Identifying Algorithmic and Domain-Specific Bias in Parliamentary Debate Summarisation

Abstract. The automated summarisation of parliamentary debates using large language models (LLMs) offers a promising way to make complex legislative discourse more accessible to the public. However, such summaries must not only be accurate and concise but also equitably represent the views and contributions of all speakers. This paper explores the use of LLMs to summarise plenary debates from the European Parliament and investigates the algorithmic and representational biases that emerge in this context. We propose a structured, multi-stage summarisation framework that improves textual coherence and content fidelity, while enabling the systematic analysis of how speaker attributes – such as speaking order or political affiliation – influence the visibility and accuracy of their contributions in the final summaries. Through our experiments using both proprietary and open-weight LLMs, we find evidence of consistent positional and partisan biases, with certain speakers systematically under-represented or misattributed. Our analysis shows that these biases vary by model and summarisation strategy, with hierarchical approaches offering the greatest potential to reduce disparity. These findings underscore the need for domain-sensitive evaluation metrics and ethical oversight in the deployment of LLMs for democratic applications.

1 Introduction

Public understanding of parliamentary activities is worryingly low, despite their fundamental role in representative democracy. Research consistently demonstrates that citizens struggle to comprehend parliamentary activities, and often see them as talking shops, where politicians grandstand, but do little else [4,8]. This points to a democratic deficit where public engagement with parliamentary processes is severely limited. Citizens' actions alone do not fully explain this situation [12,13]. Information on parliamentary activities is generally made available to the public through plenary transcripts, legislative records, and detailed minutes of parliamentary meetings. However, while this data is theoretically open, it is often not presented in a user-friendly or easily accessible format in practice. As a result, citizens rarely take the time to engage with it.

The use of large language models (LLMs) for automated summarisation of parliamentary debates offers a promising way to address the accessibility gap and enhance citizens' engagement with democratic institutions [3]. However, for such approaches to be effective, it is essential that the generated summaries accurately reflect both the content and the speakers involved. This paper investigates these challenges by evaluating the accuracy of LLMs in summarising political debates, using a sample of plenary speeches from the 9th European Parliament.

While LLMs can provide an effective means of summarising text [28], there are important methodological considerations that must be made when working with parliamentary debates. In particular, alongside challenges such as summary fidelity [5] and knowledge hallucination [20], LLMs have been shown to exhibit algorithmic biases, where language models demonstrate an 'uneven' utilisation of information in their input context [10,19], and social biases, where language models demonstrate disparate treatment or outcomes between social groups [1,7,15]. Meanwhile, the tendency of LLMs to omit details or overgeneralise in summaries has raised concerns in other domains [18].

Numerous general-purpose measures exist for assessing summarisation quality, faithfulness, and conciseness, many of which were developed before the advent of modern LLMs [5]. However, summarising political debates poses distinct challenges. It requires aggregating content from potentially long sequences of relatively brief documents or interventions, requiring a more balanced consideration of all source inputs than is typical in standard multi-document summarisation or question-answering tasks. Furthermore, there are important domain-specific considerations. Namely, the debate summary must recognise and attend to the key substantive components of each debate intervention (e.g. issue, position, argument, proposal). These components must be communicated to the reader faithfully, and with reference to the source information (i.e., the specific intervention). Crucially, summaries must also make clear not only *what* was said, but *who* said it. While many automated methods for summary evaluation exist, they do not sufficiently penalise the misattribution of arguments to speakers.

In this work, we present a new framework for generating and evaluating political debate summaries with these key requirements in mind. Furthermore, we examine algorithmic and domain-specific biases exhibited by LLMs when attending to different positions within parliamentary debates, using plenary speeches from the European Parliament (EP) as our source data. Through these investigations, we highlight methodologies for producing high-quality summaries that enhance public understanding and promote transparency in parliamentary proceedings. Our approach and findings also offer broader insights for the design and evaluation of complex multi-document summarisation systems.

2 Background

2.1 Generating and Evaluating Summaries

Text summarisation is the process of distilling the most important information from a source text to produce an abridged version for a particular user or task [28,14] While early approaches to automatic summarisation were predominantly *extractive*, relying on directly copying segments from the source text [2], many use cases benefit from *abstractive* summaries that rephrase the original content. Notably, recent advances in LLMs have enabled the generation of more natural and coherent summaries that better resemble human writing, making them well-suited for such tasks [28].

Summary evaluation methods have been developed to automatically assess how effectively a summary captures the important content of a longer text while remaining concise, coherent, and faithful to the source material. Early work in this area has focused on explicit overlaps between the summary and the source material. *ROUGE* [9] measures n-gram *recall* – the proportion of n-grams in the summary that were present in the source material. Similarly, *BLEU* [17] measures the *precision* of the n-grams in the summary compared to the source. As these methods rely on n-grams comparisons, they require that the summary text contains exact matches with the source text, and as a result, they penalise paraphrased or *abstractive* summaries. Moreover, such measures have previously been shown to correlate poorly with human judgments of summaries [16].

To evaluate abstractive summaries, embedding-based similarity measures have been developed to address the semantic limitations of the earlier methods that relied on n-gram overlap. For example, *BERTScore* [29] measures *precision* and *recall* separately, by comparing the semantic similarity of tokens in the summary and source material. Specifically, a high *BERTScore* precision (P_{BERT}) indicates that each token/term in the summary is a strong semantic match for a token in the source (indicating a low level of added information or hallucination). Conversely, a high recall (R_{BERT}) indicates that tokens/terms in the source are all matched by tokens in the summary – indicating that the summary covers the content in the original document.

Traditional n-gram-based evaluation methods treat all n-grams equally and often fail to detect subtle semantic errors in summaries. To attempt to address this limitation, more recent work has explored *factual consistency* as an evaluation criterion, which considers whether a summary accurately conveys the meaning of the source text without introducing contradictions [24]. Such approaches have been applied in the context of question answering (QA) [21] and natural language inference (NLI) [6]. However, their effectiveness relies heavily on the quality of the underlying QA and NLI models.

The brevity or succinctness of summaries is typically measured independently. For example, the compression ratio measures the ratio of tokens in the summary to tokens in the source document [5]. Alternatively, some studies will assess *information density* by normalising or penalising summary length [11].

When summarising parliamentary debates, particular attention must be given to accurately attributing arguments, positions, and proposals to the correct speakers. Existing evaluation methods do not adequately penalise the misattribution of content to speakers. Later in Section 3.2, we describe our framework for evaluating debate summaries where we address this evaluation gap.

2.2 Types of Bias in LLMs

There is existing research that finds that LLMs do not attend equally to all regions of the input context when generating outputs. For example, the "*lost-in-the-middle*" problem refers to the tendency of LLMs to favour content at the beginning or the end of the context in large context question answering (QA). In particular, Liu et al. [10] showed that LLMs consistently achieve lower QA

scores when the information relevant to the question is found in the middle of the context. Motivated by these findings, Ravaut et al. [19] performed multi-document summarisation experiments and found that LLM-based summaries contain significantly more information from the documents at the beginning of the context than from those appearing later in the context.

As well as algorithmic bias, it is also important to consider the impact of social biases. While bias is a subjective term, it can broadly be defined as disparate treatment or outcomes between social groups. In our later evaluation of bias in parliamentary debate summarisation, we adopt the definitions of *individual* and *group fairness* proposed by Gallegos et al. [7]. These definitions require similar outcomes – i.e., that summaries represent all speakers with equal accuracy and clarity – regardless of party group membership or algorithmic factors such as position within the model's input context.

3 Methods

3.1 Generating Debate Summaries

As previously noted, summarising parliamentary debates presents distinct challenges. Unlike many traditional summarisation tasks, it requires not only an accurate representation of the arguments made but also precise attribution to the correct speakers. Moreover, to promote trust, fairness, and transparency, the content of the summary must be clearly traceable to the original source material. In light of these considerations, we adopt the following framework for summarising parliamentary debates. Formally, a debate is represented by an ordered sequence of n interventions made by m different speakers

$$I = \{i_1^1, i_2^2, i_3^3 \ldots, i_n^m\}$$

where i_k^j is a contribution made by speaker j, that was the k^{th} contribution or intervention in the debate.

To support accurate attribution of content to its original source and to ensure focus on the substantive aspects of each debate intervention, we use a summarisation workflow involving two steps (see Figure 1). Firstly, we introduce an intermediary step: the creation of a set of *intervention summaries*, which is subsequently aggregated into a final *debate summary*. The motivation for adopting a *hierarchical* summarisation approach is analogous to the architecture of a convolutional neural network, in which early layers extract lower level features that are subsequently aggregated and contrasted in later layers to produce higher level representations [30]. This approach also mirrors the idea of prescribing a *chain-of-thought* (CoT) prompting for the model's summarisation process. Models fine-tuned to generate such structured reasoning have been shown to exhibit improved reasoning capabilities [26,25]. Introducing an explicit CoT through a separate intervention summary step not only guides the model's generation but also allows us to retain intermediate outputs for evaluating speaker attribution accuracy (see Section 3.2).

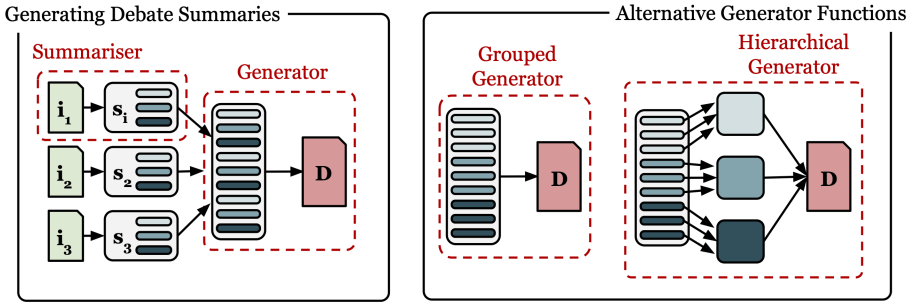

Fig. 1. Debates are summarised using a two step process. The *summariser* function generates intervention summaries from each intervention in the debate. The *generator function* aggregates the information from each intervention summary to generate a debate summary.

Intervention summaries. Each intervention is summarised by f_{SUM}, which maps an intervention to a structured summary:

$$s_i^j = f_{\text{SUM}}(i_i^j)$$

Taking inspiration from studies of policy bargaining in political institutions [23,22], we adopt the following structure for our summaries:

```
Headline: A concise, single-line summary of the speech.
Issue: A brief overview of any key issues raised by the speaker.
Position: Any stance or viewpoint expressed in the speech.
Argument: Any arguments used to justify the positions taken.
Proposal: Any proposals or policy actions mentioned by the speaker to
address the issues raised.
Quotes: 2-3 representative quotes that capture the speaker's stance.
```

This structure is intended to focus the final summary on the substantive aspects of the debate interventions, specifically the issues raised, positions expressed, arguments presented, and proposals put forward by the speakers. It also serves to guide the *evaluation* of summaries by focusing on these key components (see Section 3.2). After the first intervention summarisation step, a debate is represented as a set of structured summaries:

$$S = \{s_1^1, s_2^2, s_3^3, \ldots, s_n^m\}$$

In our experiments, we implement the summariser functions using the LLMs listed in Section 4.2.

Debate summaries. The final *debate summary* D is produced by applying a generator function f_{GEN} to the structured set of intervention summaries. This function maps the set of individual summaries to a coherent executive summary

of the overall debate:

$$D = f_{\text{GEN}}(S)$$

In our experiments, we implement a range of generator functions using LLMs to produce debate summaries from a structured set of intervention summaries. We evaluate four distinct approaches, illustrated in Figure 1:

1. **Default generator:** Stacks all intervention summaries and provides the combined text as input to the LLM.
2. **Grouped generator:** Reorganises the input by grouping content under each subheading (e.g., position, issue, argument) across speakers before inputting to the LLM.
3. **Hierarchical generator:** Introduces an additional summarisation layer in which all positions, issues, arguments, and proposals are summarised independently and then aggregated into a final summary. Hierarchical pipelines have been shown to be effective in summarising complex information, provided there is a clear organisational structure in the input [19,27,30].
4. **Prompted generator:** Makes use of an explicit prompt instructing the model to "*give equal attention to all speakers in your summary*". Prompt engineering of this nature has been shown to mitigate attention biases in certain LLMs [19].

3.2 Evaluating Debate Summaries

The goal of summarisation is to generate a concise, coherent version of a longer text that preserves key information and enables quick understanding of the main points. When summarising parliamentary debates, we place an additional emphasis on the substantive aspects of debate and policy bargaining (i.e., issue, position, argument, proposals), and the reliable attribution of each of these concepts to speakers. That is to say, when reading a debate summary, it should be clear *what* was said, and *who* said it. Many measures have been developed for the automatic evaluation of natural language summaries (see Section 2). Broadly speaking, when evaluating debate summaries these methods assess *what* was said but they do not sufficiently penalise summaries for misattributing statements to different speakers. In this section, we propose to address this evaluation gap by measuring the accuracy of the summary separately for each speaker. For the purposes of this work, we focus our evaluations on *fidelity* and *compression*.

Fidelity and attribution accuracy. To focus our evaluations on the correct attribution of issues, positions, arguments, and proposals to speakers, we must identify the regions of the debate summary that are relevant to each speaker. From each debate report D, we reconstruct the contributions of each speaker using a reconstructor function f_{REC}, where

$$\hat{s}_i^j = f_{\text{REC}}(D, \text{speaker}_j)$$

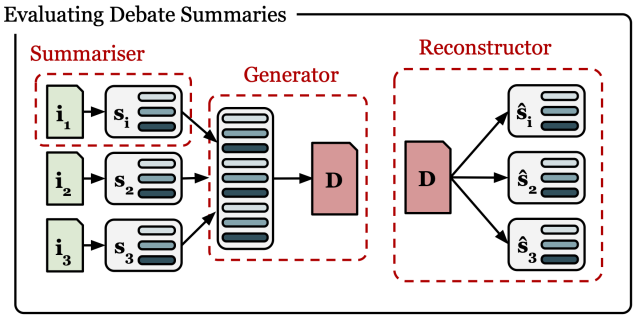

Fig. 2. Proposed two step summarisation workflow. The reconstructor function recreates a structured version of a speaker's intervention using the final debate summary as input. If the speaker's contribution can be accurately reconstructed, we conclude the summary attends to that speaker and their contributions accurately.

represents a summary of the i^{th} intervention in the debate (made by speaker j), that was *extracted* (or *reconstructed*) from information in the debate report D. Thus, we can evaluate the report (and the associated aggregator function) by comparing the reconstructed intervention summaries $\hat{S} = \{\hat{s}_1^1, \hat{s}_2^2, \ldots, \hat{s}_n^m\}$, with the original summaries S. Specifically, if a reconstructed summary $\hat{s}_i^j$ remains faithful to the initial summary s_i^j, we conclude that the summarisation has attended to that speaker and their contribution accurately and, critically, that the contribution was clearly attributed to that speaker. Further, we can assess methodological and domain-specific biases in the generated report by comparing the fidelity of the interventions given different factors such as their order in the context (i.e., when they spoke during the debate) or the political affiliations of the speaker (e.g.: party membership).

We use *BERTScore* [29] to evaluate the accuracy (fidelity) of the reconstructed intervention summaries compared to the original summaries. Both the intervention summaries S and the reconstructed summaries $\hat{S}$ share the structure described in Section 3.1. As such, $BERTScore(s_i^j, \hat{s_i}^j)$ measures how faithfully the substantive aspects of the intervention i_i^j (i.e., issue, position, argument, proposals) were communicated in the debate summary and correctly attributed to the speaker. By measuring the average intervention fidelity ($\frac{1}{n}\sum_i^n BERTScore(s_i, \hat{s_i})$) for a debate, as opposed to entire debate summary fidelity ($BERTScore(S, D)$), we build attribution accuracy into our evaluation of fidelity.

Compression. The most accurate way to convey each speaker's contribution is via a full transcript, but this contradicts the goal of summarisation, which is to provide a concise account of the source. It is therefore important to consider report length and information density alongside fidelity. We define *Compression Ratio* as the number of tokens in the final debate summary divided by the total number of tokens in the source, which includes all intermediate intervention summaries. Lower values indicate greater compression. Given our proposed

summarisation framework, we can also define a complementary metric: the *Decompression Ratio*. This is the ratio of the number of tokens in the reconstructed interventions to the number of tokens in the debate summary. A higher value for this metric indicates that more information can be effectively recovered or *extracted* from the debate summary.

4 Experimental Setup

4.1 Data

For our experiments we use a dataset of translated debates from the 9th European Parliament [1], from the period 2019 to 2025. We randomly sample 93 full debates, each containing up to 70 speeches (interventions). Although few debates in the sample exceed this length, we impose this limit to ensure all debates fit within the context window of the language models under evaluation. The final dataset comprises 2,976 interventions delivered by 719 speakers from 9 different party groups which generally consist of politicians with aligned ideologies.

4.2 Models

We use LLMs to implement the summariser, generator, and reconstructor functions in our experiments. As the evaluations are the most costly aspect of our experiments, we choose an open-weight model for reconstructing speaker interventions from debate summaries (f_{REC}). For clarity, and to avoid potential biases, we exclude the model used for evaluations from generation. Here we outline the different models used for generating and evaluating reports. For simplicity, we use the same model for generating both intervention summaries (f_{SUM}) and debate summaries (f_{GEN}). To create the latter, we consider the four generation methods outline in Section 3.1: *default*, *grouped*, *hierarchical*, *prompted*.

Generating reports. For this process, we consider a selection of openly-available and proprietary models:

- **Mistral** is a lightweight open model from *Mistral AI*. In our experiments, we use `Mistral:7b` from *Ollama*[2].
- **Claude Sonnet** is a proprietary model from *Anthropic*. In our experiments, we use `claude-3-7-sonnet-20250219`.
- **Phi4** is a compact 14 billion parameter, open model from *Microsoft*. In our experiments, we use `Phi-4:14b` from *Ollama*.
- **GPT-4.1** is a state-of-the-art proprietary model from *OpenAI*. In our experiments, we use `gpt-4.1` via *OpenAI* API.

[1] https://www.europarl.europa.eu/plenary/en/home.html
[2] https://ollama.com

Evaluating reports. To assess summaries, we make use of the open-weight model **Qwen3** model from *Alibaba*. Specifically, we use `qwen3:30b-a3b` provided by *Ollama*, which employs mixture-of-experts to offer high performance with significantly fewer active parameters.

4.3 Validation

The reconstructor function (f_{REC}) is a critical stage in our evaluation pipeline. We employ two measures to validate this aspect of the experiments. Firstly, we assess the effect of the choice of f_{REC} model on our results. In other words, if we change the model we use to reconstruct/extract speaker interventions from debate summaries, would our results change significantly? To test this, we take a sample of 7 debate summaries generated using different generation methods. For each of the 147 different debate interventions, we reconstruct intervention summaries ($\hat{s}_i = f_{\text{REC}}(D, \text{speaker}_i)$) using four different models (Mistral, Phi-4, Qwen3, and Claude-Sonnet-3-7). We then calculate the intervention fidelity ($\mathit{BERTScore}(s_i, \hat{s}_i)$) for all interventions, and measure the pairwise Pearson correlation between scores from all models. All model pairs showed high correlation in their scores, with the lowest at $r = 0.74$, indicating strong agreement across models. This suggests that the observed trends are largely robust to the choice of model used for reconstructing intervention summaries from debate summaries.

Next, we assess the *precision* of the reconstructor model. While we can *evaluate* debate summaries by comparing the extracted summary to original structured summary, we can validate the reconstructor function by considering the *BERTScore* precision between the reconstructed intervention summary and the debate summary ($P_{BERT}(D, \hat{s}_i)$). Across all reconstructed interventions in our experiments, the average BERTScore precision using Qwen3 was 0.80. This indicates that low intervention fidelity scores, ($\mathit{BERTScore}(s_i, \hat{s}_i)$), are primarily due to errors introduced during debate summarisation rather than during reconstruction.

5 Results

5.1 Fidelity and Compression

Table 1 reports the average intervention fidelity ($BERTScore$), brevity (Compression Ratio), and information density (Decompression Ratio) for all models and generation methods. Overall, the larger, proprietary models (Claude-Sonnet, GPT-4.1) communicate the core aspects of debate interventions more faithfully than the smaller, open-weight models. Prompting models to pay attention to all speakers improves the mean accuracy of the larger models, yet has no impact on the smaller models. Across most models, hierarchical summaries are the shortest and most information-dense. In all cases except Claude-Sonnet, they also demonstrate higher faithfulness than the default generator approach.

Model	Generator	$F1_{\text{BERT}}$	C_{Ratio}	D_{Ratio}
Claude-Sonnet-3-7	Default	0.239	0.24	1.01
	Grouped	0.225	0.26	0.95
	Hierarchical	0.230	0.20	1.22
	Prompt	0.258	0.26	1.11
GPT-4.1	Default	0.250	0.33	0.96
	Grouped	0.267	0.33	0.91
	Hierarchical	0.256	0.29	0.92
	Prompt	0.313	0.36	0.99
Mistral	Default	0.072	0.11	0.79
	Grouped	0.122	0.12	1.30
	Hierarchical	0.142	0.15	1.16
	Prompt	0.068	0.11	0.79
Phi-4	Default	0.096	0.16	0.58
	Grouped	0.167	0.17	1.03
	Hierarchical	0.132	0.16	0.89
	Prompt	0.098	0.18	0.58

Table 1. Average *BERTScore*, Compression, and Decompression ratios for all interventions across all debates.

5.2 Speaker Order Bias

There is existing research that finds that LLMs do not attend equally to all regions of the context when generating outputs (see Section 2.2) In light of this, in Figure 3 we assess the extent to which the temporal position of a speaker in the debate (i.e., when they speak, and thus, where they occur in the LLM context) affects how much and how accurately the debate summary attends to that speaker and their contributions. We distinguish between a speaker's *temporal position* in the debate — the order in which they speak — and their ideological position, referring to the stance they express on some issue. To avoid confusion, we will refer to the former as speaking order.

From Figure 3, it is clear that all models demonstrate some speaker order bias, with Claude-Sonnet-3-7 and GPT-4.1 attending more accurately to speakers who spoke earlier in the debate while Mistral and Phi-4 favour those at the end. In the case of Mistral and Phi-4, we note that many of the speakers from the middle of the debate are ignored entirely ($F1_{\text{BERT}} \approx 0$).

In Figure 4 we plot the speaker order bias for all generator methods, for each of our models. It is apparent from the figure that different generator methods have an impact on the speaker order bias. To evaluate these effects we estimate a beta regression model. The dependent variable is intervention fidelity (*BERTScore F1*), and predictors include relative speaker order x_i (representing when the intervention occured in the debate), its squared term x_i^2 (to capture non-linear patterns), indicators for each generator method (i.e., Method_m is a dummy variable indicating if the debate was summarised using method m), and interactions to measure the effect of generator method on the speaker order bias.

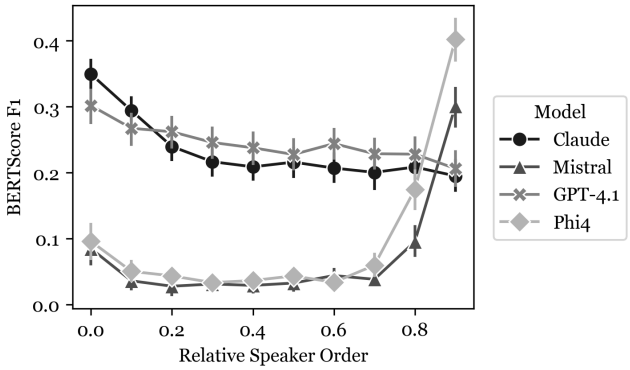

Fig. 3. Speaker Order Bias– default methods. *BERTScore* measures the similarity between a speaker's reconstructed intervention summary $\hat{s}$ (based on the information in the debate summary), and their original intervention summary s. Relative Speaker Order ($\frac{k}{n}$) represents the temporal position of their intervention in the debate (k) adjusted for the number of interventions in the debate (n) with $\frac{k}{n} \approx 0$ representing the earliest speakers in the debate and $\frac{k}{n} = 1$ representing the last intervention.

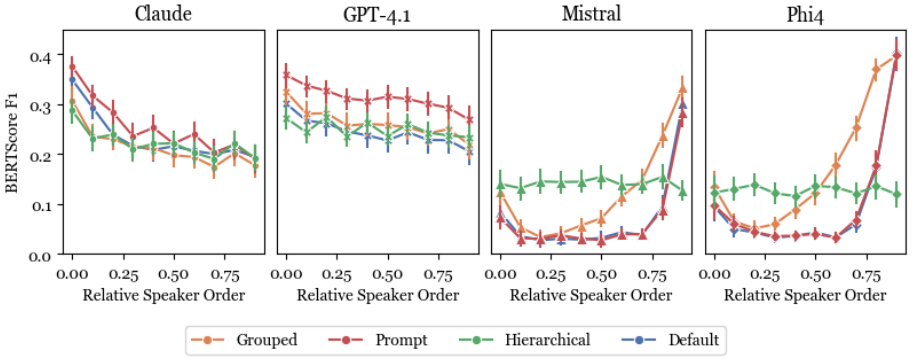

Fig. 4. Speaker Order Bias – all models and methods. *BERTScore* measures the similarity between a speaker's reconstructed intervention summary $\hat{s}$ (based on the information in the debate summary), and their original intervention summary s. Relative Speaker Order represents the temporal position of their intervention in the debate, with lower scores representing earlier contributions.

The regression is described by

$$\begin{aligned}\text{logit}(\mu_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \\ + \sum_{m \in \text{Methods}} \left[\gamma_m + \delta_m x_i + \theta_m x_i^2\right] \cdot \text{Method}_m\end{aligned}$$

where β_0, β_1 and β_2 represent coefficients for the intercept, linear and quadratic terms, while γ_m represents the method effect for method m, with β_m and θ_m describing the effects of method m on the speaker order bias. Table 2 reports

	Coef	Claude-3-7	GPT-4.1	Mistral	Phi-4
Intercept	β_0	-0.700**	-1.120**	-1.978**	-1.763**
Linear speaker order	β_1	-1.934**	-0.721*	-2.619**	-3.521**
Quadratic speaker order	β_2	1.407**	0.362	3.346**	4.523**
Hierarchical	γ_H	-0.361**	0.094	0.090	-0.220**
Grouped	γ_G	-0.334**	0.113	0.051	-0.133
Prompted	γ_P	0.171*	0.264**	-0.066	0.115
Order × Hierarchical	δ_H	1.005*	0.872*	2.950**	3.693**
Order × Grouped	δ_G	0.666	0.246	0.267	1.306**
Order × Prompted	δ_P	-0.523	0.889*	0.450	-0.262
Order^2 × Hierarchical	θ_H	-0.841*	-0.619	-3.630**	-4.694**
Order^2 × Grouped	θ_G	-0.522	-0.165	0.266	-0.479
Order^2 × Prompted	θ_P	0.236	-0.931*	-0.518	0.178
Precision		1.153**	1.009**	1.426**	1.352**

Table 2. Beta regression coefficients for measuring speaker order bias and the mitigating effects of different generation methods.

regression coefficients for all LLMs. All models show a strong positional/order bias using the default generator method, with a negative coefficient on speaker position β_1. All models but GPT-4.1 also show a significant positive quadratic term β_2 indicating that summaries favour early and late speakers, disadvantaging those in the middle.

Our proposed *hierarchical* summarisation method significantly alters the bias profile across all LLMs – δ_H consistently shows a positive interaction with speaker order, reducing the bias towards earlier speakers, and in all cases where models show a bias towards later speakers (i.e., β_2 is significant and positive), we find a significant negative θ_H term. This suggests hierarchical summarisation reduces speaker order bias. With γ_H negative and significant in the larger models (Claude and GPT), we note that the reconstruction fidelity for speakers at the beginning of the debate is lower than with the default method.

The *grouped* method shows similar but weaker effects, without significant coefficients, indicating potential but uncertain bias reduction. The *prompted* method, by contrast, does not significantly reduce bias and may trend toward slightly worse order effects in the Claude-based summaries; the positive method effect γ_P indicates a more thorough account of the earlier speakers than the other approaches. However, without a significant improvement in δ_P, this leads to increased order bias, as the model still fails to adequately attend to speakers appearing later in the debate.

Overall, the larger models (Claude and GPT) behave quite differently to the smaller models (Mistal and Phi-4). For instance, prompting the model to attend equally to all speakers has no effect on the smaller models, while in the case of the larger models it leads to improved speaker fidelity for early speakers (and increases the bias in the Claude-based summaries). While *hierarchical* summarisation reduces the speaker order bias across all models, it has the greatest effect on the smaller models as it flattens the bias *and* increases overall fidelity.

5.3 Party Group Bias

In addition to speaker order bias, LLM-based summaries may also differ in how they represent the issues, positions, arguments, and proposals put forward by speakers from different party groups. Given that all summaries demonstrate some order bias, we control for where interventions occur in the context when comparing intervention fidelity (*BERTScores*). Similar to the speaker order bias model, we fit a beta regression to model fidelity ($BERTScore(s_i, \hat{s}_i)$) as a function of the speaker's turn x_i (when they speak in the debate), the square of this x_i^2 (to account for the u-shaped attention), and indicator variables for each of the party groups in the European Parliament (e.g.: $\text{Party}_{A,i}$ is a binary variable indicating if speaker_i is a member of group A).

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 \text{Party}_{A,i} + \beta_4 \text{Party}_{B,i} + \dots$$

We centre the speaker order variable such that $x_i = 0$ refers to a speaker in the middle of a debate, and select the largest party (the European People's Party in the case of the EP) as the reference group.

Figure 5 plots the predicted marginal means for each EP party group across the different debate summarisation methods. The predicted marginal fidelity scores indicate that the choice of summarisation method can impact how well the summary reflects the positions and arguments of speakers from different political groups. For the Mistral and Phi-4 models, the *default* and *prompted* generators performed too poorly to identify any significant bias between different party groups. However, in the case of the generators that improved overall fidelity (*hierarchical* and *grouped*), we find that the Phi-4 summaries attend more faithfully to interventions made by The Left party than others. A similar pattern is apparent in Claude's *default* and *hierarchical* summaries. Notably, the Claude-based generator, which produced the most abstractive summaries (see Section 5.1), shows the greatest disparity in group means. Interventions from The Left party are represented more faithfully than those from other groups, particularly the Identity and Democracy Group.

6 Conclusions and Future Work

There is a clear benefit to summarising parliamentary data to improve accessibility, enhance transparency, and strengthen the connection between the public and the democratic institutions that represent them. However, summarising complex sources such as parliamentary debates presents distinct challenges. LLMs offer a potentially promising approach for this task, given their demonstrated ability to generate well-structured and coherent summaries across a range of domains. However, these models also raise concerns, as they can exhibit both algorithmic [10,19] and social biases [1], which must be evaluated in parallel with more conventional metrics of summary quality and accuracy. While many established methods exist for evaluating automatically-generated summaries, they are often

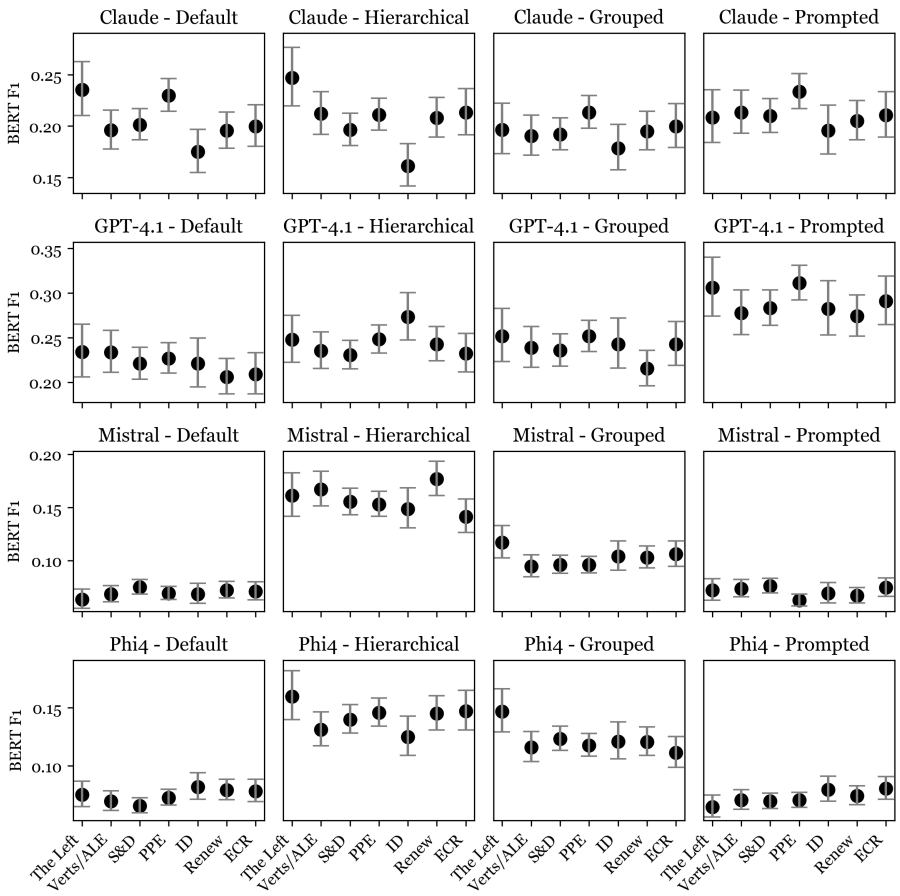

Fig. 5. Marginal mean intervention fidelity by group. Plots show the mean predicted *BERTScore* for an intervention made in the middle of a debate by speakers of different EP party groups. Error bars indicate 95% confidence interval. Higher scores indicate that the issues, positions, arguments, and proposals, made by members of that political group were more accurately/clearly communicated in the debate summary.

insufficient for assessing attribution accuracy in political debate contexts. In particular, it is crucial not only that a summary faithfully reflects the arguments, positions, and proposals expressed in the debate, but also that these elements are correctly attributed to the speakers who presented them.

To address this gap, in this work we introduced a structured framework for generating and evaluating political debate summaries. Our framework focuses on the substantive content of each intervention, namely issues, positions, arguments, and proposals, and evaluates how accurately these are attributed and communicated in the final summaries. By applying this framework in the context of plenary speeches from the European Parliament, we identified several forms

of bias present in LLM generated summaries. Hierarchical summarisation methods, which structure and aggregate content at multiple levels, proved especially effective in producing concise and faithful summaries that attend more equally to all speakers, regardless of their position in the debate. However, we also observed domain specific biases, with some models more accurately representing certain political groups than others. While the effect appears to be particularly pronounced in more abstractive summarisation methods, the exact mechanism of this bias is an important area for future work.

Going forward, our goal is to integrate these findings into a new AI-driven platform designed to transform complex parliamentary records, such as plenary speeches, into accessible, comprehensible summaries. Our aim is to support public engagement with parliamentary proceedings and help bridge the gap between citizens and their elected representatives.

References

1. M. Bartl and S. Leavy. From 'showgirls' to 'performers': Fine-tuning with gender-inclusive language for bias reduction in LLMs. In *Proc. 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 280–294, 2024.
2. J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, 1998.
3. A. Combaz, D. Mas, N. Sanders, and M. Victor. Applications of artificial intelligence tools to enhance legislative engagement: Case studies from make.org and maple. *arXiv pre-print*, (2503.04769), 2025.
4. R. J. Dalton. *Democratic challenges, democratic choices: The erosion of political support in advanced industrial democracies*. Oxford University Press, 2004.
5. A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.
6. T. Falke, L. F. Ribeiro, P. A. Utama, I. Dagan, and I. Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, 2019.
7. I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179, Sept. 2024.
8. C. Hay. *Why we hate politics*, volume 5. Polity, 2007.
9. C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.
10. N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.
11. Y. Liu, Q. Jia, and K. Zhu. Reference-free Summarization Evaluation via Semantic Correlation and Compression Ratio. In *Proc. 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2109–2115, 2022.

12. A. Lupia. *Uninformed: Why people know so little about politics and what we can do about it*. Oxford University Press, 2016.
13. P. Mair. *Ruling the void: The hollowing of Western democracy*. Verso books, 2013.
14. I. Mani and M. T. Maybury. *Advances in automatic text summarization*. MIT press, 1999.
15. K. Mei, S. Fereidooni, and A. Caliskan. Bias Against 93 Stigmatized Groups in Masked Language Models and Downstream Sentiment Classification Tasks. In *Proc. 2023 ACM Conference on Fairness Accountability and Transparency*, pages 1699–1710, 2023.
16. A. Nenkova. Summarization evaluation for text and speech: issues and approaches. In *Interspeech'06*, 2006.
17. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
18. U. Peters and B. Chin-Yee. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4):241776, 2025.
19. M. Ravaut, A. Sun, N. Chen, and S. Joty. On Context Utilization in Summarization with Large Language Models. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781, 2024.
20. S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, Y.-L. Boureau, and J. Weston. Recipes for building an open-domain chatbot. *arXiv pre-print*, (2004.13637), 2020.
21. T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, J. Staiano, A. Wang, and P. Gallinari. QuestEval: Summarization Asks for Fact-based Evaluation. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604. Association for Computational Linguistics, 2021.
22. R. Thomson, J. Arregui, D. Leuffen, R. Costello, J. Cross, R. Hertz, and T. Jensen. A new dataset on decision-making in the European Union before and after the 2004 and 2007 enlargements. *Journal of European Public Policy*, 19(4):604–622, 2012.
23. R. Thomson and F. N. Stokman. Research design: Measuring actors' positions, saliences and capabilities. *The European Union Decides*, 2006.
24. A. Wang, K. Cho, and M. Lewis. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv pre-print*, (2004.04228), 2020.
25. Y. Wang, Z. Zhang, and R. Wang. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv pre-print*, (2305.13412), 2023.
26. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv pre-print*, (2201.11903), 2023.
27. G. Zhang, M. A. N. Ahmed, Z. Hu, and A. Bulling. SummAct: Uncovering User Intentions Through Interactive Behaviour Summarisation. In *Proc. 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2025.
28. H. Zhang, P. S. Yu, and J. Zhang. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 2025.
29. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating Text Generation with BERT. *arXiv pre-print*, (1904.09675), 2020.
30. Z. Zheng, W. Chao, Z. Qiu, H. Zhu, and H. Xiong. Harnessing large language models for text-rich sequential recommendation. In *Proc. ACM Web Conference 2024*, pages 3207–3216, 2024.