Kornilova and Eidelman [36], Arregui and Perarnaud [37], KlÃijver et al. [38], Shu et al. [39] provide data on texts and the ideologies of their associated political parties, which are used for semantic analysis of texts covering different ideologies. Garzia et al. [40], Vamvas and Sennrich [41] extensively collect public comments on various political issues in Europe to study the positioning and classification of political positions. Kornilova and Eidelman [36], Shu et al. [39], Arregui and Perarnaud [37] provide a large collection of legal text data from the United States and Europe, facilitating research in the generation and summarization of legal documents. POLCA [42] collects party statements and final agreements from several European countries, providing a benchmark to evaluate whether LLMs can determine if a statement is likely to appear in the final agreement. Stammbach et al. [43], Chalkidis and Brandl [44], Batzner et al. [45] investigate whether LLMs have intrinsic political bias and explore the impact of fine-tuning and prompting on their political stance. Liang et al. [46] constructs a benchmark based on the United Nations resolution process to evaluate whether LLMs can accurately capture the political stances of member states, simulate voting, and emulate delegate speeches. Although these works offer benchmarks for political science research, their focus is not on studying the ability of LLMs to find political consensus.

3 *EuroCon* Benchmark

How to build a benchmark that can effectively evaluate the political consensus finding capabilities of LLMs is a significant challenge. To address this problem, we seek a benchmark that satisfies the following requirements: (1) **Authenticity**: All data must come from real political scenarios; (2) **Conflict**: Under each political issue, there must be a varying number of parties holding different opinions; (3) **Diverse Power Structures**: The generated political consensus need to consider the impact of different parties; (4) **Various Political Goals**: The state when political consensus is achieved; (5) **Open-ended Evaluation**: It should be capable of automatically evaluating the quality of the political consensus generated by LLMs in scenarios that meet the above requirements.

In the following sections, we will provide a detailed introduction to *EuroCon* and how it addresses the above challenges. In subsection 3.1, we will describe the data collection and processing procedures of *EuroCon*, explaining why it meets the criterion of authenticity and open-ended evaluation. In subsection 3.2, we will explain why it addresses the remaining challenges by detailing the different task definitions and how we construct these tasks based on the collected data.

3.1 Data Collection Procedure

We conduct a large-scale scraping and combine data sourced from the official website of the European Parliament[2] , HowTheyVote[3] , and the VoteWatch Europe dataset [47], to obtain a comprehensive collection of parliamentary records from the European Parliament spanning a 13-year period from 2009 to 2022. This dataset covers the full terms of the 7th and 8th Parliaments, as well as half of the 9th Parliament, and includes detailed information on issues, topics, debates, resolutions, and votes.

Unlike previous datasets that were also collected from the European Parliament or political parties [48, 47, 44, 42], we (1) do not just scrape a single aspect of the parliamentary process, such as debates [44] or votes [47], but instead collect all information corresponding to each issue from different sources separately, further aligning and integrating them more comprehensively, including information on issues, topics, debates, resolutions, and votes. (2) We perform additional cleaning and post-processing on the data to enhance its quality and readability. (3) The cleaned voting and resolution data can serve as the basis for our open-ended evaluation, allowing further verification of whether our designed evaluation framework aligns with real-world voting outcomes. These contributions not only enhance the quality and diversity of our data but also allow the data to transcend the scope of a single task (such as being used solely for text translation [48]) and further enable the construction of various complex political tasks and scenarios in *EuroCon*. We will introduce them one by one as follows:

Data Collection. We first match the URL provided for each issue's voting information in the VoteWatch Europe dataset with the corresponding issue URLs on the European Parliament's official website and HowTheyVote. This allows us to obtain the issue and resolution content corresponding to each voting record. We further match the resolution with the debate URL on the European Parliament's website using the issue name, enabling us to scrape the corresponding debate information. In this way, we obtain 30,698 raw parliamentary records. However, since many records were incomplete or duplicated, we further refine the data, retaining only those where the final vote was confirmed to be finished and all information was complete. The detailed filtering steps are provided in Appendix A.1. Furthermore, we classify all collected data by referring to the topics defined in the VoteWatch Europe dataset [47] and classify these complete parliamentary records into 5 coarse- and 19 fine-grained topics (detailed in Figure 2), such as "culture & education", "agriculture", "international trade", etc. Through this approach, we integrate different pieces of information on the same issue from various sources, ultimately selecting 2,225 complete, high-quality raw data entries, ensuring that each data entry contains a quintuple of raw information: (issue, topic, debates, resolution, votes).

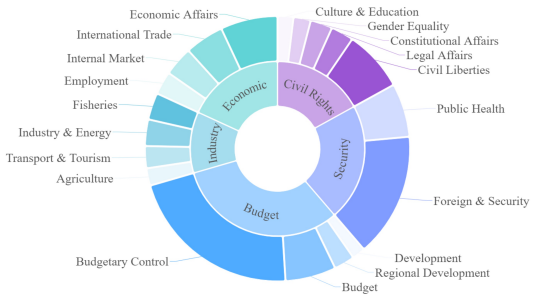
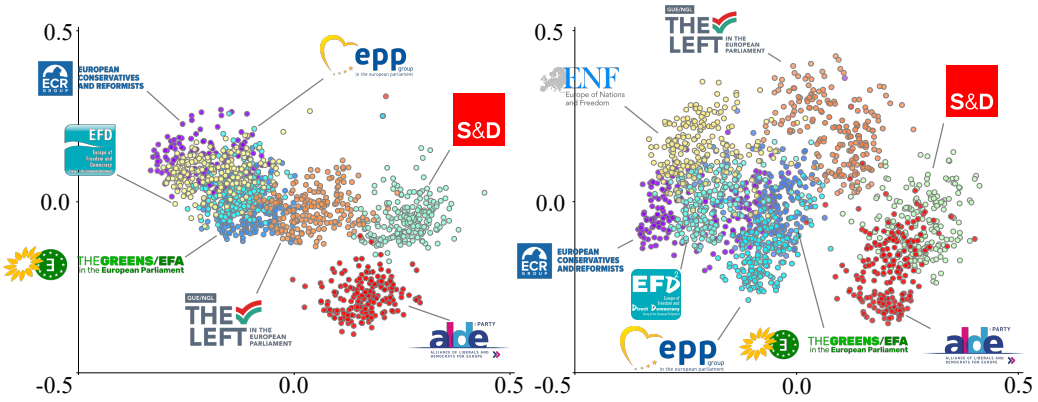


Figure 2: The 5 coarse-grained and 19 fine-grained topic categories of issues in *EuroCon*, whose definitions can be found in Appendix B.1. The shade of the color indicates the proportion of the fine-grained topic within the coarse-grained topic; the darker the color, the higher the proportion.

Data Cleaning and Post-Processing. To address raw data redundancy, we employ DeepSeek-R1 [49] and rule-based methods for cleaning and post-processing. DeepSeek-R1 is used to organize resolutions, removing redundancies while maintaining format, and summarizing parliamentary discussion background based on issue, resolution, and debate information. Voting data is processed by matching each member with their party and calculating party voting results by rounding down the proportion of MEPs within the party who voted in favor to an integer between 0 and 9. DeepSeek-R1 then summarizes party stances from debate data, removing parties without expressed stances. Rule-based methods randomly replace words to diversify data, adjusting stances on resolutions to issues. This results in cleaned sextuples of (issue, topic, background, stances, resolution, votes) containing relevant party information. Further details of the post-processing procedure and the specific prompts can be found in Appendix A.6 and Appendix C.1.

[2]https://www.europarl.europa.eu

[3]https://howtheyvote.eu



(a) Party Stances for the 7th Parliament

(b) Party Stances for the 8th Parliament

Figure 3: Semantic representation distribution of party stances (indicated by their symbols) in the 7th (2009-2014) and 8th (2014-2019) terms of the European Parliament in *EuroCon*.

Constructing the Open-ended Evaluation. Based on the sextuple data, we can perform the open-ended evaluation for each party's voting results on each issue by inputting the background of the issue, each party's stances, and the resolution generated by the evaluated LLM. This results in a scalar score between 0 and 9, indicating the proportion of the MEPs within the party voting in favor.

We define the n parties participating in each issue as $P = \{p_1, p_2, \ldots, p_n\}$. For each party p_i, its stance is represented as s_i. The corresponding voting score u_i for the party can be calculated using $u_i = LLM(\cdot \mid \text{background}, s_i, \text{resolution}), u_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. We use GPT-4o mini [50] to implement this evaluation process. Compared to GPT-4o, this is a lighter and more cost-effective model. In subsection 4.1, we verified that it is more suitable for our evaluation tasks than GPT-4o, and confirmed that its simulation of voting results is consistent with real parliamentary voting data. The specific evaluation prompt can be found in Appendix C.3.

3.2 Task Settings

After collecting and cleaning the raw data from the European Parliament, we further expand and organize these data to construct different task settings for each issue in *EuroCon*. These settings are designed to meet the evaluation needs of conflict, diverse power structures, and various political goals. In the following paragraphs, we will introduce each aspect separately:

Participating Political Parties. The core requirement of conflict is to have a different number of parties with various positions on each issue. There are clear differences in political positions among the parties in the European Parliament [51–53]. To demonstrate this point more obviously, we randomly sample 200 stance data points from each party during the 7th and 8th parliamentary terms. We then use OpenAI's text-embedding-003-small [54] model to map each party's stances into a semantic representation space, and employ Principal Component Analysis (PCA) [55] to visualize this information. As shown in Figure 3, the stances of each party form distinct clusters in the semantic space, with significant differences (detailed in Appendix B.2). We further design three different settings for the number of participating parties in *EuroCon*: 2, 4, and 6. For each issue, we select the corresponding number of parties with the highest voting variance to enhance the conflict. Since not all parties participated in every parliamentary discussion, the total number of tasks under this setting varies slightly depending on the number of parties involved.

Power Structure. One major challenge of finding political consensus is dealing with complex power structures. To more accurately simulate different parliamentary scenarios in reality, we allocate seats to each participating party in the current parliament scenario to demonstrate their political influence. We define the calculation of the total votes in favor MEP number u in this setting as:

$$u = \sum_{i=1}^{n} w_i u_i, \tag{1}$$

where w_i represents the proportion of seats occupied by party p_i in the parliament, satisfying $\sum_{i=1}^{n} w_i = 1$ and $\forall w_i \geq 0$.

In constructing the tasks, we randomly assign each party's seats in the parliament, and only assign one time for each task, which means the party seat in our task may not align with reality. For instance, in the 9th European Parliament, the EPP is the largest party, holding 24% of the total seats, while GUE/NGL is the smallest party, holding only 5%. However, in our setup, GUE/NGL might become the largest party with over 80% of the seats. This approach not only enriches our task settings but also helps mitigate the risk of data leakage and prevents the tested LLMs from using real-world prior knowledge about party seats to cheat.

Voting Mechanism. We refer to the European Parliament and the United Nations Security Council (UNSC) to set three common voting mechanisms, which are: (1) **Simple Majority**: A resolution needs to be voted through by more than 50% of the parliamentary seats. We define the boolean variable $v \in \{0, 1\}$ to indicate whether the resolution will be passed under this voting mechanism. In the setting of simple majority, $v = 1$ only if $u \geq 5$. (2) **Two-thirds Majority**: A resolution needs to be voted through by more than two-thirds of the parliamentary seats. In this setting, $v = 1$ only if $u \geq 6.67$. (3) **Veto Power**: In the UNSC, permanent members have veto power [56] . In our setting, the tested LLM needs to generate a resolution that can be passed by a simple majority of MEPs in the parliament and not be rejected by the vetoing party (in favor rate under 60%). In this setting, $v = 1$ only if $u \geq 5$ and $u_k \geq 6$, where u_k is the voting score of the vetoing party. In the actual process of constructing the task, we randomly designate which political party has the veto power.

Political Goals. In a parliament, there are often different political goals and tasks, which lead to various definitions of when the political consensus is found. In *EuroCon*, we define three different parliamentary political goals, as follows: (1) **Passing a Resolution**: This is the most common parliamentary goal, aimed at finding a political consensus that can be passed under a specific power structure and voting mechanism detailed above. (2) **Rawlsianism**: Following the Rawlsian principle [57], the party with the least current benefits receives the most attention. The political goal in this context is to formulate a resolution that maximizes the benefits for the party with the least benefits. In this setting, $u = \min_{i \in n}(u_i)$. (3) **Utilitarianism**: Following the Utilitarian principle [58], the political goal is to formulate a resolution that maximizes the sum of benefits for all parties. Under this setting, $u = \sum_{i=1}^{n} u_i$.

It is worth noting that, in our defined political goals, only the passing resolution setting requires different voting mechanisms and corresponding power structures, which return a boolean variable indicating whether a vote passes. For Rawlsianism and Utilitarianism, only the corresponding voting score needs to be considered. Therefore, by combining different power structures, voting mechanisms, and political goals, we establish five distinct settings: Passing Simple Majority (SM), Passing Two-Thirds Majority (2/3M), Passing Veto Power (VP), Rawlsianism (Rawls), and Utilitarianism (Util). These can further be combined with three party number configurations (2, 4, or 6 parties), resulting in 15 task settings. Since each data record we collected represents an independent political issue, our framework can construct 28,620 distinct political scenarios altogether.

4 Experiments

In this section, we use *EuroCon* to conduct comprehensive experiments to evaluate six current representative LLMs, specifically on their political consensus finding capability. We selected two commercial models: GPT-4o [59] and Gemini-2.5-Flash (Gemini-2.5) [60], as well as four open-sourced models from different vendors and with varying parameters: Qwen2.5-32B-Instruct (Qwen2.5-32B), Qwen2.5-72B-Instruct (Qwen2.5-72B) [61, 62], Llama-3.3-70B-Instruct (Llama-3.3-70B) [63], and 671-billion-parameter DeepSeek-R1 [49]. All LLMs are set up with standardized inference settings, including a temperature of 0.7 and top-p sampling of 0.95. For Gemini-2.5 and DeepSeek-R1, we use their thinking versions.

In the following subsections, we will demonstrate how *EuroCon* can be used to investigate these four questions: (1) Can our evaluation framework simulate the voting results well? (2) How does the performance differ among various LLMs and (3) parliament settings and issue topics? (4) What common strategies do LLMs find political consensus under different power structures?

4.1 Can GPT Simulate the Voting Results for the Political Parties?

As mentioned in subsection 3.1, to achieve open-ended evaluation, we introduced GPT-4o mini and required it to output an integer scalar between 0 and 9 to simulate the percentage of MEPs from each

Table 1: The expected Acc1, PCC, and the p-value between GPT-4o (mini) simulation score and the actual parliamentary voting results in the 7th, 8th, 9th parliament terms, and all data.

Models	7th		8th		9th		All		p-value
	Acc1	PCC	Acc1	PCC	Acc1	PCC	Acc1	PCC	
GPT-4o	0.46	0.60	0.60	0.77	0.66	0.79	0.57	0.73	0.00
GPT-4o mini	0.62	0.83	0.71	0.91	0.78	0.90	0.70	0.88	0.00

political party who voted in favor. However, before that, we need to answer two key questions: Why do we choose GPT-4o mini, and can GPT-4o mini simulate the voting results well?

GPT-4o [59] and GPT-4o mini [50] are powerful models launched by OpenAI, both achieving outstanding results in various general benchmarks such as MMLU [64], MGSM [65], and HumanEval [66]. Although GPT-4o mini is slightly at a disadvantage compared to GPT-4o on these general benchmarks, it offers advantages such as higher cost-efficiency and faster computation speed. This prompted us to conduct the following experiments to verify whether it can adequately replace GPT-4o in our evaluation tasks.

Due to the variation in political parties across different parliament terms, we randomly sample 100 issues for each party under each topic in the 7th, 8th, and 9th terms for testing. We use GPT-4o and GPT-4o mini to calculate the current party's voting approval rate conditioned on actual resolutions, and compute the expected consistency between the simulated results and the real parliamentary party voting outcomes across all topics and parties for each term and all terms combined. Specifically, we calculate the proportion of prediction errors within ± 1 (Accuracy within ± 1, Acc1), the Pearson correlation coefficient (PCC), and its p-value, with the results shown in Table 1. It is noteworthy that due to our large sample size, the p-value is close to 0, eliminating sampling interference in the results. From the results, it can be observed that although both GPT-4o and GPT-4o mini exhibit high consistency with the real data in this task, GPT-4o mini demonstrates better performance, validating that GPT-4o mini is sufficient to replace GPT-4o in our task.

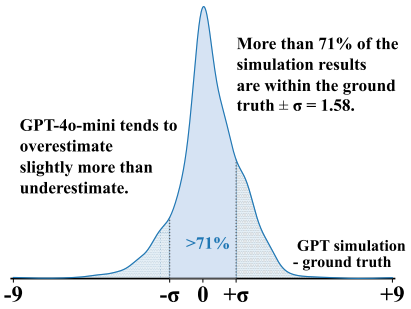


Figure 4: The error distribution between GPT-4o-mini's simulation and the ground truth voting results.

Additionally, referring to existing work [28], we plotted Figure 4 to further illustrate the distribution of computational errors for GPT-4o mini. The error is calculated by subtracting the ground truth voting score from the simulated score of GPT-4o mini. It can be observed that the majority of the simulation results (>71%) are centered within the standard deviation σ (± 1.58) around the real voting results, with more simulation results showing a slight overestimation than underestimation. Based on the above experiments, we answered the question raised in the caption of the subsection with GPT-4o mini is sufficiently capable of simulating each party's voting results for the current resolution. More detailed experimental results can be found in Appendix D.1.

4.2 Performance Analysis for Various LLMs

We utilize the *EuroCon* evaluation framework described in subsection 4.1 to assess the performance of six LLMs on *EuroCon*. The results are depicted in Table 2, which presents the average scores across all our 15 task settings described in subsection 3.2. For the SM, 2/3M, and VP, the scores represent the average passing rates ranging from 0 to 1. For Rawls and Util, the scores represent the average results obtained from the corresponding calculation methods, ranging from 0 to 9. All these metrics are higher-the-better.

We find that **Qwen-72B** and **Deepseek-R1** perform the best. Qwen-72B demonstrates exceptional performance, surpassing models of similar scale and even commercial models such as GPT-4o and Gemini, as well as the larger Deepseek-R1. This finding is consistent with some results from existing work [62], which suggests that a model's ability in specific tasks is not entirely directly related to the number of parameters but should instead focus more on the task-specific capabilities.

Table 2: Performance of different LLMs on *EuroCon*. The values in square brackets indicate the range of each metric, and all metrics follow the principle that higher values are better. The background color of the table cells deepens as the performance improves. The blue color scheme represents metrics in the 0-1 range, while the red color scheme represents metrics in the 0-9 range.

Model	SM [0-1] ↑			2/3M [0-1] ↑			VP [0-1] ↑			Rawls [0-9] ↑			Util [0-9] ↑		
	2	4	6	2	4	6	2	4	6	2	4	6	2	4	6
Qwen2.5-32B	0.63	0.64	0.76	0.37	0.38	0.45	0.43	0.45	0.57	2.70	2.28	1.66	5.00	5.60	5.61
Llama-3.3-70B	0.64	0.63	0.67	0.39	0.38	0.34	0.44	0.45	0.48	3.12	2.02	1.22	5.76	5.81	5.75
GPT-4o	0.75	0.72	0.71	0.48	0.45	0.43	0.55	0.51	0.53	4.05	2.61	1.81	6.27	5.96	5.56
Gemini-2.5	0.72	0.79	0.87	0.50	0.55	0.61	0.53	0.61	0.69	3.91	3.14	2.50	6.19	6.49	6.72
DeepSeek-R1	0.83	0.86	**0.90**	0.59	**0.63**	**0.66**	0.65	**0.67**	0.71	4.85	3.84	3.01	6.69	6.82	6.92
Qwen2.5-72B	**0.86**	**0.87**	**0.90**	**0.62**	0.61	0.62	**0.67**	**0.67**	**0.72**	**5.12**	**4.00**	**3.26**	**6.88**	**6.97**	**6.98**

We also compare the performance differences among other evaluated LLMs and identify the following trends: (1) Thinking models like Deepseek-R1 and Gemini-2.5 generally outperform no-thinking models like Llama-3.3-70B and Qwen2.5-32B. (2) Commercial models (Deepseek-R1, Gemini-2.5, GPT-4o) typically outperform non-commercial open-source models (Llama-3.3-70B and Qwen2.5-32B). (3) The minimal differences between Qwen2.5-32B and Llama-3.3-70B may further suggest that the task of political consensus finding is not strongly correlated with model size. For detailed case study comparing the outputs of different models, see Appendix E.1.

4.3 Performance Analysis for Different Parliament Settings and Issue Topics

In this section, we demonstrate how different parliament settings and issue topics in *EuroCon* influence LLMs' ability to find political consensus, which are presented separately as follows:

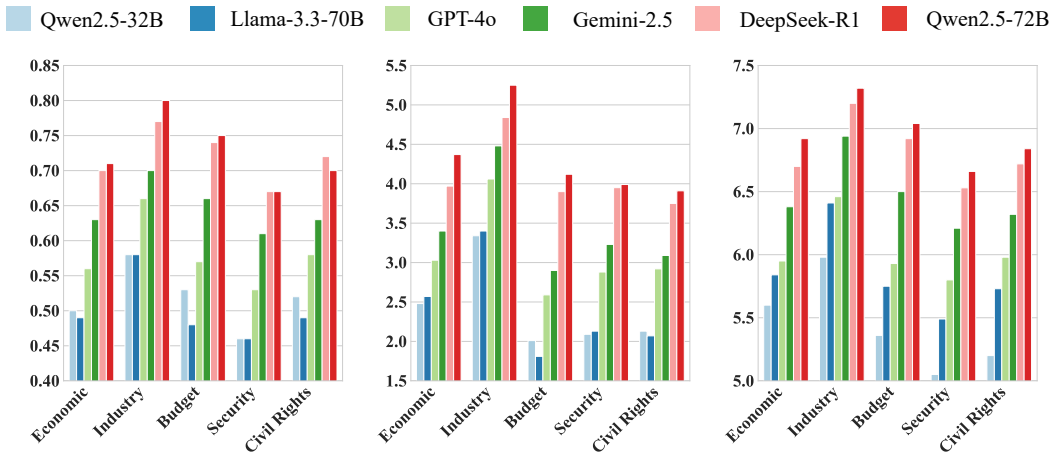
Analysis for Different Parliament Settings. As shown in Table 2, for the political goal of passing a resolution, SM is the simplest, and most models can perform well. However, in the 2/3M and VP settings, model performance declines significantly, indicating that the capabilities of existing LLMs generally lie in the gap between the increased difficulty of SM and these two settings. We further find that as the number of parties increases, the results of most models gradually rise. This could be due to our task construction prioritizing parties with the most diverse positions, complicating reconciliation with fewer parties. For the Rawls objective, however, the success rate of models decreases as the party number increases. This aligns with the task's definition, as the more participants there are, the harder it becomes to avoid neglecting any party's interests, presenting a significant challenge for current LLMs in this task. Concrete case demonstrations are given in Appendix E.2.

Analysis for Different Issue Topics. As shown in Figure 5, we analyze the experimental results of five coarse-grained topics. These results suggest that the difficulty of different topics shows certain similarities across various parliamentary settings. Specifically, topics involving policies, such as Security and Civil Rights, tend to be more challenging than those related to industrial development. This may be because these topics tend to present more complex and conflicting positions, requiring the evaluated LLM to possess stronger reasoning capabilities. For the complete experimental results of each fine-grained topic, see Appendix D.2.

Our experimental results successfully reveal the limitations of the current LLMs in political consensus finding. Although top-performing models like Qwen2.5-72B achieve a success rate of 86-90% in SM scenarios, their performance significantly drops when faced with stricter consensus requirements. In 2/3M tasks, the success rate falls to 61-62%, and in the more challenging Rawls setting, it ranges from only 3.26-5.12. Additionally, when dealing with more complex topics such as security, these models still face considerable challenges.

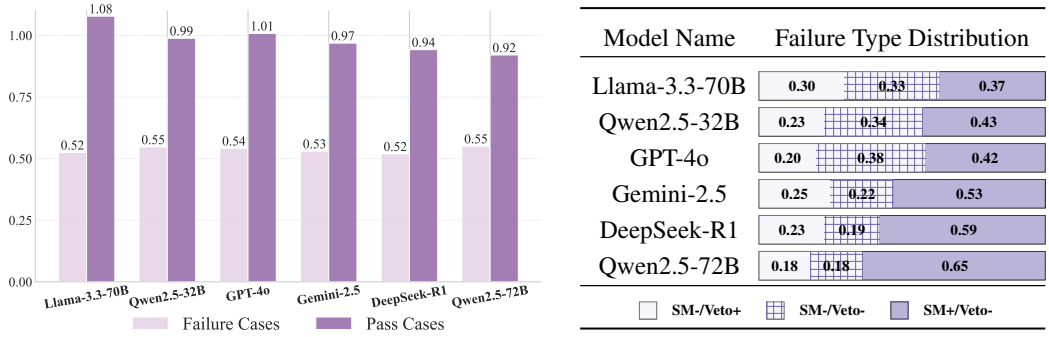
4.4 Strategies for Political Consensus Finding under Different Power Structures

As shown in Figure 6, we analyze whether a common strategy exists for LLMs to achieve political consensus under various power structures, excluding two-party scenarios. Our findings are as follows: (1) Under both simple majority and two-thirds majority systems, successful proposals often rely on the support of the largest party, indicating that dominant parties' votes are foundational for approval



(a) PR Analysis (b) Rawls Analysis (c) Util Analysis

Figure 5: The average results of the six evaluated LLMs of the five coarse-grained topics on passing resolution (PR, including SM, 2/3M, and VP), Rawls, and Util political goals.



(a) SM & 2/3M Analysis (b) VP Failure Cases Analysis

Figure 6: Strategy analysis of LLMs under different power structures. Figure (a) shows the average contribution ratio of the largest party to other parties in failed and passed cases across SM and 2/3M. Figure (b) shows the proportion of not passed SM but not vetoed (SM-/Veto+), not passed SM and vetoed (SM-/Veto-), and passed SM but vetoed (SM+/Veto-) among all failed cases in the VP setting.

and decisive in most cases. (2) Regarding the veto mechanism, models with higher passing rates experience more failures due to vetoes after simple majority approval. This suggests a strategic trade-off: prioritizing majority party support can maximize approval chances but risks overlooking veto-holding parties, leading to sudden failures when veto power is exercised.

These findings highlight *EuroCon*'s unique ability to reveal subtle flaws in the LLMs' political decision-making capabilities, which existing negotiation benchmarks [26, 32] are often hard to detect.

5 Conclusion

In this work, we introduced *EuroCon*, a novel benchmark constructed from 2,225 European Parliament deliberation records to evaluate LLMs' ability to find political consensus across diverse parliamentary settings. Our framework incorporates key factors like political issues, goals, party stances, and power structures, with a GPT-4o mini-based evaluation system that effectively simulates real voting outcomes. While our experiments demonstrate *EuroCon*'s promising evaluation capabilities, we acknowledge limitations, including potential biases in LLM-processed data and considering each political party's positions as a whole (discussed in Appendix G). To our knowledge, *EuroCon* represents the first comprehensive benchmark for assessing political consensus finding capabilities in LLMs, offering both a valuable evaluation platform and new insights into how LLMs navigate complex governance scenarios. We believe this work opens important directions for research at the intersection of AI and political deliberation.

References

[1] James W Prothro and Charles M Grigg. Fundamental principles of democracy: Bases of agreement and disagreement. *The Journal of politics*, 22(2):276–294, 1960.

[2] Arend Lijphart et al. *Patterns of democracy: Government forms and performance in thirty-six countries*. Yale university press, 1999.

[3] Robert Huckfeldt, Paul E Johnson, and John Sprague. *Political disagreement: The survival of diverse opinions within communication networks*. Cambridge University Press, 2004.

[4] John Rawls. Political liberalism. In *The new social theory reader*, pages 123–128. Routledge, 2020.

[5] Joshua Cohen. Deliberation and democratic legitimacy. In *Debates in contemporary political philosophy*, pages 352–370. Routledge, 2005.

[6] J. Citrin. Conflict/consensus. In Neil J. Smelser and Paul B. Baltes, editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 2547–2550. Pergamon, Oxford, 2001. ISBN 978-0-08-043076-8. doi: https://doi.org/10.1016/B0-08-043076-7/01115-3. URL `https://www.sciencedirect.com/science/article/pii/B0080430767011153`.

[7] William R Potapchuk and Jarle Crocker. Implementing consensus-based agreements. In *Multi-Party Dispute Resolution, Democracy and Decision-Making*, pages 429–457. Routledge, 2017.

[8] Sali Shehu. Political consensus: A crucial and key element of political organization. *Academicus International Scientific Journal*, 8(15):80–90, 2017.

[9] Howard Raiffa. *The art and science of negotiation*. Harvard University Press, 1982.

[10] Harri Ehtamo, Markku Verkama, and Raimo P Hamalainen. How to select fair improving directions in a negotiation model over continuous issues. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 29(1):26–33, 1999.

[11] Lawrence E Susskind, Sarah McKearnen, and Jennifer Thomas-Lamar. *The consensus building handbook: A comprehensive guide to reaching agreement*. Sage publications, 1999.

[12] Zachary R Baker and Zarif L Azher. Simulating the us senate: An llm-driven agent approach to modeling legislative behavior and bipartisanship. *arXiv preprint arXiv:2406.18702*, 2024.

[13] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. Enhancing ai-assisted group decision making through llm-powered devil's advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 103–119, 2024.

[14] Christopher T Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. Opportunities and risks of llms for scalable deliberation with polis. *arXiv preprint arXiv:2306.11932*, 2023.

[15] Sara Fish, Paul Gölz, David C Parkes, Ariel D Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. *arXiv preprint arXiv:2309.01291*, 2023.

[16] Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, et al. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719): eadq2852, 2024.

[17] Daniel Jarrett, Miruna Pislar, Michiel A Bakker, Michael Henry Tessler, Raphael Köster, Jan Balaguer, Romuald Elie, Christopher Summerfield, and Andrea Tacchetti. Language agents as digital representatives in collective decision-making. *arXiv preprint arXiv:2502.09369*, 2025.

[18] Andrew Konya, Luke Thorburn, Wasim Almasri, Oded Adomi Leshem, Ariel D Procaccia, Lisa Schirch, and Michiel A Bakker. Using collective dialogues and ai to find common ground between israeli and palestinian peacebuilders. *arXiv preprint arXiv:2503.01769*, 2025.

[19] Kai Chen, Zihao He, Jun Yan, Taiwei Shi, and Kristina Lerman. How susceptible are large language models to ideological manipulation? *arXiv preprint arXiv:2402.11725*, 2024.

[20] Junsol Kim, James Evans, and Aaron Schein. Linear representations of political perspective emerge in large language models. *arXiv preprint arXiv:2503.02080*, 2025.

[21] Takuma Udagawa and Akiko Aizawa. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127, 2019.

[22] Takuma Udagawa and Akiko Aizawa. An annotated corpus of reference resolution for interpreting common grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9081–9089, 2020.

[23] Takuma Udagawa and Akiko Aizawa. Maintaining common ground in dynamic environments. *Transactions of the Association for Computational Linguistics*, 9:995–1011, 2021.

[24] Koh Mitsuda, Ryuichiro Higashinaka, Yuhei Oga, and Sen Yoshida. Dialogue collection for recording the process of building common ground in a collaborative task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5749–5758, 2022.

[25] Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. Conversational grounding: Annotation and analysis of grounding acts and grounding units. *arXiv preprint arXiv:2403.16609*, 2024.

[26] Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*, 2017.

[27] Philip Paquette, Yuchen Lu, Seton Steven Bocco, Max Smith, Satya O-G, Jonathan K Kummerfeld, Joelle Pineau, Satinder Singh, and Aaron C Courville. No-press diplomacy: Modeling multi-agent gameplay. *Advances in Neural Information Processing Systems*, 32, 2019.

[28] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.

[29] Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can llms negotiate? negotiationarena platform and analysis. *arXiv preprint arXiv:2402.05863*, 2024.

[30] Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R Lyu. How far are we on the decision-making of llms? evaluating llms' gaming ability in multi-agent environments. *arXiv preprint arXiv:2403.11807*, 2024.

[31] Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring bargaining abilities of llms: A benchmark and a buyer-enhancement method. *arXiv preprint arXiv:2402.15813*, 2024.

[32] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37:83548–83599, 2024.

[33] Andrew Konya, Lisa Schirch, Colin Irwin, and Aviv Ovadya. Democratic policy development using collective dialogues and ai. *arXiv preprint arXiv:2311.02242*, 2023.

[34] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

[35] Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, et al. Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.06864*, 2024.

[36] Anastassia Kornilova and Vlad Eidelman. Billsum: A corpus for automatic summarization of us legislation. *arXiv preprint arXiv:1910.00523*, 2019.

[37] Javier Arregui and Clement Perarnaud. A new dataset on legislative decision-making in the european union: the deu iii dataset. *Journal of European Public Policy*, 29(1):12–22, 2022.

[38] Heike KlÃijver, Svenja Krauss, and Hanna BÃd'ck. COALITIONAGREE Dataset, 2023. URL `https://doi.org/10.7910/DVN/XM5A08`.

[39] Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. Lawllm: Law large language model for the us legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4882–4889, 2024.

[40] Diego Garzia, Alexander Trechsel, and Lorenzo De Sio. Party placement in supranational elections: An introduction to the euandi 2014 dataset. *Party Politics*, 23(4):333–341, 2017.

[41] Jannis Vamvas and Rico Sennrich. X-stance: A multilingual multi-target dataset for stance detection. *arXiv preprint arXiv:2003.08385*, 2020.

[42] Farhad Moghimifar, Yuan-Fang Li, Robert Thomson, and Gholamreza Haffari. Modelling political coalition negotiations using llm-based agents. *arXiv preprint arXiv:2402.11712*, 2024.

[43] Dominik Stammbach, Philine Widmer, Eunjung Cho, Caglar Gulcehre, and Elliott Ash. Aligning large language models with diverse political viewpoints. *arXiv preprint arXiv:2406.14155*, 2024.

[44] Ilias Chalkidis and Stephanie Brandl. Llama meets eu: Investigating the european political spectrum through the lens of llms. *arXiv preprint arXiv:2403.13592*, 2024.

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

[46] Yueqing Liang, Liangwei Yang, Chen Wang, Congying Xia, Rui Meng, Xiongxiao Xu, Haoran Wang, Ali Payani, and Kai Shu. Benchmarking llms for political science: A united nations perspective. *arXiv preprint arXiv:2502.14122*, 2025.

[47] Simon HIX, Doru FRANTESCU, Sara HAGEMANN, and Abdul NOURY. Votewatch europe dataset. 2022.

[48] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86, 2005.

[49] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[50] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024. URL `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence`.

[51] Gail McElroy and Kenneth Benoit. Party groups and policy positions in the european parliament. *Party politics*, 13(1):5–28, 2007.

[52] Sven-Oliver Proksch and Jonathan B Slapin. Position taking in european parliament speeches. *British Journal of Political Science*, 40(3):587–611, 2010.

[53] Gail McElroy and Kenneth Benoit. Policy positioning in the european parliament. *European Union Politics*, 13(1):150–167, 2012.

[54] OpenAI. New embedding models and api updates, 2024. URL `https://openai.com/index/new-embedding-models-and-api-updates/`.

[55] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[56] United Nations. United nations charter, chapter v: The security council. URL https://www.un.org/en/about-us/un-charter/chapter-5.

[57] John Rawls. A theory of justice. In *Applied ethics*, pages 21–29. Routledge, 2017.

[58] John Stuart Mill. Utilitarianism. In *Seven masterpieces of philosophy*, pages 329–375. Routledge, 2016.

[59] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[60] Google DeepMind. Gemini flash, 2024. URL https://deepmind.google/technologies/gemini/flash/.

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

[62] Qwen Team. Qwen2.5, 2024. URL https://qwenlm.github.io/blog/qwen2.5-max/.

[63] AI@Meta. Llama 3.3, 2024. URL https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/.

[64] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[65] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.

[66] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[67] LFM Besselink, Katja Swider, Bastian Michel, et al. The impact of the uk's withdrawal on the institutional set-up and political dynamics within the eu. *Studie im Auftrag des Europäischen Parlaments (Policy Department for Citizens' Rights and Constitutional Affairs at the request of the AFCO Committee), Brussels*, 2019.

[68] Cas Mudde. The 2019 eu elections: Moving the center. *Journal of Democracy*, 30(4):20–34, 2019.

[69] Ariadna Ripoll Servent. The european parliament after the 2019 elections: Testing the boundaries of the'cordon sanitaire'. *Journal of Contemporary European Research*, 15(4):331–342, 2019.

[70] Tarik Abou-Chadi and Markus Wagner. Electoral fortunes of social democratic parties: do second dimension positions matter? In *Domestic contestation of the European Union*, pages 86–112. Routledge, 2021.

[71] Zhaowei Zhang, Fengshuo Bai, Qizhi Chen, Chengdong Ma, Mingzhi Wang, Haoran Sun, Zilong Zheng, and Yaodong Yang. Amulet: Realignment during test time for personalized preference adaptation of llms. *arXiv preprint arXiv:2502.19148*, 2025.

[72] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

[73] Esben Kran, Hieu Minh Nguyen, Akash Kundu, Sami Jawhar, Jinsuk Park, and Mateusz Maria Jurewicz. Darkbench: Benchmarking dark patterns in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[74] Simon Hix and Bjørn Høyland. Empowerment of the european parliament. *Annual review of political science*, 16(1):171–189, 2013.

[75] Helen Wallace, Mark A Pollack, Christilla Roederer-Rynning, and Alasdair R Young. *Policy-making in the European Union*. Oxford university press, 2020.

Supplementary Material

Table of Contents

A Dataset Construction Details **16**
- A.1 Data Collection Process . . . 16
- A.2 Vote In Favor Calculation . . . 17
- A.3 Used Political Group Name Abbreviations . . . 18
- A.4 Output Data Entry Schema . . . 19
- A.5 Data Filtering Analysis . . . 19
- A.6 Data Post-Processing Details . . . 20

B Task Details **22**
- B.1 Topic Contents . . . 22
- B.2 More Stances Sematic Representation Results . . . 22

C Prompt Details **24**
- C.1 Data Processing Prompts . . . 24
- C.2 Task Prompts . . . 27
- C.3 Evaluation Prompts . . . 28

D More Experimental Results **29**
- D.1 Detailed Simulated Evaluation Consistency Results . . . 30
- D.2 Detailed Fine-grained Topics Results . . . 30

E Case Study **35**
- E.1 Case Study: LLM Performances . . . 35
- E.2 Case Study: Different Parliament Settings . . . 37

F Background of the European Parliament **41**

G Discussion and Limitations **41**

H Ethical Statement and Disclaimer **42**
- H.1 Copyright of Data Sources . . . 42
- H.2 Potential Societal Impact and Statement on the Use of *EuroCon* . . . 42

A Dataset Construction Details

In this section, we will provide a detailed explanation of the complete process of data collection and post-processing mentioned in subsection 3.1.

A.1 Data Collection Process

In this subsection, we will focus on the perspective of large-scale data crawling, introducing the methodology and process of raw data collection.

A.1.1 Data Sources

The data collection process for the *EuroCon* begins with the VoteWatch Europe dataset [47], which contains structured voting records of the European Parliament (EP) spanning 18 years from 2004 to 2022. Since the data for the five years from 2004 to 2009 is incomplete, we have excluded it. The portion of the dataset we use includes: (1) Excel files with metadata for the seventh, eighth, and half of the ninth European Parliament terms (2014-2022), including vote identifiers, titles, issue topics, etc.; (2) Roll call voting records mapping MEPs to vote outcomes, including six categories: in favor, against, abstain, absent, not voted, not an MEP; (3) URLs of the original sources from the official website of the European Parliament regarding where to obtain the voting data.

The second data source to be introduced is HowTheyVote[4]. This data source also presents roll call voting data for each MEP and provides URLs that link to the data sources. There are two main differences between this data source and the VoteWatch Europe dataset: first, it only includes data from the 9th and 10th European Parliament sessions after 2019. Second, it contains URLs for both the voting data and related records of resolutions and debates from the European Parliament's official website.

The last and most important data source is the European Parliament's official website[5]. This source lacks systematic organization of roll call voting data for each resolution (it's not absent, but it's not easy to scrape on a large scale, which makes us rely on other data sources for voting record extraction). However, it provides extensive and detailed data on resolutions and debate records for each decision.

Through HowTheyVote, we discovered how voting URLs can correspond to their respective resolution and debate records via specific web navigation. Once this information is obtained, we can cleverly combine the voting information from the VoteWatch Europe dataset and HowTheyVote, along with the voting source URLs from the European Parliament's official website, to access the resolution and debate record data corresponding to each decision. This establishes the foundation for large-scale data scraping.

A.1.2 Unified URL Parsing

On the official website of the European Parliament, some URLs have multiple redirect issues, which means the directly indexed webpage is not the original record's page. To solve this problem, we developed an automated pipeline to handle specific short URL issues in the European Parliament system, which consists of the following key steps: First, we sent HTTP HEAD requests for all short URLs (in formats such as `europarl.europa.eu/doceo/xxx`) to fully trace redirection chains. Second, the final URLs were validated against an official domain whitelist to ensure that all resolved results point to valid European Parliament resources. Finally, cryptographic hashing was employed for integrity verification, storing both original and resolved URLs while generating SHA-256 digests for audit trails.

This solution effectively addresses URL standardization issues in the European Parliament's official document system while preserving complete data provenance information. By combining the verification of the network protocol layer with cryptographic validation, a dual guarantee mechanism was established. In this way, we can ensure that every URL can index the corresponding webpage information.

[4] https://howtheyvote.eu

[5] https://www.europarl.europa.eu

A.1.3 Web Content Extraction

We employed the Python `BeautifulSoup` library[6] to parse the raw HTML content from the official European Parliament website. However, the European Parliament's web pages do not follow a uniform HTML format, especially those targeting paragraphs with distinct stylistic features (such as those with `margin-left:17.85pt` formatting). This necessitates handling these diverse special webpage structures during the data scraping process to accurately capture the resolution body text. To address this situation, we performed customized processing for each special case of uniquely occurring resolution webpage format, including identification methods like paragraph filtering using standard resolution startings (e.g., "The European Parliament"), ultimately obtaining complete raw resolution data.

For debate records, through document object model (DOM) tree traversal techniques, we identified HTML elements containing debate records (nodes with the `doceo-ring-steps-step-label` class). During speech content extraction, the system automatically filters procedural statements (e.g., chairperson remarks like "The President") while retaining substantive policy debate content. This process combines dual verification mechanisms of semantic analysis and rule-based pattern matching.

For the special requirements of the 9th European Parliament (2019 - 2022), we developed a parsing adapter based on URL path heuristic rules. By recognizing specific path patterns (such as URLs containing `/A8/` or `/B9/` identifiers), the system can automatically switch the corresponding content extraction strategies to effectively address technical challenges caused by structural changes in the websites of parliament. The framework supports dynamic loading of new parsing rules, ensuring long-term system maintainability. Key features of this implementation include: (1) Context-aware parsing for different parliamentary terms; (2) Automated detection of document structural changes; and (3) Fallback mechanisms for handling legacy formats.

These approaches leverage the standardized typography of European parliamentary document systems to reliably extract structured textual content. In this way, we obtained the 30,698 original parliamentary deliberation records mentioned in subsection 3.1.

A.1.4 Redundant Data Filtering

In the European Parliament, each issue requires careful consideration before reaching a final resolution, so clearly, no resolution can be finalized in just one meeting. As a result, the parliamentary records show that each issue typically undergoes more than ten rounds of revisions and voting. Therefore, we need to efficiently eliminate the intermediate processes of these issues, leaving only the final effective data version. To address this problem, we implemented a rigorous two-phase deduplication mechanism to ensure the uniqueness and authority of legislative data. The first phase handles duplication at the legislative level, while the second phase resolves document-level ambiguities.

Legislative Level Uniqueness Guarantee. From the perspective of the procedural legitimacy of the European Parliament, the final decision should be based on the roll-call vote results of the final vote[7]. This information is represented in the VoteWatch Europe dataset with the label `final_vote=1`. Therefore, in this paper, we only retain the voting records with this label for each issue.

Document-Level Disambiguation. When multiple entries referencing identical legislative content (identified by URL matching) were detected, we adopted the most recent-first principle, retaining the latest record according to the `vote_timestamp` field. This mechanism establishes a bijective relationship between legislative acts and their canonical representations while maintaining the temporal logic of data updates.

A.2 Vote In Favor Calculation

In the VoteWatch Europe dataset, all roll-call voting data records the voting decisions of each MEP and uses the following six labels to record their voting outcomes: in favor, against, abstain, absent, not voted, and not an MEP. For the *EuroCon* setup, we need the proportion of MEPs voting in favor of each party on each issue. Therefore, we need to further process the voting data. First, we need to

[6] https://pypi.org/project/beautifulsoup4

[7] https://www.europarl.europa.eu/doceo/document/RULES-10-2025-01-20-RULE-047_EN.html

match each MEP to their respective party, which is labeled in the VoteWatch Europe dataset. However, the names of the same parties are not consistent (due to different names and typos), so we reclassified these to accurately identify the party each MEP belongs to. We then calculated the proportion of MEPs voting in favor of each party on each issue. It's important to note that, as mentioned above, there are six voting outcome labels, but we only use the "in favor" label to calculate the proportion of votes in favor. As for the HowTheyVote data source, the proportion of votes in favor of each party is already calculated, so for the ninth parliament, we don't need to perform this operation.

A.3 Used Political Group Name Abbreviations

For each political party in the European Parliament, there are different names and abbreviations. For example, the European People's Party has official abbreviations like EPP and PPE, among other variations. Due to the different languages used in European Union countries, there are corresponding abbreviations for different languages as well. Therefore, in this document, we need to introduce the party name abbreviations used in *EuroCon* and their corresponding party names.

Table 3: Used political group name abbreviations in the 7th parliament.

Abbreviation	**Full Name**
EPP	European People's Party
EFD	Europe of Freedom and Democracy
SD	Progressive Alliance of Socialists and Democrats
ALDE	Alliance of Liberals and Democrats for Europe Party
ECR	European Conservatives and Reformists Group
GREEN/EFA (GREEN_EFA in dataset)	The Greens/European Free Alliance
GUE/NGL (GUE_NGL in dataset)	The Left in the European Parliament

Table 4: Used political group name abbreviations in the 8th parliament.

Abbreviation	**Full Name**
EPP	European People's Party
SD	Progressive Alliance of Socialists and Democrats
ECR	European Conservatives and Reformists Group
EFDD	Europe of Freedom and Direct Democracy
GREEN/EFA (GREEN_EFA in dataset)	The Greens/European Free Alliance
GUE/NGL (GUE_NGL in dataset)	The Left in the European Parliament
ALDE	Alliance of Liberals and Democrats for Europe Party
ENF	Europe of Nations and Freedom

Table 5: Used political group name abbreviations in the 9th parliament.

Abbreviation	**Full Name**
EPP	European People's Party
SD	Progressive Alliance of Socialists and Democrats
ECR	European Conservatives and Reformists Group
RENEW	Renew Europe
GREEN/EFA (GREEN_EFA in dataset)	The Greens/European Free Alliance
GUE/NGL (GUE_NGL in dataset)	The Left in the European Parliament
ID	Identity and Democracy

In the 7th parliament term, the party abbreviations we used were EPP, EFD, SD, ALDE, ECR, GREEN/EFA, and GUE/NGL, as shown in Table 3. Interestingly, the abbreviation GUE/NGL for

The Left in the European Parliament does not directly correspond to its full English name. This is because the party was originally formed by the merger of the Confederal Group of the European United Left (GUE) and the Nordic Green Left Alliance (NGL). Information on party abbreviations for the 8th and 9th parliaments is shown in Table 4 and Table 5.

A.4 Output Data Entry Schema

Finally, after the large-scale crawling and preprocessing steps described above, we obtained 2,225 high-quality complete parliamentary record data entries. For each entry, we used the following JSON format for storage:

```
{
  "excel_title": "Issue Title",
  "web_title": "HTML-Derived Title",
  "topic_select": "Fine-grained Topic Name",
  "text_url": "Canonical Document URL",
  "resolution": "Full Resolution Text",
  "votes_total": {"FOR": 75, "AGAINST": 124, ...},
  "votes": [
  {
    "group": {"code": "EPP", "label": "...", ...},
    "stats": {"FOR": 35, "AGAINST": 72, ...}
  }, ...
  ],
  "debate": {
    "title": "Debate Transcript Title",
    "views": [{"speaker": "MEP Name", "debate": "Utterance"}, ...]
  }
}
```

The JSON file contains all the quintuple raw information mentioned in subsection 3.1, namely issue, topic, debates, resolution, and votes. We will introduce which keys in the JSON field correspond to these raw pieces of information as follows: (1) issue: excel_title and web_title provide the official and HTML-derived issue titles, respectively. We use "excel_title: web_title" as the issue's final name; (2) topic: top_select indicates the policy area, and text_url links to the canonical document; (3) debates: The debate field describes the original debate record, where the title is the debate webpage's title, and views include the current speaker's name (speaker) and their speech content (debate); (4) resolution: Indicated by the resolution field; (5) votes: We have separately saved the results of two types of votes: the votes_total, which represents the overall votes for the resolution in the parliament, and the votes, which represents the votes of each party on the resolution. In the votes field, group indicates the information of the party currently voting, and stats represents the record of their votes.

A.5 Data Filtering Analysis

Our pipeline implemented rigorous quality controls across three parliamentary terms, with key metrics shown in Table 6.

Table 6: Data filtering ratio by different parliamentary terms.

Metric	7th	8th	9th
Initial Records	6,963	10,276	13,459
Duplicates Removed	5,333 (76.6%)	8,349 (81.2%)	12,414 (92.2%)
Debate Transcripts Missing	580/1,630 (35.6%)	800/1,927 (41.5%)	487/1,045 (46.6%)
Final Valid Records	1,050 (15.1%)	1,127 (11.0%)	558 (4.1%)

Initial Records. The initial records represent the total number of unprocessed voting records collected from raw data sources. For instance, the 7th term had 6,963 records, while the 9th term saw

a significant increase to 13,459 records, and that is only half of the term. This metric is significant as it reflects the original scale of data collection, illustrating a 93% growth from the 7th to the 9th term.

Duplicates Removed. Duplicates are identified through the process described in subsubsection A.1.4 and subsequently removed from the dataset. The key characteristics of this process include both absolute numbers (e.g., 8,349 removed in the 8th term) and percentages (81.2%). The duplication rate increases across terms, from 76.6% in the 7th term to 92.2% in the 9th term. Notably, the high duplication rate in the 9th term (92.2%) perhaps reflects the increased frequency of its discussion issues.

Debate Transcripts Missing. Some voting records lack corresponding parliamentary debate texts, resulting in missing debate transcripts. This issue is represented in two forms: as a numerator/denominator (e.g., the 7th term: 580/1,630) and as a percentage (ranging from 35.6% to 46.6%). There is a consistent upward trend in the missing rate, with the 9th term reaching 46.6%, indicating that nearly half of the records are devoid of contextual debate information.

Final Valid Records. The final valid records are those that are available and pass all quality checks. They are calculated by subtracting duplicates and missing records from the initial records. For example, in the 7th term, the calculation is 6,963 (initial records) - 5,333 (duplicates) - 580 (missing records) = 1,050 valid records. Despite the initial growth of the records, the number of valid records in the 9th term (558) decreased by 11% compared to the 7th term (1,050), highlighting the decline in the usability of the data.

The above analysis reveals that the data we used in *EuroCon* only accounts for 7.2% of the original data, reflecting that the data we adopted consists of carefully selected high-quality deliberation records.

A.6 Data Post-Processing Details

Due to the redundancy of the raw data, such as the large number of useless remarks in the debate, after collecting the original data, we further used DeepSeek-R1 [49] and rule-based methods for data cleaning and post-processing operations. First, we used DeepSeek-R1 to organize and summarize the resolutions, removing redundant parts while retaining the original resolution format. We further summarized the background of the current parliamentary discussion topics based on issue, resolution, and debate information using DeepSeek-R1.

Next, we processed the voting data, where the original voting information included each member's vote on each issue. We matched each member with their parliamentary party and calculated the voting information for each party on the current resolution. We calculated the proportion of members within the party who voted in favor and rounded down to an integer between 0 and 9 as the party's preference score for the resolution.

Table 7: Paraphrase word list for the data post-processing procedure.

Attitude	Word List
Support Verbs	support, agree, endorse, advocate, approve, sanction, uphold, accept, promote
Oppose Verbs	oppose, reject, disapprove, condemn, conflict, doubt, challenge, dispute, against
Support Adverbs	fully, totally, completely, absolutely, entirely, fundamentally, firmly
Oppose Adverbs	partly, slightly, partially, confitionally

Subsequently, based on the resolution and each party's voting information, we let DeepSeek-R1 summarize each party's stances on the issue from the debate data. If a party did not express a stance or opinion in the debate, we removed the party from the issue. The detailed prompt can be found in Appendix C.1. Then we used rule-based methods to perform synonym replacement on tone words expressing political party stances. For example, "strongly agree" can be replaced with "fully endorse" or "totally support", among others (detailed in Table 7). This approach increases data diversity and helps reduce the bias in word choices introduced by the LLM. Additionally, since all stances in the debate data are related to the current committee proposal or submitted resolution, and we need the

LLM to provide new resolutions when using this data, we replaced the word "resolution" in each party stance with the synonym "issue" to adjust the stances on the resolution to stances on the issue. This eliminates conflicts in referential terms between the new resolution generated by the tested LLMs and the word "resolution" in the stances during practical data usage.

We applied the process to each data entry. Through this approach, we cleaned the raw data into sextuples of (issue, topic, background, stances, resolution, votes), where stances and votes contain relevant information from all parties involved in the discussion of the issue.

Table 8: Overview of the fine-grained topics and their contents (with some topic names abbreviated for convenience in the table).

Topic Name	**Detailed Content**
Agriculture	Agricultural policy, rural development, and food security.
Budget	Budget negotiations, annual budget adoption, and financial reforms.
Budgetary Control	Budget implementation, ensures financial transparency, combats fraud, and promotes accountability.
Civil Liberties	Policies on civil liberties, justice, and home affairs, focusing on fundamental rights, migration, data protection, and security.
Constitutional Affairs	Constitutional affairs, focusing on treaty implementation, institutional reforms, and democratic governance .
Culture & Education	Policies on culture, education, media, youth, and sports, managing flagship programs to promote cultural diversity, education, and cross-border cooperation.
Development	Global sustainable development, overseeing EU aid budgets, combating poverty, and strengthening partnerships to tackle inequality and humanitarian challenges.
Economic Affairs	Regulation of financial services, the free movement of capital, payments, taxation, competition policies, and the international financial system.
Employment	Employment policies, workers' rights, social inclusion, and addressing challenges like economic transitions and inequality through legislative oversight.
Public Health	Environmental policies, climate action, and food safety, prioritizing Green Deal implementation, biodiversity, and sustainable transition, public health issues, including pharmaceutical reforms, disease prevention (e.g., cancer, mental health), health data governance, and reducing EU health inequalities.
Fisheries	Sustainable fisheries management, marine conservation, and socio-economic support for coastal communities under the Common Fisheries Policy reform.
Foreign & Security	Common Foreign and Security Policy and international agreements, defense strategies, hybrid threats, and military resilience in response to security challenges like Russia's war in Ukraine.
Gender Equality	Gender equality, combats violence/discrimination, and ensures women's inclusion in decision-making to address democratic deficits and societal fairness.
Industry & Energy	Legislation for energy transition, industry competitiveness, research innovation, digital/telecom policies, cybersecurity, and space policy to drive sustainable prosperity and EU strategic autonomy.
Internal Market	Single market rules, including digital integration and consumer protection, aiming to align with Green Deal objectives and high social/environmental standards.
International Trade	International trade agreements, WTO compliance, and scrutiny of trade policy implementation to strengthen the EU's global economic role.
Legal Affairs	Legal affairs, corporate law, intellectual property, and EU law simplification while ensuring institutional compliance and judicial oversight.
Regional Development	Cohesion policy, regional development, and solidarity through structural funds and multilevel governance to address disparities and future enlargement challenges.
Transport & Tourism	Transport/tourism decarbonization, digital transformation (e.g., autonomous vehicles), and sustainable mobility to meet climate goals and social equity.

B Task Details

In this section, we will present the definitions of the coarse-grained and fine-grained topics we have categorized for each issue mentioned in subsection 3.1, as well as a more detailed display of the distribution of various political parties' stances in the semantic space.

B.1 Topic Contents

We categorize all collected data based on the topics outlined in the VoteWatch Europe dataset, which are derived from the committees of the European Union[8]. These 19 topics are then grouped into 5 coarse-grained categories:

Economics. Focuses on macroeconomic strategies. The fine-grained topics in this category are International Trade[9], Internal Market & Consumer Protection[10], Employment & Social Affairs[11], and Economic & Monetary Affairs[12].

Industry. Covers policies for specific industries. The fine-grained topics in this category are Agriculture[13], Fisheries[14], Transport & Tourism[15], and Industry, Research & Energy[16].

Budget. Encompasses budget policies for development. The fine-grained topics in this category are Development[17], Regional Development[18], Budget[19], and Budgetary Control[20].

Security. Addresses basic security guarantees, including military and health aspects. The fine-grained topics in this category are Environment & Public Health[21][22], and Foreign & Security Policy[23][24].

Civil Rights. Pertains to political and cultural issues. The fine-grained topics in this category are Culture & Education[25], Gender Equality[26], Civil Liberties, Justice & Home Affairs[27], Constitutional and Inter-institutional Affairs[28], and Legal Affairs[29].

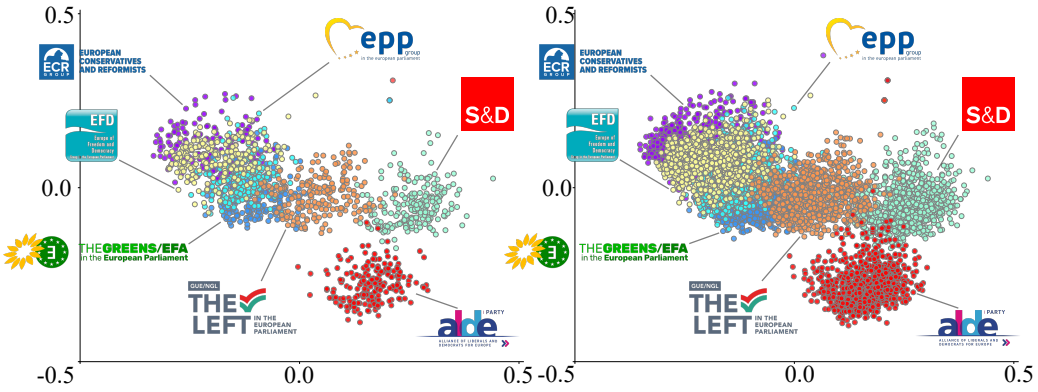
We provide an overview of the main content covered under each topic in Table 8.

B.2 More Stances Sematic Representation Results

In subsection 3.2, we have previously provided a rough overview of the diversity of stances between parties in each parliamentary session. For illustration simplicity, we only displayed the distribution of

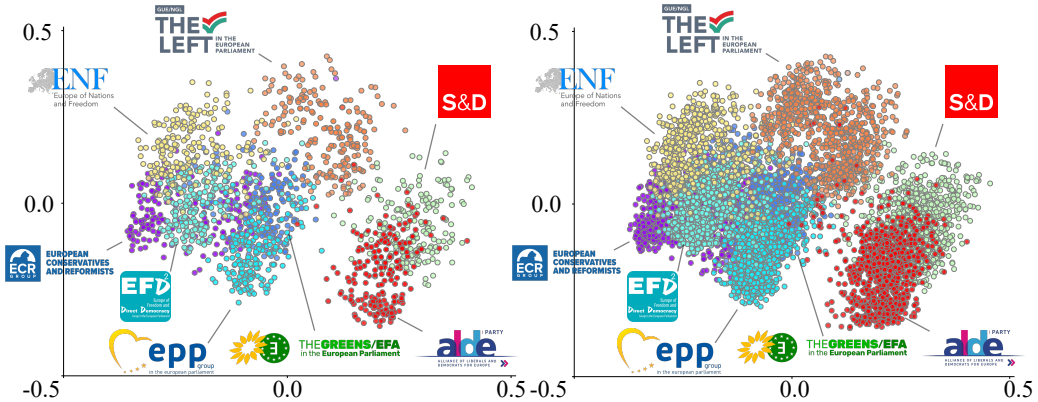
[8] https://www.europarl.europa.eu/committees/en/about/list-of-committees
[9] https://www.europarl.europa.eu/committees/en/inta/about
[10] https://www.europarl.europa.eu/committees/en/imco/about
[11] https://www.europarl.europa.eu/committees/en/empl/about
[12] https://www.europarl.europa.eu/committees/en/econ/about
[13] https://www.europarl.europa.eu/committees/en/agri/about
[14] https://www.europarl.europa.eu/committees/en/pech/about
[15] https://www.europarl.europa.eu/committees/en/tran/about
[16] https://www.europarl.europa.eu/committees/en/itre/about
[17] https://www.europarl.europa.eu/committees/en/deve/about
[18] https://www.europarl.europa.eu/committees/en/regi/about
[19] https://www.europarl.europa.eu/committees/en/budg/about
[20] https://www.europarl.europa.eu/committees/en/cont/about
[21] https://www.europarl.europa.eu/committees/en/envi/about
[22] https://www.europarl.europa.eu/committees/en/sant/about
[23] https://www.europarl.europa.eu/committees/en/afet/about
[24] https://www.europarl.europa.eu/committees/en/sede/about
[25] https://www.europarl.europa.eu/committees/en/cult/about
[26] https://www.europarl.europa.eu/committees/en/femm/about
[27] https://www.europarl.europa.eu/committees/en/libe/about
[28] https://www.europarl.europa.eu/committees/en/afco/about
[29] https://www.europarl.europa.eu/committees/en/juri/about

200 sampled data points in the semantic space for each party in the seventh and eighth parliaments. In this section, we will present a more detailed analysis of the sample data distribution and the complete data distribution for each party in every parliamentary session of *EuroCon*. This will further reveal the significant semantic diversity and stance conflicts between parties in *EuroCon*.



(a) Sampled Stances for the 7th Parliament (b) All Stances for the 7th Parliament

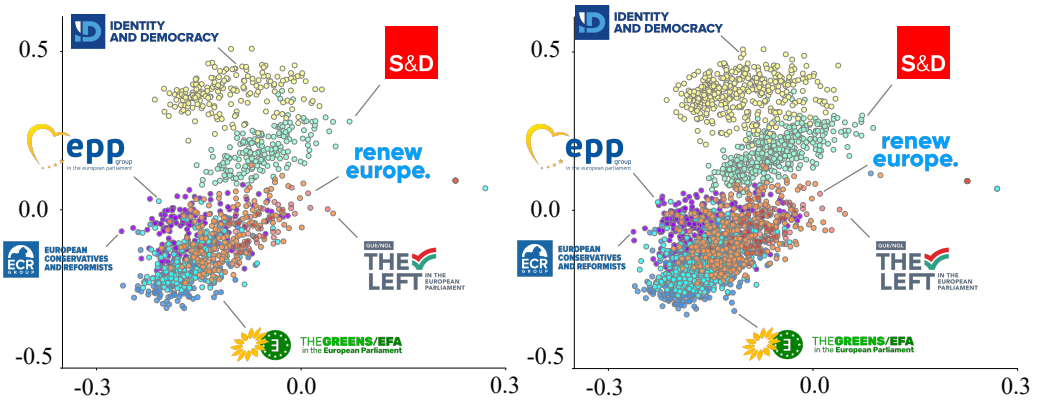
Figure 7: Semantic representation distribution of party stances (indicated by their symbols) in the 7th (2009-2014) term of the European Parliament in *EuroCon*. Figure (a) shows the sampled stances while Figure (b) illustrates all the stances.



(a) Sampled Stances for the 8th Parliament (b) All Stances for the 8th Parliament

Figure 8: Semantic representation distribution of party stances (indicated by their symbols) in the 8th (2014-2019) term of the European Parliament in *EuroCon*. Figure (a) shows the sampled stances while Figure (b) illustrates all the stances.

As shown in Figure 7, Figure 8, and Figure 9, we present the sampled stances and all stances of all political parties during the 7th, 8th, and 9th terms of the parliament. From these three figures, it can be observed that the distribution results after sampling 200 data points for each party closely resemble those of the entire dataset, providing a strong reference value. Additionally, we can see that the distribution of party stances in the 7th and 8th terms of the European Parliament is more diverse compared to the 9th term. This may be due to factors such as Brexit [67] and the rise of right-wing forces [68–70], which highlights that our data analysis aligns with actual political trends.



(a) Sampled Stances for the 9th Parliament

(b) All Stances for the 9th Parliament

Figure 9: Semantic representation distribution of party stances (indicated by their symbols) in the half of the 9th (2019-2022) term of the European Parliament in *EuroCon*. Figure (a) shows the sampled stances while Figure (b) illustrates all the stances.

C Prompt Details

In this section, we will demonstrate the details of all the prompts involved in this paper.

C.1 Data Processing Prompts

First, we will introduce the prompts used in the data post-processing process. In this process, the prompts required include those for obtaining resolution, background, and extracting stances from the debate data of each party's MEPs. We will explain each of these in detail below.

> Summarize the key points of this European Parliament resolution in one continuous paragraph, without any formatting or line breaks. Begin the summary with 'The European Parliament raised' and focus on the resolution's substantive content, decisions, and numerical data where applicable. Omit procedural details like voting records and amendments, focusing only on the original resolution text. Ensure the output is concise yet comprehensive. Here's the resolution: {resolution}

As shown above is our resolution summarization prompt template. Its primary purpose is to condense lengthy resolution texts into a usable length while preserving their original format. As a resolution of the European Parliament, its most distinctive linguistic feature is starting with "The European Parliament", and here we require that it is immediately followed by the verb "raised". We also require it to focus on the resolution's substantive content, decisions, and numerical data where applicable. Omit procedural details like voting records and amendments, focusing only on the original resolution text. Additionally, we require it to ensure the output is concise yet comprehensive.

> **Title:**
> {title}
>
> **Resolution:**
> {resolution}
>
> **Debate:**
> {debate}
>
> **Instructions:**
> Based on the provided Title, Resolution, and Debate, compose a neutral background summary

(under 50 words) objectively describing the contextual factors that led to this issue being raised in the European Parliament. The summary must:
1. Focus solely on documented events and conditions prior to parliamentary consideration
2. State the general topic area for parliamentary discussion 3. Avoid all reference to debate content or resolution outcomes

Output Requirements:
- Strict 50-word maximum, in one paragraph, without title or line changes.
- First part: Factual description of pre-existing conditions (events/institutional/geopolitical context)
- Second part: Clear statement of the general discussion topic ("The Parliament will discuss...")
- Use only verified facts - no speculative language ("may reflect"/"could indicate")
- Maintain complete neutrality, exclude any reference to:
Parliamentary proceedings
Debate positions
Resolution content
Political motivations

As shown above is our background prompt template, which summarizes relevant background knowledge related to the issue based on the issue title, resolution, and full debate record, clarifying the problems the European Parliament needs to address. We require the generated background to meet the following criteria: Based on the provided Title, Resolution, and Debate, compose a neutral background summary (under 50 words) objectively describing the contextual factors that led to this issue being raised in the European Parliament. The summary must focus solely on documented events and conditions prior to parliamentary consideration, state the general topic area for parliamentary discussion, and avoid all reference to debate content or resolution outcomes. The output must adhere to a strict 50-word maximum, consist of one paragraph without title or line changes, begin with a factual description of pre-existing conditions (events/institutional/geopolitical context), and conclude with a clear statement of the general discussion topic ("The Parliament will discuss..."). Use only verified facts, no speculative language ("may reflect"/"could indicate"), while maintaining complete neutrality and excluding any reference to parliamentary proceedings, debate positions, resolution content, or political motivations.

Topic:
{topic}
Resolution:
{resolution}
Debate:
{debate}
Score (0 - 9):
{score}
Instructions:
{instructions}

Our opinion summarization prompt template is quite simple, and just needs to summarize each party's stances conditioned on the issue title, resolution summary, all the debate records, and the party's voting score. The key point is the instructions, which have been outlined below:

If the debate is empty or the {party} party has no arguments, output: "None"
Otherwise:
1. **Score-Specific Requirements**:

9-10	Perfect alignment	None	Forbidden	“fully endorses”, “perfectly aligns”
7-8	Strong support	$\leq$1 minor suggestion	Forbidden	“strongly supports”, “approves”
5-6	General support	$\leq$2 constructive mods	$\leq$1 phrased as concern	“supports with suggestions”, “advises”
3-4	Reserved approval	$\leq$3 major changes	$\leq$2 objections	“conditionally accepts”, “requests revisions”
0-2	Explicit opposition	N/A	Primary focus	“rejects”, “opposes fundamentally”

2. **Argument Processing Rules**:
- For scores $\geq$7:
- Convert all criticism to “enhancement opportunities” (e.g., “opposes X” $\rightarrow$ “proposes strengthening X”)
- Minimum 3:1 support-to-modification ratio
- For scores $\leq$3:
- Highlight contradictions with party principles
- Use comparative language: “fails to address”, “inconsistent with”

3. **Language Enforcement**:
- **High Scores (7-10)**:
- Mandatory reinforcement phrases:
“This aligns perfectly with party’s longstanding commitment to...”
“The resolution effectively advances party’s priority of...”
- **Low Scores (0-3)**:
- Required framing:
“This fundamentally conflicts with party’s position that...”
“The proposal overlooks critical aspects such as...”

4. **Output Validation Checklist**:
- All viewpoints begin with “{party} [score-appropriate verb]...”
- Modification proposals include concrete wording (e.g., “amend Article 3 to specify...”)
- For scores $\geq$5, $\geq$80% of content must directly affirm resolution goals
- Opposition arguments (if allowed) must reference specific resolution clauses

5. **Examples (Score=8)**:
- {party} strongly supports the transparency measures in Articles 1-3, particularly the lobbyist disclosure requirements.
- {party} proposes extending document publication deadlines by 15 days to ensure thorough review without opposing the principle.
- {party} applauds the anti-fraud provisions as matching their 2023 manifesto commitments.

6. **Special Cases**:
- Empty debate with score $\geq$7 $\rightarrow$ “Consistent silent endorsement”
- Contradictory arguments $\rightarrow$ Flag with: “[Note: Reconcile with score {score}]”
7. Output format (one viewpoint per line):
viewpoint_1
viewpoint_2
...
(Max 5 viewpoints, no numbering or bullets)

The instructions outlined above illustrate how to determine different parties’ stances based on parliamentary debate records. If the debate lacks content or the party has no arguments, the output is “None”. Otherwise, the system categorizes responses into specific score ranges with criteria for alignment, support, and opposition. High scores (9-10) indicate perfect alignment, while low scores (0-2) reflect explicit opposition. The prompt includes rules for processing arguments, emphasizing positive reinforcement for higher scores and highlighting contradictions for lower scores. An output validation checklist ensures all viewpoints are appropriately framed and modifications are clearly articulated. Examples illustrate these rules, and special cases address unique situations. The output format presents up to five viewpoints per line for clarity and coherence.

C.2 Task Prompts

In this section, we will present the prompts we used in the process of political consensus finding with LLMs using *EuroCon*.

> You are an AI designed to provide the most helpful, clear, and concise responses. Focus on giving actionable information, ensuring accuracy and detail without overwhelming the user. You should also be patient, polite, and calm. Avoid unnecessary complexity and always prioritize practical, user-friendly advice.

The system prompt displayed above demonstrates good performance and has been widely used in previous work [71].

In the following, we will present our task prompt. We require the creation of a consensus European Parliament resolution statement that adheres to the specified criteria in a single, unbroken paragraph. The statement must begin with "The European Parliament raised" and concentrate on the substantive content, decisions, and numerical data where applicable. It should address opposing stances by providing detailed solutions and mitigations for the concerns raised, while moderating supporting stances with appropriate qualifications and limitations. Procedural details such as voting records and amendments should be omitted, focusing solely on the original resolution text. The output must be concise yet comprehensive.

> Background: {background}
> A group of {party_num} political parties in the European Parliament was required to find consensus on this topic: {topic}
> Below is each party's stance:
> {stances}
> {task_requirements}
> Your task is to write a consensus European Parliament resolution statement that meets the upper requirements in one continuous paragraph, without any formatting or line breaks. Begin the resolution statement with 'The European Parliament raised' and focus on the resolution's substantive content, decisions, and numerical data where applicable. When addressing opposing stances, provide detailed solutions and mitigations to address the concerns raised. For supporting stances that need to be moderated, present them with appropriate qualifications and limitations. Omit procedural details like voting records and amendments, focusing only on the original resolution text. Ensure the output is concise yet comprehensive. Here's an example of the resolution:
> {resolution}
> Now is your turn:

As for our opinion prompt template, we just simply use the following format to illustrate each party's positions:

> Party {party_name}: {stance}

C.2.1 Task Requirement Templates

In this section, we will present our prompts for conveying the specific task requirements to the tested LLMs.

First, we introduce our task description for different political goals. The first is the prompt for passing the resolution under specific voting requirements:

> The resolution should be based on the seat proportions of each party, where the proportions are defined by the following weights rather than actual parliamentary seats: {sea_apportionmen_weights}

Then are the requirements for Rawlsianism and Utilitarianism:

> The resolution should be based on the Rawlsian principle, which means maximizing the benefits for the political parties with the weakest positions to achieve the greatest possible fairness.

> The resolution should be based on the Utilitarianism principle, which means maximizing the preferences of all political parties while ensuring their sum is maximized.

Finally, we convey each party's seat apportionment information through the following simple template:

> Party {party_name}'s seat proportion is {seat_proportion}%.

C.2.2 Voting Mechanism Templates

In this section, we introduce the prompts used in the paper to describe three different voting mechanisms. Below are the details for simple majority, two-thirds majority, and veto power, respectively:

> The resolution should be supported by more than 50% of the parliament members. The resolution should reflect each party's stance in proportion to their seat allocation.

> The resolution should be supported by more than two-thirds of the parliament members. The resolution should reflect each party's stance in proportion to their seat allocation.

> Because {veto_party_name} has veto power, the resolution should meet the stances of {veto_party_name} as much as possible while also being supported by more than 50% of the parliament members. The resolution should reflect each party's stance in proportion to their seat allocation.

C.3 Evaluation Prompts

In this section, we will present our evaluation framework's prompts.

> Your task is to give a score for the following statement based on the score guidelines and your knowledge of the party.
> {{"{wait_for_judge_statement}", {{ "choices": ["not at all", "very poorly", "poorly", "somewhat poorly", "somewhat", "moderately", "somewhat well", "well", "very well", "perfectly"],
> "choice_numbers": [0, 1, 2, 3, 4, 5, 6, 7, 8, 9],
> YOU SHOULD ONLY RESPONSE IN [0, 1, 2, 3, 4, 5, 6, 7, 8, 9],
> "choice_number":

The template presented above outlines our comprehensive evaluation framework, focusing primarily on the content awaiting assessment and the instructions for evaluation. The subsequent prompt provides a structured approach for scoring statements based on specific guidelines and party knowledge. It includes a scoring system with choices ranging from "not at all" to "perfectly", corresponding to numerical values from 0 to 9. The task requires a precise response within this numerical range, ensuring alignment with the party's principles and facilitating consistent evaluation outcomes. The most crucial parts of the template are the content awaiting evaluation and the evaluation guidelines, which we will present separately below.

The first is the template we provide to our evaluation framework for assessing content. This framework is designed to analyze the degree to which a given resolution encapsulates the specific implementation

details that reflect the stances of a particular party, regardless of whether these stances support or oppose the issue. By focusing on the alignment between the resolution and the party's expressed opinions, the framework aims to ensure a comprehensive evaluation of how well the resolution addresses the party's principles and priorities. The guidelines included in the prompt serve to direct the assessment process, ensuring consistency and accuracy in evaluating the alignment between the resolution and the party's stance.

Background:
{background}
Consider the following statement:
{resolution}
The {party_name}'s opinion is:
{stance}
To what extent does this resolution provide specific implementation details that capture {party_name}'s stances?
Regardless of whether the stances itself is supportive or opposing to the issue.
{guidelines}

The next one is our evaluation guidelines, which aim to assess resolutions based on their alignment with the European Parliament's stances. The scoring system ranges from 0 to 9, evaluating resolutions on their specificity, feasibility, and comprehensiveness in addressing key points from various stances. Scores from 0-3 indicate resolutions that lack proper format, omit critical details, or undermine intended benefits. Scores from 4-6 reflect partial alignment with party interests, addressing some but not all key aspects. Scores from 7-9 recognize fully detailed and practical implementation measures that comprehensively address all stance points, ensuring no compromise or dilution of objectives.

Please follow this scoring guideline:
- **Score 0-3**: If the resolution does not start with "The European Parliament", or if the resolution only rephrases content from the stances without providing specific implementation details, contains impractical/unfeasible implementation proposals, omits key points mentioned in the stances, or if it contains elements that weaken/dilute the benefits sought in supportive stances (for opposing stances, if it promotes/strengthens what the party opposes). IF THE CONTENT IS EVEN NOT IN A RESOLUTION FORMAT, YOU SHOULD GIVE 0 DIRECTLY.
- **Score 4-6**: If the resolution provides some feasible implementation details for the stances' requirements but lacks comprehensiveness (e.g. only addresses some aspects, missing some points from the stances) or contains minor conflicts with party interests (e.g. implementation approach differs slightly from party's preferred method, timeline not fully aligned with party's urgency level). The resolution should cover at least half of the key points mentioned in the stances.
- **Score 7-9**: If the resolution provides detailed, concrete and practically feasible implementation measures that fully strengthen and implement supportive stances (for opposing stances, score high if the resolution effectively addresses and resolves the opposition's concerns) without any dilution or compromise. The resolution must comprehensively address ALL points raised in the stances, with higher scores for more detailed coverage of each point.

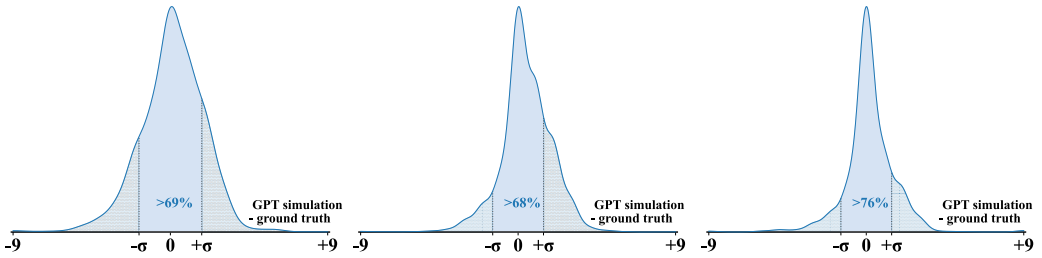
The experiments in subsection 4.1 and Appendix D.1 demonstrate a strong consistency between our evaluation method and the real voting results.

D More Experimental Results

In this section, we will illustrate more experimental results, especially more simulated consistency results of our open-ended evaluation framework and detailed performance on all the fine-grained topics.

D.1 Detailed Simulated Evaluation Consistency Results

In this section, we will provide a more detailed presentation and supplement to the experimental results from subsection 4.1.



(a) Error Distribution for the 7th Term (b) Error Distribution for the 8th Term (c) Error Distribution for the 9th Term

Figure 10: Error distribution on the GPT-4o mini simulated votes and the ground truth for the 7th (2009 - 2014), 8th (2014 - 2019), and half of the 9th (2019 - 2022) parliament terms.

As shown in Figure 10, we employed the same method as in subsection 4.1 to further illustrate the consistency between the simulated voting results of our open-ended evaluation framework and the actual voting results for the 7th, 8th, and 9th terms of the European Parliament.

The error is calculated by subtracting the ground truth voting score from the simulated score of GPT-4o mini. It can be observed that for each term of the parliament, most of our simulation results fall within the ground truth's σ range: 69% for the 7th, 68% for the 8th, and 76% for the 9th. From this more detailed error analysis, we can also see that in almost every parliament, GPT-4o mini tends to overestimate slightly rather than underestimate, particularly evident in the 8th parliament. This may be related to factors such as the sycophancy of LLMs [72, 73].

D.2 Detailed Fine-grained Topics Results

In this section, we will demonstrate the performance of different LLMs on each fine-grained topic, as shown in Table 9 to Table 13.

As shown in Table 9, in the economic topics of the euro, including international trade, internal market & consumer protection, employment & social affairs, and economic & monetary affairs (with some topic names abbreviated for convenience in the table), there are significant performance differences among various LLMs. Overall, Qwen2.5-72B and DeepSeek-R1 perform the best, especially in SM tasks, achieving high pass rates of 0.78-0.94 and 0.85-0.93, respectively. In contrast, commercial models like GPT-4o and Gemini-2.5 show moderate performance. Notably, as task difficulty increases (such as with the 2/3M and Veto tasks), the performance of all models declines significantly. For example, the pass rate of Qwen2.5-72B in 2/3M tasks drops to 0.43-0.80, reflecting the limitations of LLMs under strict consensus requirements. In the Rawls and Util tasks, the general trend remains consistent. Additionally, the type of topic significantly impacts performance, with policy-related topics like employment & social affairs and economic & monetary affairs generally being more challenging than industry development topics like international trade and internal market & consumer protection, highlighting the challenges LLMs face with complex political issues.

Table 9: Performance of different LLMs on *EuroCon*'s Economic topic. The values in square brackets indicate the range of each metric, and all metrics follow the principle that higher values are better. The background color of the table cells deepens as the performance improves. The blue color scheme represents metrics in the 0-1 range, while the red color scheme represents metrics in the 0-9 range.

Topic	Model	SM [0-1] ↑			2/3M [0-1] ↑			VP [0-1] ↑			Rawls [0-9] ↑			Util [0-9] ↑		
		2	4	6	2	4	6	2	4	6	2	4	6	2	4	6
international trade	QWen-2.5-32B	0.73	0.57	0.57	0.46	0.41	0.31	0.59	0.43	0.44	3.10	2.69	1.57	5.79	5.69	5.20
	Llama-3.3-70B	0.67	0.64	0.56	0.45	0.39	0.18	0.58	0.48	0.46	3.65	2.31	1.34	5.99	6.07	5.32
	GPT-4o	0.82	0.70	0.61	0.52	0.48	0.25	0.69	0.57	0.51	4.40	2.92	1.51	6.50	5.97	5.21
	Gemini-2.5	0.73	0.79	0.84	0.50	0.61	0.51	0.67	0.60	0.70	4.38	3.24	2.57	6.37	6.47	6.34
	DeepSeek-R1	0.88	0.84	0.79	0.69	0.66	0.51	0.74	0.67	0.67	4.60	3.92	2.48	6.67	6.75	6.45
	QWen-2.5-72B	0.91	0.92	0.80	0.69	0.61	0.48	0.75	0.64	0.72	5.34	4.52	3.11	7.04	7.08	6.64
internal market	QWen-2.5-32B	0.78	0.70	0.86	0.50	0.51	0.57	0.57	0.64	0.66	4.44	3.97	2.55	6.41	6.69	6.24
	Llama-3.3-70B	0.84	0.75	0.77	0.65	0.56	0.43	0.69	0.59	0.50	5.40	4.13	2.50	6.95	6.84	6.40
	GPT-4o	0.84	0.82	0.77	0.72	0.57	0.48	0.74	0.67	0.57	5.66	4.10	2.52	7.29	6.53	5.84
	Gemini-2.5	0.85	0.87	0.91	0.70	0.67	0.70	0.71	0.80	0.70	5.84	5.08	3.39	7.16	7.25	7.39
	DeepSeek-R1	0.90	0.90	0.93	0.84	0.75	0.73	0.85	0.77	0.70	6.20	5.31	4.20	7.44	7.51	7.27
	QWen-2.5-72B	0.94	0.93	0.95	0.80	0.75	0.66	0.80	0.74	0.77	6.61	5.30	4.11	7.91	7.49	7.20
employment	QWen-2.5-32B	0.56	0.62	0.82	0.43	0.31	0.46	0.39	0.35	0.54	2.39	2.09	1.23	5.05	5.79	5.76
	Llama-3.3-70B	0.57	0.60	0.74	0.38	0.38	0.31	0.44	0.38	0.54	2.95	2.27	1.46	5.88	5.85	5.94
	GPT-4o	0.70	0.67	0.77	0.48	0.29	0.46	0.59	0.44	0.49	4.02	2.35	1.64	6.25	5.65	5.73
	Gemini-2.5	0.67	0.67	0.87	0.46	0.51	0.62	0.51	0.56	0.59	4.11	2.96	2.03	6.17	6.43	6.61
	DeepSeek-R1	0.75	0.82	0.82	0.56	0.45	0.69	0.64	0.58	0.74	5.13	3.35	2.10	6.57	6.70	6.70
	QWen-2.5-72B	0.90	0.84	0.85	0.56	0.55	0.67	0.62	0.64	0.59	5.23	3.64	2.97	6.80	6.80	6.92
economic affairs	QWen-2.5-32B	0.57	0.55	0.61	0.32	0.34	0.25	0.38	0.33	0.45	2.74	1.93	0.96	5.31	5.46	4.82
	Llama-3.3-70B	0.61	0.58	0.55	0.36	0.30	0.16	0.32	0.34	0.35	2.59	1.79	0.82	5.49	5.49	5.12
	GPT-4o	0.72	0.67	0.60	0.45	0.37	0.22	0.46	0.37	0.42	3.39	2.22	1.15	6.02	5.48	5.10
	Gemini-2.5	0.68	0.73	0.76	0.43	0.53	0.41	0.43	0.49	0.56	3.30	2.63	1.81	5.85	6.08	6.00
	DeepSeek-R1	0.83	0.86	0.83	0.51	0.58	0.44	0.59	0.56	0.62	4.47	3.35	2.23	6.43	6.45	6.36
	QWen-2.5-72B	0.87	0.88	0.78	0.58	0.50	0.43	0.57	0.57	0.64	4.90	3.67	2.54	6.64	6.70	6.44
Average	Qwen2.5-32B	0.65	0.60	0.68	0.41	0.38	0.35	0.48	0.41	0.50	3.12	2.52	1.43	5.61	5.79	5.30
	Llama-3.3-70B	0.67	0.63	0.62	0.44	0.38	0.24	0.48	0.43	0.43	3.49	2.42	1.33	5.97	5.94	5.52
	GPT-4o	0.77	0.70	0.66	0.53	0.42	0.31	0.60	0.49	0.48	4.21	2.75	1.55	6.44	5.82	5.35
	Gemini-2.5	0.73	0.76	0.82	0.51	0.57	0.52	0.56	0.59	0.63	4.22	3.27	2.30	6.30	6.45	6.42
	DeepSeek-R1	0.84	0.86	**0.83**	0.63	**0.61**	**0.54**	**0.69**	**0.63**	**0.67**	4.95	3.85	2.61	6.72	6.75	6.59
	Qwen2.5-72B	**0.90**	**0.89**	**0.83**	**0.65**	0.58	0.52	0.67	**0.63**	**0.67**	**5.41**	**4.18**	**3.02**	**7.03**	**6.96**	**6.69**

Table 10: Performance of different LLMs on *EuroCon*'s Industry topic. The values in square brackets indicate the range of each metric, and all metrics follow the principle that higher values are better. The background color of the table cells deepens as the performance improves. The blue color scheme represents metrics in the 0-1 range, while the red color scheme represents metrics in the 0-9 range.

Topic	Model	SM [0-1] ↑			2/3M [0-1] ↑			VP [0-1] ↑			Rawls [0-9] ↑			Util [0-9] ↑		
		2	4	6	2	4	6	2	4	6	2	4	6	2	4	6
agriculture	QWen-2.5-32B	0.63	0.68	0.74	0.48	0.42	0.44	0.52	0.57	0.44	3.33	2.38	1.89	5.27	5.76	6.07
	Llama-3.3-70B	0.63	0.62	0.70	0.50	0.50	0.41	0.52	0.50	0.44	4.17	3.02	1.44	6.53	6.08	5.60
	GPT-4o	0.72	0.80	0.81	0.54	0.53	0.41	0.63	0.68	0.67	5.15	3.50	2.44	6.86	6.59	6.09
	Gemini-2.5	0.72	0.80	0.89	0.59	0.62	0.67	0.50	0.78	0.70	5.39	3.92	2.81	7.02	6.79	6.77
	DeepSeek-R1	0.80	0.85	0.96	0.67	0.68	0.59	0.61	0.78	0.74	5.67	4.22	2.93	7.27	7.09	6.66
	QWen-2.5-72B	0.91	0.93	0.96	0.74	0.68	0.56	0.74	0.85	0.81	5.89	4.58	3.19	7.16	7.16	6.91
fisheries	QWen-2.5-32B	0.80	0.78	0.86	0.57	0.57	0.57	0.61	0.57	0.66	4.51	4.69	3.41	7.01	6.73	6.52
	Llama-3.3-70B	0.86	0.77	0.75	0.61	0.51	0.39	0.74	0.68	0.61	5.43	4.15	2.59	7.07	7.07	6.16
	GPT-4o	0.88	0.82	0.75	0.74	0.62	0.64	0.80	0.71	0.66	5.94	4.94	3.32	7.31	6.94	6.09
	Gemini-2.5	0.90	0.89	0.86	0.67	0.72	0.68	0.72	0.78	0.70	5.91	5.37	4.91	7.14	7.49	7.36
	DeepSeek-R1	0.93	0.97	0.93	0.78	0.75	0.77	0.86	0.85	0.75	6.45	5.66	4.61	7.61	7.73	7.31
	QWen-2.5-72B	0.99	0.98	1.00	0.83	0.74	0.75	0.88	0.88	0.77	6.80	6.09	4.91	7.72	7.92	7.49
transport & tourism	QWen-2.5-32B	0.66	0.75	0.62	0.40	0.44	0.30	0.53	0.55	0.42	3.90	3.53	1.98	5.59	5.97	4.89
	Llama-3.3-70B	0.71	0.76	0.55	0.47	0.49	0.30	0.57	0.53	0.35	4.40	2.93	1.65	6.43	6.25	5.59
	GPT-4o	0.84	0.78	0.72	0.62	0.51	0.38	0.74	0.58	0.42	5.03	3.40	2.33	6.81	6.20	5.69
	Gemini-2.5	0.76	0.82	0.80	0.66	0.64	0.42	0.62	0.67	0.65	5.02	3.89	2.55	6.75	6.80	6.63
	DeepSeek-R1	0.86	0.89	0.82	0.64	0.78	0.60	0.72	0.75	0.72	5.88	4.73	3.12	7.34	7.04	6.71
	QWen-2.5-72B	0.91	0.95	0.78	0.74	0.69	0.55	0.84	0.73	0.75	6.24	5.11	3.73	7.35	7.20	7.01
industry & energy	QWen-2.5-32B	0.72	0.66	0.73	0.48	0.45	0.55	0.52	0.52	0.53	3.51	2.77	2.27	5.50	6.21	5.64
	Llama-3.3-70B	0.72	0.66	0.73	0.42	0.40	0.39	0.51	0.52	0.55	3.83	2.55	2.16	6.40	6.44	6.21
	GPT-4o	0.86	0.72	0.73	0.54	0.43	0.55	0.62	0.52	0.53	4.54	3.28	2.55	6.61	6.05	5.56
	Gemini-2.5	0.83	0.83	0.88	0.52	0.55	0.67	0.51	0.62	0.69	4.39	4.23	3.41	6.77	6.71	6.94
	DeepSeek-R1	0.83	0.86	0.92	0.61	0.62	0.69	0.65	0.71	0.73	5.15	4.14	3.27	7.01	7.06	6.98
	QWen-2.5-72B	0.93	0.88	0.94	0.66	0.63	0.65	0.69	0.69	0.71	5.51	4.62	4.02	7.04	7.23	7.16
Average	Qwen2.5-32B	0.71	0.72	0.74	0.48	0.48	0.47	0.55	0.55	0.53	3.85	3.44	2.44	5.91	6.22	5.77
	Llama-3.3-70B	0.74	0.71	0.69	0.50	0.47	0.37	0.59	0.56	0.50	4.48	3.19	2.03	6.62	6.51	5.94
	GPT-4o	0.84	0.78	0.75	0.61	0.52	0.51	0.70	0.62	0.56	5.17	3.83	2.69	6.90	6.44	5.83
	Gemini-2.5	0.81	0.84	0.86	0.61	0.64	0.61	0.59	0.71	0.69	5.16	4.42	3.51	6.92	6.97	6.95
	DeepSeek-R1	0.86	0.90	0.91	0.68	**0.71**	**0.68**	0.72	0.77	0.74	5.79	4.74	3.54	7.31	7.25	6.95
	Qwen2.5-72B	**0.94**	**0.93**	**0.92**	**0.74**	0.68	0.64	**0.79**	**0.78**	**0.76**	**6.12**	**5.16**	**4.05**	**7.33**	**7.41**	**7.17**

The results in Table 10 demonstrate the performance differences of various LLMs on the *EuroCon* benchmark's Industry theme, which includes four sub-themes: agriculture, fisheries, transport & tourism, and industry, research & energy. The results indicate that the pass rates for agriculture and fisheries topics are generally higher, possibly due to the relatively clear stance conflicts in these traditional industry topics, making it easier to reach compromises. In contrast, the performance on transport and tourism topics is slightly weaker (e.g., Llama-3.3-70B scores only 0.30 on the 2/3M task), suggesting that when it comes to cross-regional resource allocation, LLMs struggle to effectively safeguard the interests of the most vulnerable parties, directly related to the complexity of multiple stakeholders. Notably, topics like industry, research & energy, which involve technological transformation and policy coordination, score the lowest in Rawls tasks, reflecting the limitations of LLMs in handling issues at the intersection of technology and policy.

The results in Table 11 focus on budget-related topics in *EuroCon*, including development, regional development, budget, and budgetary control. They reveal distinct performance differences of LLMs on fiscal topics. Regional development topics demonstrate the highest consensus-building ability, likely due to their involvement with specific infrastructure projects, where benefit distribution schemes are easier to quantify and compromise on. In contrast, pure budget allocation topics (such as the budget fine-grained topic) show the weakest performance, reflecting the difficulty LLMs face in balancing multiple demands in abstract fiscal rule-making. Notably, budget control topics perform well in the Util task, indicating that LLMs are better at achieving technical consensus through quantifying overall benefits (such as fund usage efficiency) rather than resolving conflicts over political principles.

Table 11: Performance of different LLMs on *EuroCon*'s Budget topic. The values in square brackets indicate the range of each metric, and all metrics follow the principle that higher values are better. The background color of the table cells deepens as the performance improves. The blue color scheme represents metrics in the 0-1 range, while the red color scheme represents metrics in the 0-9 range.

Topic	Model	SM [0-1] ↑			2/3M [0-1] ↑			VP [0-1] ↑			Rawls [0-9] ↑			Util [0-9] ↑		
		2	4	6	2	4	6	2	4	6	2	4	6	2	4	6
development	QWen-2.5-32B	0.72	0.48	0.85	0.44	0.35	0.50	0.47	0.45	0.65	3.06	1.97	1.40	5.08	5.24	5.38
	Llama-3.3-70B	0.47	0.61	0.55	0.19	0.26	0.35	0.56	0.35	0.55	3.84	1.55	0.70	5.97	5.52	5.93
	GPT-4o	0.84	0.65	0.55	0.38	0.52	0.50	0.72	0.45	0.70	4.47	2.68	1.35	6.52	5.62	5.40
	Gemini-2.5	0.69	0.81	0.80	0.56	0.45	0.55	0.69	0.61	0.75	3.50	3.19	2.35	6.17	6.23	6.67
	DeepSeek-R1	0.81	0.84	0.90	0.62	0.68	0.65	0.75	0.61	0.80	5.00	3.52	2.40	6.80	6.68	6.83
	QWen-2.5-72B	0.84	0.90	0.90	0.72	0.61	0.75	0.78	0.71	0.65	5.47	4.06	2.85	7.16	6.73	6.79
regional development	QWen-2.5-32B	0.83	0.90	0.74	0.56	0.45	0.49	0.54	0.55	0.57	3.83	2.55	1.83	5.95	6.01	5.21
	Llama-3.3-70B	0.71	0.86	0.69	0.52	0.45	0.34	0.54	0.55	0.49	3.06	2.02	1.00	6.25	6.31	5.98
	GPT-4o	0.73	0.90	0.74	0.58	0.52	0.46	0.69	0.57	0.54	4.73	2.88	2.29	6.61	6.49	5.89
	Gemini-2.5	0.83	0.95	0.91	0.58	0.64	0.69	0.67	0.71	0.80	4.04	3.33	2.66	6.70	6.74	6.98
	DeepSeek-R1	0.85	0.98	0.94	0.71	0.81	0.60	0.73	0.74	0.71	5.54	4.38	3.14	7.08	6.90	7.12
	QWen-2.5-72B	0.85	0.98	0.97	0.65	0.67	0.60	0.81	0.71	0.69	5.56	4.38	3.86	7.09	7.25	7.14
budget	QWen-2.5-32B	0.62	0.62	0.60	0.27	0.38	0.32	0.33	0.42	0.41	1.85	1.77	1.07	4.88	5.50	5.16
	Llama-3.3-70B	0.59	0.64	0.52	0.29	0.38	0.27	0.29	0.51	0.35	2.46	1.53	0.99	5.46	5.94	5.24
	GPT-4o	0.68	0.71	0.61	0.38	0.48	0.28	0.38	0.56	0.43	3.26	2.05	1.28	5.99	5.81	5.08
	Gemini-2.5	0.72	0.83	0.85	0.46	0.54	0.51	0.46	0.67	0.55	3.35	2.99	1.97	5.97	6.63	6.30
	DeepSeek-R1	0.82	0.90	0.87	0.49	0.61	0.52	0.54	0.70	0.61	3.98	3.42	2.37	6.14	6.82	6.50
	QWen-2.5-72B	0.85	0.91	0.88	0.57	0.64	0.43	0.54	0.74	0.64	4.96	3.61	2.69	6.73	7.02	6.52
budgetary control	QWen-2.5-32B	0.60	0.72	0.87	0.29	0.37	0.59	0.36	0.47	0.65	2.35	1.98	1.67	4.55	5.61	6.00
	Llama-3.3-70B	0.61	0.61	0.74	0.33	0.37	0.44	0.33	0.42	0.54	2.70	1.61	0.95	5.61	5.66	5.97
	GPT-4o	0.77	0.75	0.77	0.44	0.46	0.54	0.47	0.50	0.55	3.68	2.22	1.71	6.12	5.96	5.74
	Gemini-2.5	0.68	0.82	0.93	0.46	0.56	0.74	0.47	0.63	0.75	3.23	2.76	2.50	6.06	6.55	7.04
	DeepSeek-R1	0.83	0.89	0.95	0.55	0.64	0.79	0.65	0.68	0.78	4.76	3.80	3.31	6.73	7.04	7.32
	QWen-2.5-72B	0.85	0.89	0.96	0.61	0.66	0.75	0.68	0.68	0.77	4.91	3.92	3.44	6.93	7.07	7.29
Average	Qwen2.5-32B	0.63	0.70	0.83	0.31	0.38	0.55	0.37	0.47	0.61	2.39	1.97	1.59	4.73	5.60	5.82
	Llama-3.3-70B	0.60	0.63	0.70	0.33	0.37	0.41	0.35	0.44	0.52	2.73	1.62	0.95	5.64	5.75	5.88
	GPT-4o	0.75	0.75	0.74	0.43	0.47	0.50	0.48	0.51	0.54	3.70	2.25	1.68	6.15	5.95	5.65
	Gemini-2.5	0.70	0.83	0.92	0.47	0.56	0.70	0.49	0.64	0.73	3.32	2.86	2.44	6.09	6.56	6.93
	DeepSeek-R1	0.83	**0.90**	0.94	0.55	**0.65**	**0.74**	0.64	0.68	**0.76**	4.67	3.75	3.15	6.64	6.97	**7.19**
	Qwen2.5-72B	**0.85**	**0.90**	**0.95**	**0.61**	**0.65**	0.70	**0.66**	**0.70**	0.74	**4.99**	**3.90**	**3.35**	**6.91**	**7.05**	7.16

Table 12: Performance of different LLMs on *EuroCon*'s Security topic. The values in square brackets indicate the range of each metric, and all metrics follow the principle that higher values are better. The background color of the table cells deepens as the performance improves. The blue color scheme represents metrics in the 0-1 range, while the red color scheme represents metrics in the 0-9 range.

Model	Topic	SM [0-1] ↑			2/3M [0-1] ↑			VP [0-1] ↑			Rawls [0-9] ↑			Util [0-9] ↑		
		2	4	6	2	4	6	2	4	6	2	4	6	2	4	6
public health	QWen-2.5-32B	0.70	0.74	0.79	0.46	0.44	0.55	0.53	0.56	0.69	3.62	3.15	2.70	5.77	6.33	6.13
	Llama-3.3-70B	0.72	0.76	0.76	0.48	0.52	0.43	0.58	0.57	0.55	3.91	2.93	2.04	6.27	6.44	6.10
	GPT-4o	0.79	0.79	0.79	0.56	0.55	0.51	0.67	0.63	0.65	4.81	3.40	2.62	6.81	6.41	5.92
	Gemini-2.5	0.80	0.86	0.94	0.60	0.62	0.68	0.65	0.70	0.81	4.67	4.07	3.05	6.50	6.94	7.16
	DeepSeek-R1	0.85	0.91	0.94	0.69	0.70	0.71	0.72	0.77	0.78	5.76	4.51	3.87	7.00	7.30	7.34
	QWen-2.5-72B	0.87	0.93	0.96	0.69	0.72	0.75	0.76	0.78	0.82	5.71	4.74	4.29	7.24	7.39	7.29
foreign & security	QWen-2.5-32B	0.53	0.47	0.65	0.31	0.31	0.29	0.32	0.34	0.43	1.86	1.62	0.88	4.06	4.78	5.07
	Llama-3.3-70B	0.55	0.52	0.51	0.31	0.30	0.16	0.35	0.38	0.37	2.44	1.42	0.74	4.95	5.11	5.38
	GPT-4o	0.66	0.58	0.61	0.37	0.35	0.31	0.43	0.43	0.41	3.51	2.13	1.22	5.66	5.56	5.09
	Gemini-2.5	0.67	0.71	0.81	0.44	0.45	0.45	0.46	0.50	0.58	3.56	2.67	1.88	5.69	5.93	6.26
	DeepSeek-R1	0.76	0.75	0.82	0.50	0.53	0.52	0.54	0.58	0.57	4.36	3.35	2.40	6.18	6.13	6.39
	QWen-2.5-72B	0.77	0.75	0.79	0.50	0.47	0.51	0.57	0.55	0.62	4.47	3.19	2.29	6.26	6.40	6.46
Average	Qwen2.5-32B	0.58	0.56	0.70	0.36	0.35	0.39	0.39	0.41	0.53	2.39	2.10	1.58	4.58	5.26	5.48
	Llama-3.3-70B	0.60	0.60	0.60	0.36	0.37	0.26	0.42	0.44	0.44	2.89	1.89	1.23	5.35	5.53	5.65
	GPT-4o	0.70	0.65	0.68	0.43	0.41	0.38	0.50	0.49	0.50	3.90	2.53	1.76	6.01	5.82	5.41
	Gemini-2.5	0.71	0.76	0.86	0.49	0.50	0.54	0.52	0.56	0.67	3.90	3.10	2.32	5.93	6.25	6.60
	DeepSeek-R1	0.79	**0.80**	**0.87**	**0.55**	**0.59**	0.59	0.60	**0.64**	0.65	4.78	**3.72**	2.96	6.42	6.49	6.75
	Qwen2.5-72B	**0.80**	**0.80**	0.86	**0.55**	0.55	**0.60**	**0.63**	0.62	**0.69**	**4.85**	3.67	**3.05**	**6.56**	**6.71**	**6.77**

Table 12 illustrates the significant differences among various LLMs on two fine-grained topics under the Security theme: environment & public health and foreign & security policy. The environment & public health topic demonstrates the highest consensus-building ability, likely due to its technical and non-political nature, which allows models to reconcile different positions more easily. In contrast, the foreign & security policy topic performs the weakest across all task settings, highlighting the limitations of LLMs when handling highly sensitive issues like national sovereignty and geopolitics. Notably, in the Rawls task, the environment & public health topic scores significantly higher than foreign & security policy, indicating that LLMs achieve better consensus in healthcare fields, while struggling to overcome established power structures in complex political issues related to national security. This disparity supports the conclusion throughout the text regarding how topic complexity affects model performance, especially with the value conflicts and zero-sum nature unique to security topics.

Table 13: Performance of different LLMs on *EuroCon*'s Civil Rights topic. The values in square brackets indicate the range of each metric, and all metrics follow the principle that higher values are better. The background color of the table cells deepens as the performance improves. The blue color scheme represents metrics in the 0-1 range, while the red color scheme represents metrics in the 0-9 range.

Topic	Model	SM [0-1] ↑			2/3M [0-1] ↑			VP [0-1] ↑			Rawls [0-9] ↑			Util [0-9] ↑		
		2	4	6	2	4	6	2	4	6	2	4	6	2	4	6
culture & education	QWen-2.5-32B	0.78	0.67	0.69	0.54	0.36	0.50	0.51	0.33	0.65	3.10	1.67	1.04	4.91	5.51	4.96
	Llama-3.3-70B	0.76	0.72	0.81	0.44	0.42	0.54	0.46	0.39	0.58	2.49	1.36	1.08	5.79	5.88	6.39
	GPT-4o	0.85	0.86	0.77	0.61	0.53	0.62	0.68	0.53	0.65	3.88	2.50	2.00	6.28	6.42	6.43
	Gemini-2.5	0.85	0.78	0.92	0.51	0.61	0.73	0.61	0.61	0.81	3.80	3.36	2.23	6.48	6.64	7.01
	DeepSeek-R1	0.85	0.94	0.96	0.73	0.75	0.88	0.66	0.64	0.77	4.44	3.44	2.54	6.85	6.83	6.97
	QWen-2.5-72B	0.90	0.92	1.00	0.59	0.64	0.81	0.68	0.58	0.81	4.56	3.56	2.58	6.82	6.71	7.22
gender equality	QWen-2.5-32B	0.43	0.64	0.74	0.28	0.36	0.48	0.30	0.45	0.56	2.13	1.73	1.52	4.55	5.66	5.68
	Llama-3.3-70B	0.43	0.61	0.74	0.30	0.34	0.44	0.32	0.41	0.52	2.04	1.66	0.70	5.33	5.93	5.77
	GPT-4o	0.51	0.75	0.70	0.38	0.41	0.44	0.51	0.45	0.63	3.64	2.07	1.93	6.14	6.03	5.54
	Gemini-2.5	0.55	0.77	0.89	0.36	0.57	0.67	0.36	0.55	0.67	3.68	2.34	2.30	5.89	6.46	6.61
	DeepSeek-R1	0.68	0.86	0.96	0.47	0.50	0.74	0.47	0.55	0.70	3.91	3.30	2.93	6.55	6.54	6.64
	QWen-2.5-72B	0.70	0.84	0.89	0.51	0.66	0.59	0.36	0.55	0.70	3.70	3.20	2.07	6.45	6.89	6.79
civil liberties	QWen-2.5-32B	0.62	0.61	0.78	0.37	0.34	0.34	0.45	0.48	0.60	2.20	2.13	1.58	4.43	5.16	5.40
	Llama-3.3-70B	0.61	0.59	0.64	0.36	0.35	0.29	0.48	0.39	0.46	2.79	1.77	1.28	5.67	5.53	5.44
	GPT-4o	0.78	0.70	0.68	0.44	0.39	0.39	0.57	0.49	0.52	3.88	2.52	1.65	6.20	5.70	5.44
	Gemini-2.5	0.69	0.73	0.83	0.56	0.51	0.53	0.57	0.57	0.70	3.64	2.48	2.19	5.97	6.18	6.39
	DeepSeek-R1	0.82	0.81	0.87	0.61	0.57	0.63	0.68	0.68	0.72	4.52	3.57	2.80	6.74	6.60	6.87
	QWen-2.5-72B	0.86	0.81	0.88	0.59	0.54	0.51	0.67	0.65	0.69	4.79	3.69	3.00	6.68	6.80	6.95
constitutional affairs	QWen-2.5-32B	0.60	0.63	0.61	0.35	0.28	0.34	0.53	0.44	0.50	1.47	1.48	1.20	4.85	5.23	4.97
	Llama-3.3-70B	0.68	0.52	0.55	0.37	0.22	0.25	0.49	0.44	0.41	2.32	1.22	0.77	5.39	5.44	5.35
	GPT-4o	0.74	0.54	0.64	0.53	0.37	0.32	0.61	0.46	0.41	3.19	1.57	1.23	5.84	5.72	4.91
	Gemini-2.5	0.72	0.74	0.80	0.35	0.52	0.43	0.54	0.59	0.61	3.81	2.17	2.18	5.62	6.18	6.08
	DeepSeek-R1	0.86	0.81	0.86	0.56	0.56	0.52	0.63	0.67	0.59	3.98	2.81	1.91	6.11	6.23	6.35
	QWen-2.5-72B	0.86	0.85	0.86	0.56	0.50	0.45	0.65	0.63	0.66	4.46	3.11	2.68	6.47	6.48	6.51
legal affairs	QWen-2.5-32B	0.66	0.82	0.81	0.51	0.51	0.54	0.53	0.59	0.65	3.97	3.51	2.70	5.86	6.24	6.60
	Llama-3.3-70B	0.75	0.78	0.84	0.47	0.59	0.32	0.59	0.61	0.65	4.44	3.10	2.05	6.54	6.39	6.42
	GPT-4o	0.85	0.80	0.84	0.63	0.63	0.57	0.64	0.61	0.68	5.36	3.88	3.27	6.92	6.73	6.17
	Gemini-2.5	0.75	0.88	0.86	0.54	0.65	0.57	0.59	0.71	0.70	4.69	4.16	3.43	6.75	6.95	7.24
	DeepSeek-R1	0.86	0.94	0.92	0.71	0.80	0.73	0.76	0.73	0.78	5.73	4.63	4.05	7.03	7.51	7.21
	QWen-2.5-72B	0.92	0.92	0.95	0.66	0.80	0.76	0.75	0.73	0.78	5.68	4.80	4.43	7.46	7.55	7.37
Average	Qwen2.5-32B	0.62	0.65	0.74	0.40	0.36	0.40	0.46	0.47	0.59	2.46	2.12	1.62	4.80	5.44	5.49
	Llama-3.3-70B	0.64	0.62	0.68	0.38	0.37	0.33	0.48	0.44	0.50	2.85	1.82	1.22	5.74	5.73	5.71
	GPT-4o	0.76	0.71	0.71	0.50	0.44	0.43	0.60	0.50	0.55	3.98	2.50	1.89	6.26	5.98	5.57
	Gemini-2.5	0.71	0.77	0.85	0.49	0.55	0.56	0.55	0.60	0.69	3.86	2.76	2.40	6.09	6.38	6.56
	DeepSeek-R1	0.82	**0.85**	**0.90**	**0.62**	**0.61**	**0.67**	**0.65**	**0.66**	**0.71**	4.54	3.55	2.81	6.68	6.69	6.81
	Qwen2.5-72B	**0.85**	**0.85**	**0.90**	0.59	0.60	0.58	0.64	0.64	**0.71**	**4.71**	**3.68**	**3.01**	**6.76**	**6.86**	**6.95**

Table 13 focuses on these five subtopics: culture & education, gender equality, civil liberties, justice & home affairs, constitutional & inter-institutional affairs, and legal affairs. These fine-grained topics reveal significant differences in how LLMs handle various Civil Rights issues. The topic of culture & education shows the strongest consensus-building ability, possibly because its relatively neutral cultural attributes make it easier for models to find compromise solutions. In contrast, the topic of constitutional affairs performs the weakest, reflecting the difficulty LLMs face in overcoming

opposing stances when fundamental constitutional principles are involved. Notably, the gender equality topic exhibits the most fluctuation in scores on the Rawls task (Qwen2.5-32B scores only 1.52 while DeepSeek-R1 reaching 3.91), indicating that this issue is the most sensitive to the models' value orientations. Meanwhile, the legal affairs topic performs best in the Util task (Qwen2.5-72B scores 7.55), confirming that LLMs may be more adept at maximizing benefits through procedural justice and technical terms.

E Case Study

In this section, we will demonstrate specific cases from our evaluation process by examining two aspects: the differences in political consensus finding capability among different LLMs on the same issue and the capability differences of the same LLM in different parliamentary settings (CAUTION FOR THE AI-GENERATED CONTENT). This will more clearly demonstrate the potential of using *EuroCon* to assess LLMs' ability to find political consensus.

E.1 Case Study: LLM Performances

Here, we present an example issue, illustrating its title, background and stances of each party. We then compare the response performance of different LLMs on this example. By comparing Response 1.1 and Response 1.2, we find that both models demonstrate strong support for the major political party (ALDE). However, Qwen2.5-72B clearly excels at reconciling the positions of the other party (EFD). For instance, Qwen2.5-72B's responses repeatedly emphasize "safeguarding national sovereignty and the integrity of border control", explicitly aligning this stance with European solidarity to directly address EFD's core concerns. This approach demonstrates a more confrontational yet compromising stance. Additionally, Qwen2.5-72B employs technical terminology (e.g., "strong external border management support" and "coordinated approach to Schengen zone challenges") to depoliticize sensitive sovereignty issues. Consequently, Qwen2.5-72B achieves higher alignment scores with EFD compared to Gemini-2.5.

We also found that the task becomes easier as the number of parties increases in majority voting, as demonstrated by comparing Response 1.1 and Response 1.3. This is because polarized stances have limited room for compromise when there are limited parties involved.

Topic: Civil Liberties Justice & Home Affairs

Title:
European Refugee Fund for the period 2008 to 2013 (amendment of Decision No 573/2007/EC): REPORT on the proposal for a decision of the European Parliament and of the Council amending Decision No 573/2007/EC establishing the European Refugee Fund for the period 2008 to 2013 as part of the General programme "Solidarity and Management of Migration Flows" and repealing Council Decision 2004/904/EC.
Background:
The European Refugee Fund (2008-2013) was established under Decision 573/2007/EC to support member states in asylum and migration management, forming part of the General Programme "Solidarity and Management of Migration Flows". The Parliament will discuss amendments to the fund's implementation framework and financial provisions to address operational challenges during its designated period.

Stances

EFD: EFD rejects the proposal's alignment with Lisbon Treaty frameworks as absolutely conflicting with national sovereignty principles. EFD opposes the issue's solidarity mechanisms for failing to address systemic border control failures highlighted in Schengen area operations. EFD rejects provisions encouraging refugee resettlement, citing unresolved risks of economic migration exploitation under current EU policies. EFD absolutely conflicts with the issue's expanded asylum commitments, emphasizing Turkey's potential accession as exacerbating uncontrolled migration pathways. EFD opposes the legislative continuity

approach, demanding treaty re-ratification to halt further EU integration as outlined in Article 294(3) TFEU procedures.
ALDE: ALDE strongly agrees the expansion of refugee resettlement programs through enhanced ERF allocations, aligning perfectly with ALDE's longstanding commitment to European solidarity and human rights protection. ALDE approves directing funds to local authorities for integration infrastructure, effectively advancing ALDE's priority of empowering municipalities to deliver housing and social services. ALDE endorses mandatory involvement of cities in implementation structures, recognizing their frontline role in refugee integration as emphasized in ALDE's local governance principles. ALDE commends the focus on vulnerable groups like women and children, mirroring ALDE's 2023 manifesto commitments to gender-sensitive asylum policies. ALDE proposes adding explicit safeguards in Article 4 to prevent national budget absorption, ensuring funds reach grassroots integration projects without opposing the overall funding mechanism.

Qwen2.5-72B *Simple Majority (7.0)* Response 1.1
EFD (weight: 0.4 score: 4) ALDE (weight: 0.6 score: 9)

The European Parliament raised its position on the proposal for a decision of the European Parliament and of the Council amending Decision No 573/2007/EC establishing the European Refugee Fund for the period 2008 to 2013 as part of the General Programme "Solidarity and Management of Migration Flows" and repealing Council Decision 2004/904/EC, recognizing the importance of balanced and principled approaches to asylum and migration management. The Parliament acknowledges the need for solidarity mechanisms and the expansion of refugee resettlement programs, aligning with the ALDE party's commitment to human rights and local integration, while addressing the EFD party's concerns about national sovereignty and border control. The Parliament supports enhanced ERF allocations for local authorities to develop integration infrastructure, ensuring that funds are directed to grassroots projects with explicit safeguards to prevent national budget absorption, as proposed by ALDE. The Parliament emphasizes the role of cities in implementation structures, recognizing their frontline role in refugee integration, and commits to focusing on vulnerable groups like women and children, in line with ALDE's priorities. However, the Parliament also stresses the importance of robust border control measures and the need to address systemic failures in the Schengen area, as highlighted by EFD, to prevent economic migration exploitation. The Parliament further calls for a cautious approach to expanded asylum commitments, particularly in light of Turkey's potential accession, and demands a comprehensive strategy to manage migration pathways. The Parliament underscores the necessity of maintaining national sovereignty and the integrity of border control, advocating for a balanced approach that respects both European solidarity and national interests. The resolution is supported by more than 50% of the Parliament members, reflecting the proportional stances of the EFD (40.0%) and ALDE (60.0%) parties.

Gemini-2.5 *Simple Majority (5.4)* Response 1.2
EFD (weight: 0.4 score: 0) ALDE (weight: 0.6 score: 9)

The European Parliament raised its position on the proposal for a decision amending Decision No 573/2007/EC establishing the European Refugee Fund for the period 2008 to 2013, acknowledging the importance of supporting member states in migration management under the General programme "Solidarity and Management of Migration Flows" while recognizing concerns regarding implementation and effectiveness, and strongly supporting the expansion of refugee resettlement programs through enhanced ERF allocations as a demonstration of European solidarity and commitment to human rights protection, although emphasizing the need for robust screening and measures to mitigate risks of economic migration exploitation and ensure genuine need, further approving the direction of funds towards local authorities to strengthen integration infrastructure, including housing and social services, thereby empowering municipalities, and endorsing the mandatory involvement of cities in implementation structures given their crucial frontline role in refugee integration, and commending the nec-

essary focus on vulnerable groups such as women and children within the fund's activities, while proposing explicit safeguards, potentially within Article 4, to prevent the absorption of fund allocations into national budgets, ensuring resources effectively reach grassroots integration projects, understanding that effective solidarity mechanisms must be complemented by support for robust external border management and coordinated approaches to address systemic challenges within the Schengen area, and proceeding within the established legislative framework to address operational challenges during the fund's designated period.

Qwen2.5-72B *Simple Majority (7.7)* Response 1.3
EFD (weight: 0.1 score: 3) ALDE (weight: 0.3 score: 9)
SD (weight: 0.4 score: 8) GUE_NGL (weight: 0.2 score: 8)

The European Parliament raised its position at first reading on the European Commission's proposal to amend Decision No 573/2007/EC establishing the European Refugee Fund for the period 2008 to 2013, informed by the Committee on Civil Liberties, Justice and Home Affairs' report, and incorporating the perspectives of the European Free Democratic (EFD), Alliance of Liberals and Democrats for Europe (ALDE), Social Democrats (SD), and the Confederal Group of the European United Left/Nordic Green Left (GUE_NGL). The Parliament emphasized the need to balance the principles of national sovereignty and EU solidarity, recognizing the EFD's concerns about the alignment with the Lisbon Treaty and the potential impact on national border control. To address these concerns, the Parliament proposed a detailed review of the Schengen area operations to ensure robust border management and prevent economic migration exploitation. The Parliament strongly supports the expansion of refugee resettlement programs and the enhanced allocation of funds to local authorities for integration infrastructure, aligning with ALDE's commitment to human rights and municipal empowerment. It endorses the mandatory involvement of cities in implementation structures and proposes explicit safeguards in Article 4 to prevent national budget absorption, ensuring funds reach grassroots integration projects. The Parliament also strongly agrees with SD's emphasis on harmonizing resettlement programs through the European Asylum Support Office and enhancing parliamentary oversight to ensure compliance with humanitarian standards. It proposes reinforcing interinstitutional cooperation on delegated acts to prevent procedural delays while maintaining full support for the ERF's core objectives. The Parliament strongly approves the issue's alignment with the Lisbon Treaty's solidarity principles, as highlighted by GUE_NGL, and endorses enhanced parliamentary oversight to ensure that migration funds prioritize humanitarian needs over securitized systems. It proposes amending implementation guidelines to mandate rigorous data protection safeguards and auditing technical frameworks to avoid wasteful expenditures and ensure accountability in migration management tools. The Parliament underscores the necessity of strict adherence to free movement principles and urges mechanisms to address disproportionate border checks that undermine Schengen commitments. The resolution reflects a balanced approach to address the diverse concerns and priorities of the political parties, ensuring that the European Refugee Fund effectively supports member states in asylum and migration management.

E.2 Case Study: Different Parliament Settings

In the previous subsection, we observed that task difficulty tends to decrease as the number of parties increases. Now we examine another example that demonstrates additional effects of parliamentary settings.

By comparing Response 2.1 and Response 2.2, we notice a reversed effect concerning Rawlsianism settings. Viewing from the content of Response 2.1, it becomes evident that accommodating all four stances within a single resolution is more challenging.

By comparing Response 2.3 and Response 2.4, we observe that achieving a 2/3 Majority is a little bit more challenging than dealing with the setting of Veto Power. In this case, the 2/3 Majority requires accommodating three out of four stances, making it difficult to craft a resolution that is sensible and acceptable to a larger group.

Topic: Foreign & Security Policy

Title:
Negotiations of the EU-Georgia Association Agreement: REPORT containing the European Parliament's recommendations to the Council, the Commission and the EEAS on the negotiations of the EU-Georgia Association Agreement.
Background:
Ongoing EU-Georgia Association Agreement negotiations followed the 2008 Georgia-Russia conflict, existing Partnership and Cooperation Agreement (1999), Eastern Partnership initiatives, and ENP Action Plan commitments. The Parliament will discuss advancing the Association Agreement to deepen political-economic ties, including trade integration and addressing post-conflict territorial disputes.

Stances

EFD: EFD rejects the issue's assumption that EU influence can effectively stabilize Georgia, citing unresolved geopolitical tensions like Abkhazia's alignment with Russia as fundamentally disputeing with EFD's skepticism about EU capacity in the region. EFD opposes the prioritization of economic integration through DCFTA talks, arguing the proposal overlooks critical aspects of post-Soviet governance challenges and ingrained instability inconsistent with EFD's emphasis on sovereignty-first approaches. EFD fundamentally disputes with the issue's reliance on technical assistance for democratic reforms, asserting that Georgia's Soviet-era institutional legacies and civil unrest require deeper structural changes beyond EU frameworks. EFD criticizes the issue's failure to address Russian influence in breakaway regions as a direct contradiction to EFD's stance on prioritizing territorial integrity over aspirational trade alignment. EFD opposes the emphasis on EU-aligned legislative reforms, deeming it inconsistent with EFD's principle that Georgia's democratic development must precede external economic integration.
GREEN_EFA: GREEN_EFA strongly accepts the issue's integration of human rights and governance reforms, which aligns perfectly with their commitment to conflict accountability and sustainable development in EU partnerships. GREEN_EFA applauds the focus on Georgia's economic recovery through DCFTA conditions, urging additional EU technical assistance to accelerate labor rights alignment with ILO standards. GREEN_EFA proposes amending the issue to explicitly reference the Tagliavini Commission's findings on the 2008 war, enhancing historical clarity while maintaining full support for Georgia's territorial integrity. GREEN_EFA approves the emphasis on ICC cooperation as critical to addressing unresolved war crime allegations, matching their manifesto priorities on international justice mechanisms. GREEN_EFA highlights the need for Georgia to address the Norwegian Helsinki Committee's concerns through transparent investigations, reinforcing institutional reforms under the Agreement's governance pillar.
SD: SD strongly sanctions the issue's emphasis on advancing Georgia's economic reforms and alignment with EU standards, particularly in rule of law and social market economy, which aligns perfectly with SD's commitment to democratic governance and sustainable development. SD applauds the call for inclusive political dialogue and media accessibility for opposition parties, reflecting SD's priority of fostering democratic pluralism and transparency in EU partner states. SD proposes enhancing technical assistance to explicitly include minority integration programs, ensuring Georgia's reforms address social cohesion as highlighted in the debate. SD endorses the issue's focus on conflict mediation and WTO cooperation, critical for stabilizing Georgia's regional context and advancing economic recovery post-2008. SD emphasizes the need to accelerate visa liberalization to deepen people-to-people ties, a step congruent with the issue's human rights and governance objectives.
GUE_NGL: GUE_NGL opposes fundamentally the Agreement's neglect of territorial autonomy and democratic principles in South Ossetia, Abkhazia, and the omitted third autonomous unit, conflicting with their commitment to self-determination. GUE_NGL rejects the issue's prioritization of anti-Russia rhetoric over addressing Georgia's internal governance failures, inconsistent with their stance on peaceful conflict issue. GUE_NGL highlights the Agreement's omission of citizenship rigging concerns, undermining its credibility on rule-of-law

reforms central to the issue's goals. GUE_NGL condemns the assimilation-driven economic demands as premature, clashing with their principle of equitable integration over coercive alignment. GUE_NGL notes the proposal's silence on toxic waste disposal accountability, failing to align with their environmental justice priorities outlined in the issue.

Qwen2.5-72B *Rawlsianism 4 Parties (3.0)* Response 2.1
EFD (weight: n/a score: 3) GREEN_EFA (weight: n/a score: 7)
SD (weight: n/a score: 9) GUE_NGL (weight: n/a score: 4)

The European Parliament raised concerns over the ongoing negotiations of the EU-Georgia Association Agreement, emphasizing the need to balance economic integration with territorial integrity and democratic reforms. The Parliament recognizes the importance of advancing the Deep and Comprehensive Free Trade Area (DCFTA) to bolster Georgia's economic recovery and align with EU standards, but acknowledges EFD's skepticism about the EU's capacity to stabilize Georgia and the need to address unresolved geopolitical tensions, particularly in Abkhazia and South Ossetia. To mitigate these concerns, the Parliament proposes a phased approach to economic integration, prioritizing the resolution of territorial disputes and the reduction of Russian influence in breakaway regions. The Parliament supports GREEN_EFA's focus on human rights and governance reforms, including the integration of labor rights aligned with ILO standards and the explicit reference to the Tagliavini Commission's findings on the 2008 war. The Parliament also endorses SD's emphasis on inclusive political dialogue and media accessibility, and the need for enhanced technical assistance to include minority integration programs. To address GUE_NGL's concerns, the Parliament calls for a nuanced approach to territorial autonomy, ensuring that democratic principles are respected in South Ossetia and Abkhazia, and for the inclusion of measures to address internal governance failures and citizenship rigging. The Parliament further urges the EU to provide transparent and accountable mechanisms for addressing war crime allegations and toxic waste disposal, aligning with GUE_NGL's environmental justice priorities. The Parliament commits to accelerating visa liberalization to deepen people-to-people ties and to enhancing conflict mediation efforts, recognizing the critical role of the WTO in stabilizing Georgia's regional context. The Parliament stresses the importance of ensuring that all reforms and agreements are grounded in Georgia's democratic development and sovereignty, and calls for a comprehensive and balanced approach to the negotiations that maximizes benefits for all parties involved.

Qwen2.5-72B *Rawlsianism 2 Parties (8.0)* Response 2.2
SD (weight: n/a score: 8) GUE_NGL (weight: n/a score: 8)

The European Parliament raised concerns about the ongoing negotiations of the EU-Georgia Association Agreement, emphasizing the need to balance economic and political reforms with the principles of democratic governance and social cohesion. The Parliament strongly supports the advancement of Georgia's economic reforms and alignment with EU standards, particularly in the rule of law and the social market economy, recognizing these as crucial for sustainable development and democratic pluralism. It applauds the call for inclusive political dialogue and media accessibility for opposition parties, fostering transparency and democratic participation. To address the concerns of the GUE_NGL, the Parliament underscores the importance of respecting territorial autonomy and democratic principles in South Ossetia, Abkhazia, and any other autonomous units, advocating for a balanced approach that includes dialogue and cooperation with local authorities to ensure self-determination. The Parliament also highlights the need to address internal governance failures, including concerns over citizenship rigging, to enhance the credibility of rule-of-law reforms. To mitigate the risk of premature assimilation-driven economic demands, the Parliament proposes a phased approach to economic integration, focusing on equitable and inclusive growth that supports minority integration and social cohesion. The Parliament further emphasizes the importance of environmental justice, calling for the inclusion of provisions on toxic waste disposal accountability to align with the EU's environmental standards. The Parliament endorses the focus on conflict mediation and WTO cooperation, recognizing these as essential for

stabilizing Georgia's regional context and advancing economic recovery post-2008. It also calls for accelerated visa liberalization to deepen people-to-people ties, aligning with the human rights and governance objectives of the Agreement. The Parliament urges the Council, the Commission, and the EEAS to incorporate these recommendations into the negotiations to ensure a fair and comprehensive Association Agreement that benefits all stakeholders.

Qwen2.5-72B *2/3 Majority (6.6)* Response 2.3
EFD (weight: 0.2 score: 3) GREEN_EFA (weight: 0.4 score: 9)
SD (weight: 0.2 score: 9) GUE_NGL (weight: 0.2 score: 3)

The European Parliament raised its recommendations to the Council, the Commission, and the EEAS on the negotiations of the EU-Georgia Association Agreement, emphasizing the need to balance economic integration with political stability and territorial integrity by a weighted consensus of 20% from Party EFD, 40% from Party GREEN_EFA, 20% from Party SD, and 20% from Party GUE_NGL. The Parliament supports the advancement of Georgia's economic reforms and alignment with EU standards, particularly in the rule of law and social market economy, as proposed by Party SD, while also acknowledging the critical importance of addressing unresolved geopolitical tensions and Russian influence in breakaway regions, as highlighted by Party EFD. To address these concerns, the Parliament recommends that the EU provide targeted technical assistance for Georgia's democratic reforms, including labor rights alignment with ILO standards, as advocated by Party GREEN_EFA, and explicitly reference the findings of the Tagliavini Commission to enhance historical clarity and maintain full support for Georgia's territorial integrity. The Parliament further endorses the inclusion of minority integration programs and the acceleration of visa liberalization to deepen people-to-people ties, as proposed by Party SD, while also recognizing the need for conflict mediation and WTO cooperation to stabilize Georgia's regional context. In response to the concerns raised by Party GUE_NGL, the Parliament emphasizes the importance of respecting the principles of self-determination and internal governance, ensuring that economic demands are equitable and not coercive, and addressing environmental justice issues, such as toxic waste disposal accountability. The Parliament calls for a comprehensive approach that integrates human rights, governance reforms, and sustainable development, ensuring that the Agreement's goals are achieved through transparent and inclusive processes that respect Georgia's sovereignty and the well-being of all its citizens.

Qwen2.5-72B *Veto Power (5.0)* Response 2.4
EFD (weight: 0.2 score: 0) GREEN_EFA (weight: 0.4 score: 8)
SD (weight: 0.2 score: 7) GUE_NGL (weight: 0.2 score: 2)

The European Parliament raised strong support for the EU-Georgia Association Agreement, recognizing it as a critical framework for deepening political and economic ties, while emphasizing the need to address unresolved geopolitical tensions and the complexities of Georgia's post-Soviet governance challenges. The Parliament underscored the importance of Georgia's sovereignty and territorial integrity, particularly in the context of Russian influence in Abkhazia and South Ossetia, and called for a balanced approach that prioritizes these concerns over aspirational trade alignment. It acknowledged the potential of the Deep and Comprehensive Free Trade Area (DCFTA) to foster economic recovery and integration, but stressed that such efforts must be accompanied by comprehensive structural reforms and a focus on democratic development. The Parliament strongly supported the integration of human rights and governance reforms, including the alignment with ILO labor standards and the transparent investigation of concerns raised by the Norwegian Helsinki Committee. It emphasized the need for inclusive political dialogue and media accessibility for opposition parties, and proposed enhancing technical assistance to include minority integration programs to ensure social cohesion. The Parliament also highlighted the importance of conflict mediation and WTO cooperation to stabilize the region and accelerate visa liberalization to deepen people-to-people ties. However, it recognized the necessity of addressing internal governance failures and the need for equitable integration, cautioning against premature economic de-

mands and the potential for social and environmental dumping. The Parliament called for a nuanced approach that respects Georgia's sovereignty and emphasizes the importance of self-determination, while ensuring that the Agreement addresses the concerns of all regions, including the third autonomous unit. It urged the inclusion of provisions for citizenship rigging concerns and toxic waste disposal accountability to align with environmental justice priorities. The Parliament reiterated Georgia's European perspective and its strategic role in the Southern Corridor, advocating for constructive regional dialogue and the extension of the EU Monitoring Mission (EUMM) mandate to ensure long-term stability and security.

F Background of the European Parliament

The European Parliament is one of the principal legislative and supervisory bodies[30] of the European Union, which is composed of MEPs who are directly elected by citizens of European Union member states [74]. It plays a crucial role in shaping European Union policies, exercising legislative powers in cooperation with the Council of the European Union, approving the European Union budget, and overseeing the work of the European Commission [75].

A defining characteristic of the European Parliament is its multi-party, cross-cultural deliberative environment. Unlike national parliaments, the European Parliament brings together political groups that transcend national boundaries, fostering a pluralistic debate that integrates diverse political ideologies, from green and far-left parties to center-right and eurosceptic factions. This structure can more comprehensively reflect diverse political viewpoints.

Moreover, the European Parliament's deliberative procedures are relatively open and comprehensive, making its data more complete and easier to obtain. This openness to data and relatively complete records make it a valuable resource for political and academic research.

G Discussion and Limitations

As discussed in section 5, although *EuroCon* has demonstrated an excellent ability to evaluate LLMs in finding political consensus, we acknowledge that it still faces some limitations. In this section, we will discuss these in more detail.

Firstly, we introduced LLMs in the data cleaning process, which may lead to the introduction of its specific biases, as well as AI-generated content that contains risks or offensive language toward certain groups. Secondly, our work treats all political parties as a whole, but in reality, the parties in the European Parliament are inherently complex political groups with internal conflicts. This complexity can be considered in future work. Additionally, there is a risk of data leakage in our dataset. However, not only have we mitigated this effect by setting task configurations different from the real world, but our experiments also show that current state-of-the-art LLMs are not very effective at handling tasks that involve finding political consensus across different tasks. This suggests that the impact of data leakage might not be significant.

Moreover, regarding the setting of the veto system, in the real-world UNSC, there are five permanent members with veto power, whereas in our setup, only one party has veto power at a time. This setting can be improved in future work by increasing the number of parties with veto power under conditions involving more parties. Finally, for convenience of illustration, *EuroCon* only used the task settings defined in subsection 3.2 to generate one round of data. In fact, since generating task scenarios incurs no cost, we can customize a large number of test scenarios flexibly and diversely according to specific needs. This can further enable our work to be applied to broader research settings, such as Pareto improvements and multi-objective optimization research, as well as research on different deliberation algorithms, and our evaluation framework can even retain its algorithm-agnostic feature, which can also be considered in future work.

[30] https://european-union.europa.eu/institutions-law-budget/institutions-and-bodies/search-all-eu-institutions-and-bodies/european-parliament_en

H Ethical Statement and Disclaimer

In this section, we will discuss the copyright issues of the data sources in this paper, the potential social risks, and the statement regarding the proper use of the data in *EuroCon*.

H.1 Copyright of Data Sources

The data in this paper is sourced and organized from the official website of the European Parliament[31], HowTheyVote[32], and the VoteWatch Europe dataset [47]. Both the official website of the European Parliament and HowTheyVote allow the use of their data as long as the source is cited, while the VoteWatch Europe dataset follows the CC 4.0 license.

H.2 Potential Societal Impact and Statement on the Use of *EuroCon*

EuroCon, as an AI project with the potential to influence social governance processes, carries certain social risks. For instance, it might generate biased or offensive statements towards specific groups when producing consensus decisions. Additionally, the use of AI systems in social governance processes could have both short-term and long-term impacts. Short-term effects might include generating persuasive rhetoric or exploiting cognitive biases of government officials, such as the anchoring effect, thereby reinforcing legislators' existing biases. It could also lead to legislators becoming overly reliant on automated tools, neglecting more comprehensive research, consultation, and deliberation. In the long term, it might amplify social issues, lock in certain values and knowledge, or lead to unpredictable risks and adverse outcomes. Before applying it to real-world governance processes, it is crucial to extensively consider its potential social risks.

The data in *EuroCon* has undergone processing using LLMs, including filtering, summarizing, and translating, as well as expanded settings for specific tasks, such as adjusting the distribution of seats among different parties and adding additional voting rules. During the LLM data processing, although the content is directly related to the original text, inherent biases and harmful statements may still be introduced from the LLMs. Additionally, we do not rule out the possibility of omissions during data collection. These factors mean that our benchmark does not necessarily have a direct correlation with real-world European Parliament decisions and cannot be used to represent or predict any political outcomes or statements of the European Parliament.

It is worth noting that *EuroCon* should be only used for scientific research and academic purposes. If any third party uses *EuroCon* to make inappropriate statements, actions, or harmful legal suggestions regarding political, ethical, or other issues, this paper is not responsible for such actions. Additionally, since the data sources of *EuroCon* are real parliamentary data, they may contain politically sensitive statements from certain countries and regions, which do not represent any political views of the authors of this article.

[31] https://www.europarl.europa.eu

[32] https://howtheyvote.eu