What They Say vs How They Say It: Rhetorical Polarization in the European Parliament (2015–2025) from Multilingual Speech Analysis

Paul Gerrit Koslowsky

Dissertation written under the supervision of professor Nicolò Bertani

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa, 02.01.2025.

What They Say vs How They Say It: Rhetorical Polarization in the European Parliament (2015–2025) from Multilingual Speech Analysis

Paul Gerrit Koslowsky

Supervisor: Professor Nicolò Bertani

Abstract

This dissertation examines how rhetorical polarization between political groups in the European Parliament evolved between 2015 and mid-2025. Using a newly constructed corpus of around 160,000 plenary speeches, it combines multilingual sentence embeddings with Mehlhaff's cluster polarization coefficient (CPC) to measure between-party separation in semantic space, and applies the same framework to large-language-model-based scores of conflict, outgroup tone, and extremity. This yields two complementary views of polarization: what parties talk about and how hostile their rhetoric sounds. The results show that the Parliament is not semantically hyper-polarized. Party labels explain only a small share of variation in speech embeddings, and increases in semantic CPC are concentrated in specific policy domains and crisis moments rather than in a uniform chamber-wide trend. By contrast, CPC in the hostile-rhetoric space is substantially larger: party identity strongly structures the deployment of conflictual and alarmist language, with a stable pro-EU core and more distinctive radical left and radical right flanks. Topic-specific analysis reveals that peaks in both content and tone are closely associated with the refugee crisis, COVID-19, and Russia's invasion of Ukraine. These spikes remain topic-specific and do not erase a shared semantic ground. Methodologically, the study demonstrates how speech-based CPC and LLM-derived tone measures can be combined to track elite polarization in a multilingual, multiparty legislature.

Keywords: European Parliament; rhetorical polarization; hostile rhetoric; cluster polarization coefficient (CPC); text-as-data; large language models; plenary speeches; crisis politics.

What They Say vs How They Say It: Rhetorical Polarization in the European Parliament (2015–2025) from Multilingual Speech Analysis

Paul Gerrit Koslowsky

Supervisor: Professor Nicolò Bertani

Abstract

Esta dissertação analisa como a polarização retórica entre grupos políticos no Parlamento Europeu evoluiu entre 2015 e meados de 2025. Com um corpus recém-construído de cerca de 160.000 discursos em plenário, combina embeddings de frases multilingues com o coeficiente de polarização por clusters (CPC) de Mehlhaff para medir a separação entre partidos no espaço semântico, e aplica o mesmo enquadramento a pontuações, geradas por modelos de linguagem de grande escala (LLMs), de conflito, tom face ao exogrupo e radicalidade. Daqui resultam duas perspetivas: sobre o que os partidos falam e quão hostil soa a sua retórica. Os resultados indicam que o Parlamento não é semanticamente hiperpolarizado. As etiquetas partidárias explicam apenas uma pequena fração da variação nos embeddings, e os aumentos do CPC semântico concentram-se em áreas políticas específicas e em momentos de crise. Em contraste, o CPC no espaço da retórica hostil é muito maior: a identidade partidária estrutura fortemente o uso de linguagem conflituosa e alarmista, com um núcleo pró-UE estável e flancos mais distintivos da esquerda radical e da direita radical. A análise por tópicos mostra que os picos de conteúdo e de tom estão associados à crise dos refugiados, à COVID-19 e à invasão russa da Ucrânia. Estes picos permanecem específicos de certos temas e não eliminam um terreno semântico partilhado. Metodologicamente, o estudo demonstra como combinar CPC baseado em discursos com medidas de tom derivadas de LLMs para acompanhar a polarização das elites numa legislatura multilingue e multipartidária.

Keywords: European Parliament; rhetorical polarization; hostile rhetoric; cluster polarization coefficient (CPC); text-as-data; large language models; plenary speeches; crisis politics.

Contents

1 Introduction

Political polarization has become a central concern for contemporary democracies, not only because parties diverge on policy, but because citizens increasingly view politics as a struggle between hostile social camps. In the United States, partisan attachments have fused with racial, religious, and cultural identities into partisan “mega identities,” so that party affiliation summarizes where individuals stand in a broader social conflict (Mason 2018). Experimental work shows that many partisans now treat politics as competition between rival “teams” and express open animus toward supporters of the opposing party, on a scale comparable to or exceeding prejudice based on race (Iyengar and Westwood 2015). These developments have motivated a distinction between ideological polarization, understood as distance in policy positions, and affective polarization, understood as hostility and social distance between camps. This dissertation takes that broader concern as a starting point, but shifts attention from the American mass public to elite rhetoric in the European Parliament between 2015 and mid 2025. The core question is how far polarization is visible in plenary debates, both in what party groups say and in how confrontational and hostile their language becomes.

Existing empirical measures of polarization are not well suited to this task. Survey based indices, such as gaps between in party and out party ratings on feeling thermometers, capture affective polarization among citizens but do not directly address elite rhetoric in a multilingual, multiparty legislature. Roll call based estimates of ideological positions summarize legislative conflict, yet they are typically designed for two party systems and are observed only for recorded votes on a subset of decisions. In the European Parliament, where coalition patterns vary across policy areas and where debates are conducted in more than twenty official languages, such measures have difficulty distinguishing semantic disagreement from institutional procedure and agenda choice. Proksch and Slapin show that plenary speeches in the European Parliament provide a rich record of how parties position themselves on European integration and national dimensions, and that speech data can recover latent conflict structures that roll call votes obscure (Proksch and Slapin 2010). What is still missing is a speech based, multilingual, multiparty measure that separately tracks divergence in semantic content and in hostile or confrontational tone over time and across policy domains.

The dissertation addresses this gap by analysing a corpus of around 160,000 plenary speeches delivered in the European Parliament between 2015 and mid 2025 across roughly 25 official languages. At a first layer, each speech is represented by a multilingual text embedding, party group centroids are constructed over time and by topic, and a group based polarization statistic (cluster polarization coefficient, CPC) is used to quantify how much of the variation in this semantic space is structured by party labels (Mehlhaff 2024). At a second layer, a large language model is used as a coder of rhetorical style, assigning each speech scores on three 0–10 scales for conflict, tone toward outgroups, and extremity, and the same party by time and party by time by topic aggregation logic is applied to these tone dimensions, again using CPC to summarize party structured divergence. Both layers are implemented in pooled and language balanced variants that control for the distribution of languages across parties. Robustness checks include permutation based null models that randomize party labels within language and topic strata, alternative embedding architectures, and English only subsets, providing a common uncertainty framework for the semantic and tone based indices.

Methodologically, the thesis contributes a unified measurement design for tracking rhetorical polarization in a multilingual, multiparty legislature. It separates, in a transparent way, two dimensions that are often conflated: what parties talk about, captured by high dimensional semantic embeddings, and how hostile or alarmist their rhetoric sounds, captured by LLM derived tone scores. The same CPC framework is applied in both spaces, which allows direct comparison of how much variance in semantic content and in hostile tone is structured by party identity, and how these layers evolve over time and across topics. The approach is explicitly group based, focusing on between party separation while preserving within party heterogeneity, and it accommodates multiple languages and issues through language and topic de-meaning and topic specific semantic spaces. In this way, the study extends text based position taking work in the European Parliament beyond the estimation of static ideological dimensions to the dynamic tracking of group based polarization in both content and style.

On the applied side, the dissertation delivers a speech based polarization tracker for the European Parliament that covers the decade from 2015 to mid 2025 and is organized by macro topic, time, and party group. The tracker is implemented on top of Eurowatch, a custom data infrastructure and dashboard that aggregates plenary transcripts into a searchable database and provides the research-ready exports used in all analyses.[1] Using CPC on embeddings and on hostile tone scores, the tracker maps how semantically measured rhetorical polarization and party structured hostility evolve over time and identifies which topics and party pairs drive the main peaks. The headline empirical result is that the

[1] A public demo of Eurowatch is available at https://www.koslowsky.net/app/ (accessed 2 January 2026). Eurowatch is an unofficial research tool and is not affiliated with the European Union.

Parliament is not semantically “hyper polarized”: party labels explain only a small share of variation in speech embeddings, and increases in semantic CPC are concentrated in specific domains such as migration, foreign policy, health, and social policy at particular crisis moments rather than in a uniform chamber wide trend. By contrast, CPC in the hostile rhetoric space is substantially larger, indicating that party identity plays a much stronger role in structuring how often and how intensely conflictual, outgroup oriented, and alarmist language is used. Peaks in both content and tone are closely associated with major crises, including the refugee crisis, COVID 19, and Russia’s full scale invasion of Ukraine, yet these spikes remain topic specific and do not erase a shared semantic ground in which mainstream pro EU parties show more overlap than the ideological flanks. The core contrast between modest semantic distances and sharper differences in rhetorical temperature is robust across pooled and language balanced specifications, permutation based nulls, alternative embedding models, and English only subsets. In line with comparative work that documents divergent trajectories of affective polarization across advanced democracies (Boxell, Gentzkow, and Shapiro 2024; Gidron, Adams, and Horne 2020), the European Parliament therefore appears as a selectively polarized arena rather than a simple analogue of the American case. The distinction between semantic and tonal polarization also speaks to the normative ambiguity highlighted by Ahler and Broockman, who show that ideologically “polarized” representatives can in some settings represent heterogeneous citizen preferences more accurately than centrists (Ahler and Broockman 2018). The results here suggest that in the European Parliament, modest semantic distances coexist with sharper differences in rhetorical style, which may have different implications for representation, accountability, and the quality of democratic discourse.

The remainder of the thesis is structured as follows. The next chapter situates the study within the literatures on polarization, its measurement, and event driven dynamics, and develops the conceptual distinction between semantic disagreement and hostile tone in more detail. Chapter 3 describes the Eurowatch infrastructure, the construction of the multilingual speech corpus, and the descriptive statistics that characterize speaking activity by party, language, and topic. Chapter 4 presents the methodological framework, including the embedding based CPC, the LLM based coding of rhetorical tone, and the implementation of the permutation null and bootstrap procedures. Chapter 5 reports the main results on rhetorical intensity, semantic CPC, and hostile tone CPC across topics and over time. Chapter 6 discusses limitations of the data and methods, and Chapter 7 concludes by drawing out the substantive and methodological implications of the findings and outlining directions for future research.

2 Literature Review

2.1 What is Polarization?

Following Mehlhaff (2024), polarization is defined as a configuration with two core features: greater heterogeneity between groups and greater homogeneity within groups. This view is closely aligned with DiMaggio, Evans, and Bryson's classic four dimensional framework, in which polarization is diagnosed along dispersion, bimodality, constraint, and consolidation (DiMaggio et al. 1996). Dispersion refers to how far apart people's opinions are spread in a distribution. Bimodality indicates whether opinions cluster into two opposing camps rather than around a single middle position. Constraint captures how strongly views on different issues hang together as a consistent package. Consolidation describes the extent to which opinions are tied to social or group identities, such as party, religion, or class. In their analysis of US survey data from the 1970s to the early 1990s, DiMaggio et al. find limited evidence of strong mass polarization on most social issues, with abortion and the Democrat-Republican divide as important exceptions, and in doing so establish polarization as a multidimensional, group structured phenomenon rather than a single scalar.

Esteban and Ray provide a complementary, formal account that places group clustering at the center of the concept (Esteban and Ray 1994). They define a society as polarized when individuals can be grouped into a small number of internally similar but mutually distant clusters, and they emphasize that this structure is distinct from inequality: a distribution can be unequal without clear camps, and it can be highly polarized with two antagonistic blocs even if overall inequality is modest. In their framework, polarization is normatively salient because such clustered configurations are closely linked to social tension and the potential for conflict.

On this conceptual foundation, subsequent work distinguishes more sharply between ideological and affective polarization. Fiorina, Abrams, and Pope argue that most American voters remain near the center on many cultural issues and that much of the apparent change reflects partisan sorting rather than mass movement toward extremes (Fiorina et al. 2004). By contrast, Iyengar, Sood, and Lelkes characterize contemporary US polarization as primarily affective: feeling thermometer data show stable warmth toward one's own party but sharply declining evaluations of the opposing party (Iyengar et al. 2012). Mason's account of partisan "mega identities" highlights how partisanship has come to fuse with racial, religious, and cultural cleavages, deepening social distance between camps (Mason 2018), while Iyengar and Westwood demonstrate experimentally that party cues alone now trigger automatic negative affect and discriminatory behavior toward out partisans on a scale comparable to, or exceeding, discrimination based on race (Iyengar and Westwood 2015). Taken together, this literature recasts polarization less as simple ideological extremity along a left right continuum and more as the consolidation of internally cohesive, mutually distrustful camps.

These developments also underline the normative ambiguity of polarization. Much of the discussion treats rising polarization as a threat to compromise, institutional trust, and democratic stability. Ahler and Broockman, however, show that ideologically polarized representatives can, under plausible conditions, more accurately reflect heterogeneous citizen preferences than centrist politicians, because they offer clearer and more consistent positions that match voters' issue bundles (Ahler and Broockman 2018). They describe this as a "delegate paradox", in which legislators who appear extreme in ideological space may in practice be closer to their constituents on individual issues. Polarization thus emerges as a multi dimensional and normatively mixed phenomenon: it can increase social hostility and weaken cross cutting ties, but it can also sharpen political choices and clarify lines of representation.

2.2 Measurement of Polarisation

2.2.1 Classical Approaches

Classical approaches to polarization measurement fall into three broad families that differ in what they treat as the relevant object of polarization. A first tradition focuses on ideological dispersion in party systems. Dalton develops a party system polarization index that uses voters' placements of parties on a 0-10 left right scale and weights these positions by each party's electoral strength (Dalton 2008). Parties are first located at their mean left right positions; deviations from the vote share weighted system mean are then squared, multiplied by the corresponding vote shares, summed, and square rooted to yield a single polarization score. The index is designed to capture how far electorally important parties are spread along the ideological continuum, independently of the sheer number of parties, and has become the standard comparative measure of party polarization based on vote share weighted ideological distances.

A second line of work conceptualizes polarization as a property of group structure within a continuous distribution. Duclos, Esteban, and Ray develop a family of indices in which polarization is defined as the aggregate of "effective antagonisms" between parts of the distribution, combining alienation (how far apart groups are) and identification (how many individuals cluster at each location) into a single measure (Duclos, Esteban, and Ray 2006). Their framework yields polarization indices that generalize inequality measures by putting additional weight on large, internally cohesive groups and provides estimators and inference tools for comparing polarization across countries or over time.

A third tradition shifts attention from ideological or material attributes to affective relations between partisan groups. Iyengar, Sood, and Lelkes argue that contemporary mass polarization is best understood as affective hostility between partisans and operationalize it as the gap between in party and out party evaluations on feeling thermometers and related items (Iyengar, Sood, and Lelkes 2012). Iyengar and Westwood extend this approach by combining implicit association tests, feeling thermometers, and behavioral allocation tasks, consistently defining polarization as the distance between co partisans and opponents in affective and discriminatory responses (Iyengar and Westwood 2015).

2.2.2 Group-based (CPC) approach

Mehlhaff (2024) treats polarization as a fundamentally group based phenomenon with two core features: groups move farther apart from one another and become more internally similar. To capture both elements in a single statistic, he proposes the cluster polarization coefficient (CPC), which decomposes total variance into between group and within group components and defines polarization as the share of total variance accounted for by between group differences. The CPC is defined for multiple groups and multiple dimensions, comes with a small sample corrected version, and can be applied to any multivariate data structure where observations can be assigned to groups. Because it is formulated over generic multidimensional variables, it can be used not only with ideal points and survey scales but also with high dimensional text embeddings, by treating each document's embedding as an observation and grouping by party, party family, or topic. This dissertation adopts this framework for rhetorical polarization in the European Parliament and returns to the formal definition and implementation details in the methodology section.

2.2.3 Text-based and speech-based measures

A growing body of work measures political positions and partisan conflict directly from texts. Slapin and Proksch introduce Wordfish, an unsupervised scaling model that uses word frequencies in documents such as party manifestos to estimate latent positions of parties on one or more dimensions (Slapin and Proksch 2008). Word counts are modeled as Poisson outcomes with document specific positions, word specific discrimination parameters, and controls for overall word frequency and document length. The resulting latent scales provide time series estimates of party locations, from which party system polarization can be derived as changing distances between parties in the estimated space.

Lauderdale and Herzog extend this logic to legislative speech by developing Wordshoal, a two stage model that operates on entire speech corpora (Lauderdale and Herzog 2016). In a first step, each debate is scaled separately, using a Wordfish type model to recover a debate specific dimension that captures systematic differences in word use among speakers. In a second step, these debate level scales are linked through a Bayesian factor model that yields a common latent position for each legislator and a loading parameter for each debate. This structure allows speech based polarization to be characterized both as separation in party means on the common legislator scale and as variation over time in how strongly individual debates load on the partisan dimension.

Other contributions focus more directly on measuring linguistic divergence between groups. Monroe, Colaresi, and Quinn propose a Bayesian log odds framework that identifies "fightin words" by estimating, for each term, a smoothed log odds contrast between groups and an associated z score, which can be summarized over time to track shifts in the intensity of partisan language differences (Monroe, Colaresi, and Quinn 2008). Gentzkow, Shapiro, and Taddy develop a general method for quantifying group differences in high dimensional choice settings and apply it to congressional speech (Gentzkow, Shapiro, and Taddy 2019). They define partisanship as the ease with which an ideal observer could infer a speaker's party from a single utterance under a multinomial logit model of phrase choice, and show that naive divergence measures suffer from severe finite sample bias in large vocabularies. To address this, they introduce leave out and penalized estimators that correct the bias while remaining computationally feasible, yielding a text based index of partisan distinctiveness that is explicitly grounded in predictive separability of linguistic choices.

2.2.4 LLM-based measurement

Recent work shows that large language models can be used as general purpose measurement tools for political texts without custom training. Ornstein, Blasingame, and Truscott demonstrate that few shot prompting of models such as GPT 3 and GPT 4 can deliver sentiment scores, tone classifications, ideological content measures, and topic labels that typically outperform standard automated approaches and achieve accuracy comparable to human coders, at far lower cost and with minimal preprocessing (Ornstein, Blasingame, and Truscott 2025). In their applications to political tweets, campaign ads, party manifestos, and congressional speeches, each task is framed as a constrained completion problem and continuous measures are constructed from the models' predicted probabilities over labels, treating LLMs as flexible coders rather than as objects of study in their own right.

The study also formalizes a set of best practices for LLM based text analysis: prompts should resemble instructions to a research assistant, include clear definitions and balanced few shot examples, and be validated against human coded gold standards before being scaled to full corpora. At the same time, the authors emphasize limitations, including dependence on proprietary models, calibration and reproducibility concerns, and the risk that biases in training data are reproduced in measurement. Taken together, this work positions prompt based LLM coding as a powerful but validation dependent strategy for generating fine grained, low cost annotations of political text, a role that is directly relevant for measuring rhetorical tone and targeting in parliamentary speeches.

2.3 Empirical trends and event driven polarization

Event centered research shows that major shocks can push political conflict toward more extreme and polarized configurations. Funke, Schularick, and Trebesch use a long historical panel of 20 advanced democracies from 1870 to 2014 to study political developments after systemic banking crises, combining detailed data on electoral outcomes, parliamentary fragmentation, and protest with crisis chronologies and local projection event studies (Funke, Schularick, and Trebesch 2016). They find that, relative to normal recessions and equally severe non financial downturns, banking crises are followed by systematic gains for far right parties, shrinking government majorities, more fragmented parliaments, and heightened social unrest, implying that financial crises are "special" events with particularly strong polarizing effects.

Other work links polarization directly to responses during crises and to elite rhetoric around such events. Milosh et al. show that in the United States during the first wave of COVID 19, local mask use is most strongly predicted by Trump's 2016 vote share, and that state and local mask mandates do little to close this partisan gap, while presidential messaging produces sharp shifts in online attention and sentiment (Milosh et al. 2021). Lauderdale and Herzog measure polarization from legislative speech using their Wordshoal framework and document that government-opposition polarization in the Irish Dáil and partisan polarization in the US Senate exhibit pronounced spikes and dips around major shocks, including the 2008 financial crisis, the 9/11 attacks, and the 2009-2010 health care reform debates (Lauderdale and Herzog 2016). Taken together, these studies suggest that crises and salient political events are closely associated with changes in both electoral and rhetorical polarization, and that speech based measures are particularly sensitive to such event driven dynamics.

3 Data and Descriptive Statistics

3.1 Eurowatch

Eurowatch is a custom data infrastructure and web based dashboard that I developed to provide a workable environment for exploring plenary speeches and producing descriptive statistics on European Parliament debates. It consists of a SQLite database built from the official plenary transcripts, an API layer, and a lightweight browser interface. Users can filter speeches by date range, political group, language, topic, and individual MEP, inspect the underlying text through a speech browser and session level listings, and view basic charts and tables summarising speaking activity.

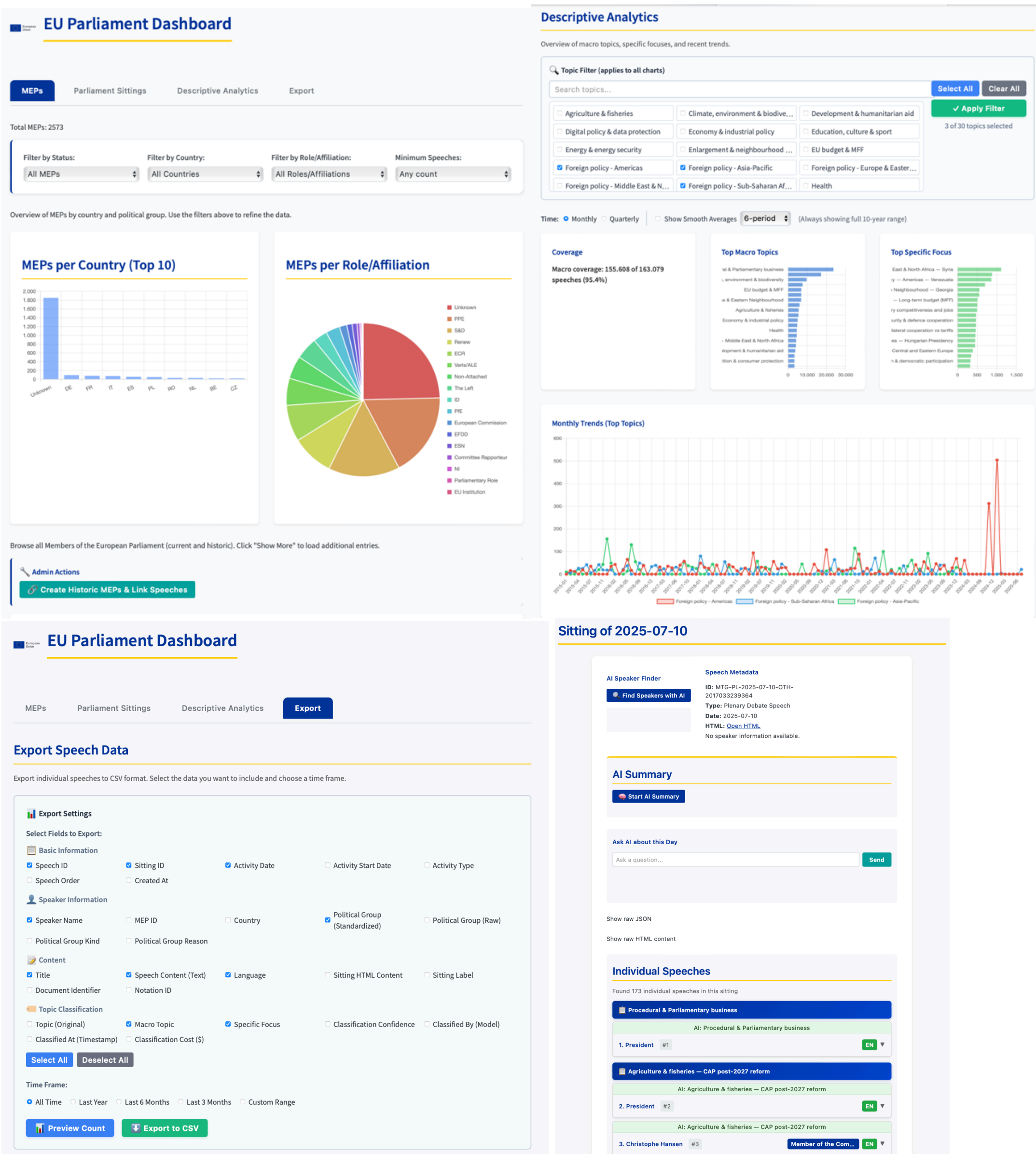

Figure 1: Eurowatch Tool

Eurowatch is intended both as a research aid and as a transparency mechanism. It supports multi dimensional filtering and simple aggregations, such as yearly or monthly trends in speaking activity, topic distributions across party groups, and language coverage over time. Full text search and type ahead speaker and topic search functions make it possible to move quickly between specific interventions and more aggregated views. The same infrastructure underlies the research ready exports used in the quantitative analyses in later chapters, so that exploratory plots, descriptive statistics, and the CPC based polarization measures are all derived from a single, documented data pipeline. This improves reproducibility of the analyses and makes it straightforward to extend the tracker to new time periods or additional topics.

All scripts, configuration files, and the web application code are available in the public Eurowatch GitHub repository[2], which provides the main technical documentation for the system and can be cited as an online resource.

3.2 Data Retrieval and Preprocessing

The empirical analysis builds on the official plenary transcripts published on the European Parliament website. For each sitting, the Parliament publishes an HTML record of the debates on its website, for example for the sitting of 10 July 2025 [3]. The Eurowatch pipeline uses custom Node.js scripts to query the Parliament API for sitting metadata and to download these raw HTML files for all sessions between January 2015 and mid 2025. The raw HTML pages are stored in a sittings table together with the sitting date, language, and basic contextual information.

From these sitting level HTML files, individual speeches are parsed using a combination of CSS selectors and pattern matching. The parser identifies speaker turns based on recurring patterns in the transcripts (for example "Name, Role. - Speech" or "Name (Party). - Speech") and extracts the speaker name, any role or title, the declared political affiliation, and the speech text. These segments are written to an individual_speeches table and linked to Members of the European Parliament using fuzzy name matching against a separate meps table. A second parsing pass improves the handling of parentheses and multi part titles, which is important for reliably recovering political groups from strings such as "Name (Party), Role".

---

[2] https://github.com/paulkoslo/eurowatch (accessed 28.11.2025)

[3] European Parliament, "Verbatim report of proceedings", available at: https://www.europarl.europa.eu/doceo/document/CRE-10-2025-07-10_EN.html (accessed 28.11.2025).

Each speech is then enriched with language, political group, and topic information. Language is detected using a combination of CLD3 [4] and franc [5] constrained to the main EU languages, with script based heuristics for Greek and Cyrillic. Political group information, which appears in several hundred distinct string variants in the raw data, is normalized to a small set of canonical group labels (PPE, S&D, Renew, Greens/EFA, ECR, ID, The Left, and others), while institutional roles such as the President or Commission representatives are flagged and treated as non attached (NI). Topics are first recovered from the HTML by parsing agenda headers and document identifiers and aligning them to speeches; these original parliamentary topic labels are then mapped to a controlled vocabulary of 29 macro topics by an LLM based classifier, with macro topic and specific focus labels propagated to all speeches sharing the same original topic string.

Quality control steps remove duplicate sittings and speeches and produce coverage statistics by period, language, and political group. The final output is a set of research ready CSV exports derived from the SQLite database, including full speech level files and group and speaker summaries. These exports form the basis for the descriptive statistics and polarization measures reported in the remainder of the thesis, and their construction is fully documented in the Eurowatch GitHub repository.

3.3 Descriptive Analytics

The final Eurowatch database contains around 163,000 plenary speech segments between January 2015 and mid 2025. Language detection succeeds for 161,858 speeches, corresponding to a classification rate of roughly 99.9 percent. English is the single most common language (41,580 speeches), followed by German (17,476), French (16,175), Italian (13,309), and Spanish (12,369). The remaining speeches are distributed across roughly twenty additional official EU languages. This distribution is consistent with the composition of the Parliament and its working languages, but still leaves substantial material in smaller linguistic communities, which is important for the multilingual analyses in later chapters.

Macro topic labels are available for 155,608 of 163,079 speeches, or 95.4 percent of the corpus. The topic taxonomy is intentionally broad and organised around approximately thirty macro themes. Procedural & Parliamentary business is by far the largest category, reflecting the routine adoption of agendas, rules of procedure, and institutional reports that accompany

[4] https://github.com/google/cld3 (accessed 10.12.2025)
[5] https://www.npmjs.com/package/franc (accessed 10.12.2025)

almost every sitting. Rule of law & fundamental rights forms the second largest block, followed by Climate, environment & biodiversity, Social policy & employment, EU budget & MFF, and Migration & asylum. Each of these high volume topics attracts several thousand speeches over the period and together they account for the majority of classified contributions. The remaining macro topics, such as Trade & globalisation, Security & defence, or Digital policy & data protection, form a long tail of more specialised but still sizeable domains.

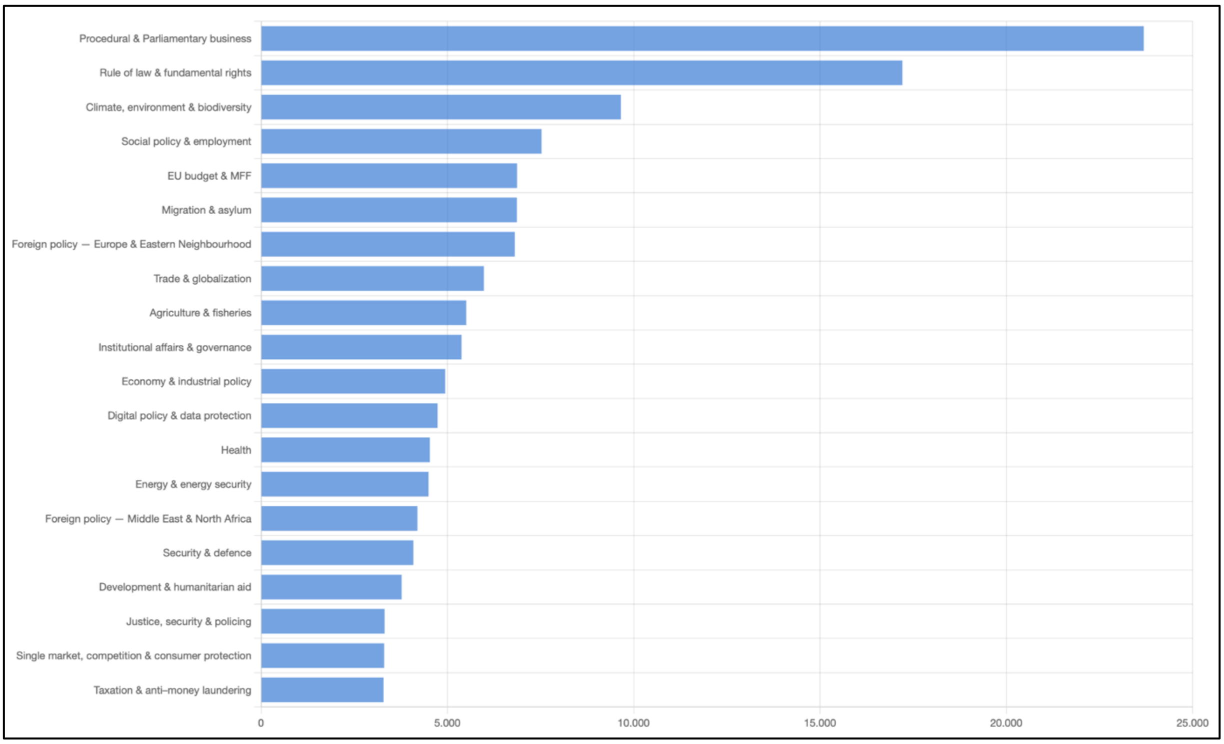

Figure 2: Top Macro Topics

The more fine grained specific focus labels highlight the concrete issues that dominate parliamentary attention. Among the most frequently discussed dossiers are external crises such as the war in Syria, the situation in Ukraine, and developments in Venezuela, as well as long running internal conflicts around the rule of law in Hungary and Poland. Other prominent files include negotiations on the long term EU budget and own resources (MFF), the EU Canada PNR data transfer and privacy safeguards, debates over the TikTok platform and digital security, and the European Media Freedom Act. This distribution confirms that the corpus captures both high profile geopolitical crises and technically dense legislative files, many of which later appear as focal points in the polarization analyses.

Monthly time series of speech counts for selected macro topics illustrate how plenary attention shifts over time.

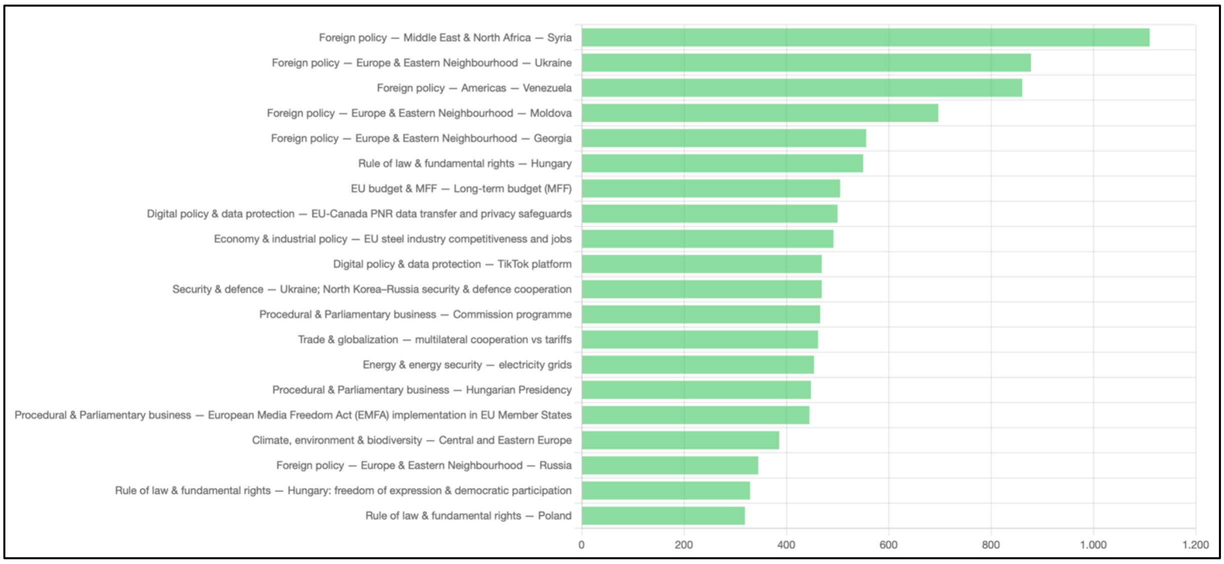

Figure 3: Top Specific Topics

Security and defence displays a long period of relatively low and stable activity, with only occasional peaks linked to specific debates. From late 2023 and especially in 2024 and early 2025, however, the series jumps to a much higher level, suggesting that defence policy becomes a more central and recurrent concern toward the end of the legislature.

Foreign policy in Eastern Europe follows almost the opposite pattern. Volumes are modest and volatile up to 2020, but increase sharply from 2021 onward, with sustained high levels and repeated spikes in 2022, 2023, and 2024 that coincide with the escalation of conflict and instability in the region.

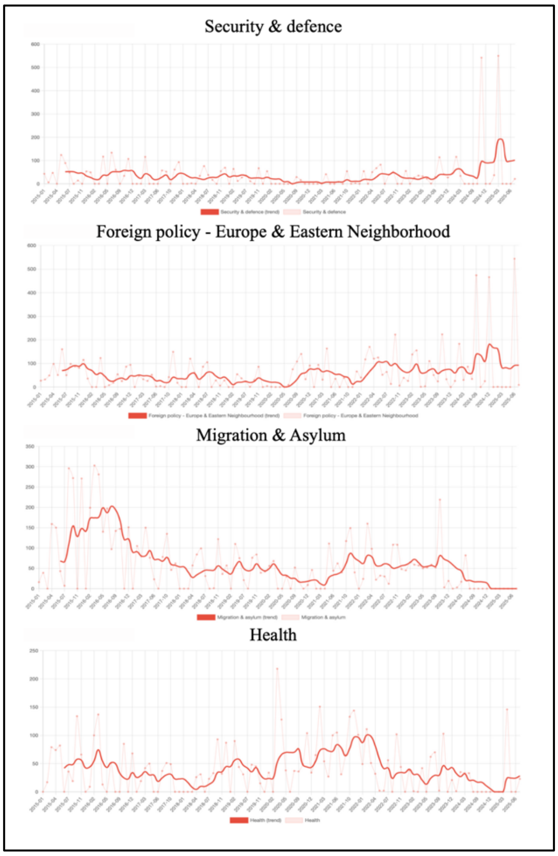

Figure 4: Selected Macro Topics over Time

Migration and asylum is highly concentrated in the first part of the period, with very large peaks in 2015 and 2016 during the refugee crisis, followed by a steady decline and only intermittent smaller spikes around 2019 and 2021 to 2022. By 2024 and 2025, speeches on migration become rare, indicating that the issue has largely moved off the plenary agenda.

Health shows a moderate baseline in the early years, followed by a marked increase beginning in

2019 and peaking in 2020 and 2021 during the COVID 19 pandemic and its aftermath. After 2021 the number of health related speeches declines again, but remains somewhat above its pre pandemic level, indicating a lasting, though reduced, prominence of health on the parliamentary agenda.

4 Methodology

This chapter sets out a two-layer measurement strategy for rhetorical polarization in the European Parliament. The first layer captures what parties say by embedding speeches in a multilingual semantic space and assessing the extent to which party labels structure that space using Mehlhaff's cluster polarization coefficient (CPC). The second layer captures how parties say it by using a large language model to score speeches on conflict, outgroup tone, and extremity, and applying the same CPC framework to this interpretable tone space. Each element of the design – semantic embeddings, CPC, and LLM based coding – is established in its own right; the methodological contribution of the thesis lies in integrating them into a unified, multilingual, multiparty tracker that compares content based and tone based polarization under a common scheme for aggregation, balancing, and uncertainty assessment. Section 4.1 describes the embedding based CPC pipeline, and Section 4.2 presents the tone scoring rubric and the parallel CPC pipeline in the tone space.

4.1 Embedding-based polarization in semantic space (CPC)

Rhetorical polarization is measured using Mehlhaff's cluster polarization coefficient (CPC), which treats polarization as the share of total variance in a multivariate space that is accounted for by party membership. Conceptually, CPC increases when party centroids move farther apart and when speeches within each party cluster more tightly around their centroid. Formally, CPC is introduced in *Equation (1)* as

$$CPC = \frac{BSS}{TSS} = 1 - \frac{WSS}{TSS}, \quad CPC \in [0,1]$$

where total scatter (TSS), within party scatter (WSS), and between party scatter (BSS) are defined in terms of squared deviations from the overall mean and party means. *Equation (2)* gives the explicit formulas for TSS, WSS, and BSS, using the notation for speech embeddings and party centroids.

$$TSS = \sum_{i=1} \sum_{j=1} \left(x_{\{ij\}} - \mu_j\right)^2$$

$$WSS = \sum_{k=1} \sum_{i=1} \sum_{j=1} \left(x_{ikj} - \mu_{kj}\right)^2$$

$$BSS = TSS - WSS = \sum_{\mathrm{k}=1} \sum_{\mathrm{j}=1} \left(\mu_{\{\mathrm{kj}\}} - \mu_{\mathrm{j}}\right)^2$$

A small sample corrected version CPC_{adj} is provided in Equation (4);

$$CPC_{adj} = 1 - (1 - CPC)\frac{n_i - n_j}{n_i - n_k}$$

CPC_{adj} is a small sample correction analogous to an adjusted R^2, intended to remove the upward bias that arises when polarization is estimated with many dimensions and a limited number of observations relative to the number of parties (Mehlhaff 2024). In this setting, speeches are projected into a relatively high dimensional PCA space and many macro topic by year bins contain only moderate numbers of speeches and speakers per party. As a result, the correction term often becomes large, so that CPC_{adj} frequently takes values near zero or even negative despite clearly positive CPC values, which undermines its interpretation as a share of variance explained by party membership. Moreover, CPC_{adj} is undefined for weighted designs and for bins with very small N, which would lead to internally inconsistent time series across modes and years. For these reasons, the analysis relies on raw CPC as the primary polarization statistic and uses the permutation based null model and cluster bootstrap intervals described below to characterize sampling uncertainty.

In all cases, CPC is interpreted as the proportion of semantic variance that is explained by party labels in a given macro topic and time bin.

The input for this metric is a plenary speech dataset containing at minimum the speech text, date, MEP name, political group, language, micro topic, and macro topic. Basic pre processing removes rows with missing information on any of these variables, drops extremely short interventions (for example texts shorter than 40 characters), and can exclude selected political groups if required. Before embedding, all speech texts are lowercased and standard boilerplate phrases such as “mr president” or “honourable members” are stripped so that the subsequent analysis focuses on substantive content. Each speech is then assigned to a time bin, by default a calendar year, although longer multi year bins can be defined if needed. Speeches are embedded using the *paraphrase-multilingual-mpnet-base-v2* sentence transformer, which maps each speech (up to 512 tokens) into a high dimensional semantic vector (Reimers and Gurevych 2019). Embeddings are normalized to unit length. To prevent systematic language and agenda differences from dominating party separation, embeddings are demeaned by language and, in the default specification, by language by micro topic cells. This step removes stable offsets associated with translation and issue mix. The adjusted embeddings are then globally centered, and for each macro topic a single principal components analysis is fitted on all speeches in that topic across all time bins. All speeches in the macro topic are projected into this shared PCA space, retaining up to 100 components subject to the usual rank constraint. The definitions of the multivariate positions used in *Equation (2)* are based on these macro topic specific principal component scores.

Within each macro topic and time bin, three alternative designs construct the matrices on which CPC is computed.

In the speaker averaged mode, all speeches by a given MEP within a bin are averaged to form a single embedding, and each MEP is assigned to the party for which they most frequently speak in that period; guardrails require that each party has at least a minimum number of speakers and that a minimum number of parties are present before a bin is retained. This mode is best interpreted as treating each MEP as one unit of analysis, so that polarization reflects how far apart party delegations are in terms of their typical members’ speech profiles.

The speech capped mode treats individual speeches as observations but limits the number of speeches per MEP in a bin to a fixed maximum (for example five), again with party level guardrails on the minimum number of speeches and speakers. Here, polarization is interpreted as party separation when each MEP’s voice is prevented from dominating, approximating a scenario in which all representatives contribute a bounded number of interventions.

The speech weighted mode includes all speeches without a per speaker cap and can weight observations by speech length as a proxy for speaking time, subject to the same guardrails; bins that fail these criteria are flagged as low sample and their CPC values are treated as missing. In this mode, polarization is interpreted as rhetorical distance “as heard” in the chamber, with parties exerting influence in proportion to how much and how long they actually speak.

For each valid bin and each mode, the party specific centroids and overall mean are computed in the PCA space, and the scatter components in *Equation (2)* are calculated. Within party scatter is the sum of squared deviations of observations from their party centroid; between party scatter is the sum over parties of group size multiplied by the squared distance between the party centroid and the overall mean; total scatter is their sum. *Equation (1)* is then used to obtain CPC for that bin. The CPC values, together with the underlying BSS and TSS components, form the core time series of rhetorical polarization by topic and weighting mode.

Two additional procedures quantify uncertainty and provide diagnostic information. A label shuffle null model generates a benchmark for each bin by permuting party labels at the speaker level within language by topic strata and recomputing CPC many times, holding the semantic structure of the speeches fixed. The resulting null mean and standard deviation are used to compute a z score that summarizes how strongly speech content is clustered by party relative to random assignment of party labels. A cluster bootstrap resamples speakers with replacement, rebuilds the design matrices for each bootstrap sample, and recomputes CPC, yielding empirical confidence intervals for the polarization series. Finally, for the most recent valid bin in each macro topic and mode, pairwise standardized distances between party centroids are computed in the PCA space by dividing the Euclidean distance between centroids by the average within party spread. These distances are used to identify which party pairs are most distinct in rhetorical space and to complement the aggregate CPC index in the descriptive reporting.

4.2 LLM-based coding of rhetorical tone and tone-space CPC

The LLM based component treats a large language model as a fast, consistent coder of rhetorical tone for the entire corpus of around 160,000 plenary speeches from 2015 to 2025. Each speech is scored on three 0-10 dimensions: conflict, outgroup tone, and extremity. Conflict captures visible contestation and confrontational language; outgroup tone captures how speakers talk about political opponents, migrants, member states, or other salient groups; extremity captures alarmist or crisis framing and calls for radical action. Each scale is anchored with verbal descriptions at 0, 5, and 10, distinguishing neutral or technical interventions from clearly adversarial, hostile, or alarmist rhetoric and clarifying common European Parliament edge cases such as criticism without hostility or dramatic but non extremist language. The full anchor descriptions and the complete prompt (including worked examples) are reproduced in Appendix.

Prompt design follows recent “LLM as coder” practice in political text as data work (Ornstein, Blasingame, and Truscott 2025): the model is instructed to act as a political text coder, to return only a single JSON object with three integer scores, and to follow explicit “do” and “do not” guidance, with multiple worked examples illustrating low, medium, and high levels of each dimension. Lightweight metadata on language, party group, and macro topic are passed before the truncated speech text to situate rhetoric in context. An initial pilot on a random sample of 1,000 speeches, using an Azure hosted small model, produced numeric scores that were checked for face validity; definitions and examples were refined until high scoring cases consistently matched intuitive notions of adversarial, hostile, and alarmist speech.

For the full run, the prompt was simplified by removing free text notes and tightening instructions, while retaining strict JSON output and safety language about coding but not reproducing explicit sexual or violent content. Because Azure content moderation blocked reliable classification for some sensitive debates, notably speeches on the Israel-Gaza conflict, scoring was moved to an OpenAI hosted small model (*“gpt-5-nano-2025-08-07”*). A custom calling routine tracked errors, distinguished hard from transient failures, implemented limited retries, and processed only speeches with missing scores, with parallelization and periodic checkpointing to ensure robustness.

The resulting file (eu_speeches_with_scores_final.csv) combines the original speech metadata with three LLM derived scores per speech. After basic face validity checks and pilot validation, these scores provide a compact, three dimensional representation of rhetorical style that serves as an alternative input space for polarization analysis. The same CPC machinery used for the embedding based analysis is then applied to these three dimensional score vectors, with identical macro topic slices, yearly time bins, language and language-topic demeaning, guardrails, and the three weighting modes (speaker averaged, speech capped, and speech weighted). For each topic, year, and mode, CPC (and, where relevant, its adjusted form) is computed on the LLM score space, accompanied by stratified label shuffle nulls, speaker clustered bootstrap intervals, and pairwise standardized centroid distances between party centroids. In this way, the LLM based polarization tracker mirrors the embedding based pipeline in design and outputs, while operating on an explicitly interpretable set of rhetorical dimensions.

5 Results

5.1 Rhetorical intensity: LLM-based speech scores

To describe overall patterns of rhetorical intensity, the analysis collapses the three LLM based scores into a simple speech level index. For each speech, a rhetorical polarization score is defined as the unweighted average of the conflict, outgroup tone, and extremity scores,

$$polarization\_index = (conflict + outgroup\_tone + extremity)/3.$$

Higher values indicate more conflictual, hostile, and alarmist rhetoric in the plenary record, but this remains a within speech intensity measure rather than a between party separation measure; the latter is captured later by CPC. The marginal distributions of the three underlying scores show that conflict is moderate on average and relatively dispersed, with most speeches falling between 2 and 6 and comparatively few at the extremes. Outgroup tone is more strongly skewed toward low values, with a large mass of speeches at 0-1, indicating either no clear outgroup or only mild criticism, and only a small minority of observations exhibiting strongly hostile language. Extremity scores cluster mainly in the 2-4 range, with a long but thin right tail, suggesting that explicitly alarmist or maximalist rhetoric is present but concentrated in a limited subset of interventions.
The evolution of average rhetorical intensity over time is captured by the mean

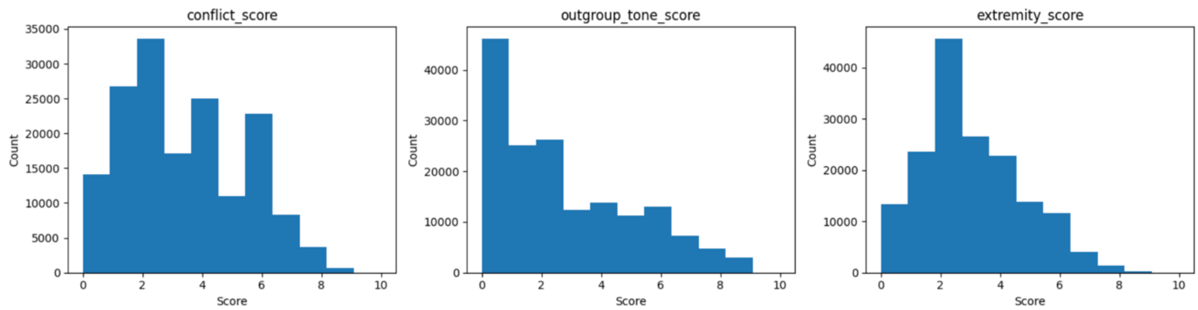

Figure 5: Score Distributions

polarization_index for all speeches by year. This series rises steadily from about 2.6 in 2015 to around 3.6 by 2025, indicating a gradual warming of plenary rhetoric over the decade rather than a pattern of isolated spikes followed by full reversion to baseline. All three components contribute to this upward drift, but the outgroup tone and extremity scores accelerate particularly after 2018-2019, consistent with an increasing use of hostile and alarmist framing alongside procedural and substantive conflict.

Party level averages reveal systematic differences across groups. Identity and Democracy (ID) and non attached MEPs (NI) display the highest rhetorical intensity throughout, starting from elevated levels and increasing further by 2025. The Left and the European Conservatives and Reformists (ECR) also maintain consistently higher scores than the main pro integration groups, indicating more adversarial and alarmist rhetoric on average. The mainstream groups (PPE, S&D, Renew, Greens/EFA) begin from noticeably lower levels in 2015 but exhibit a

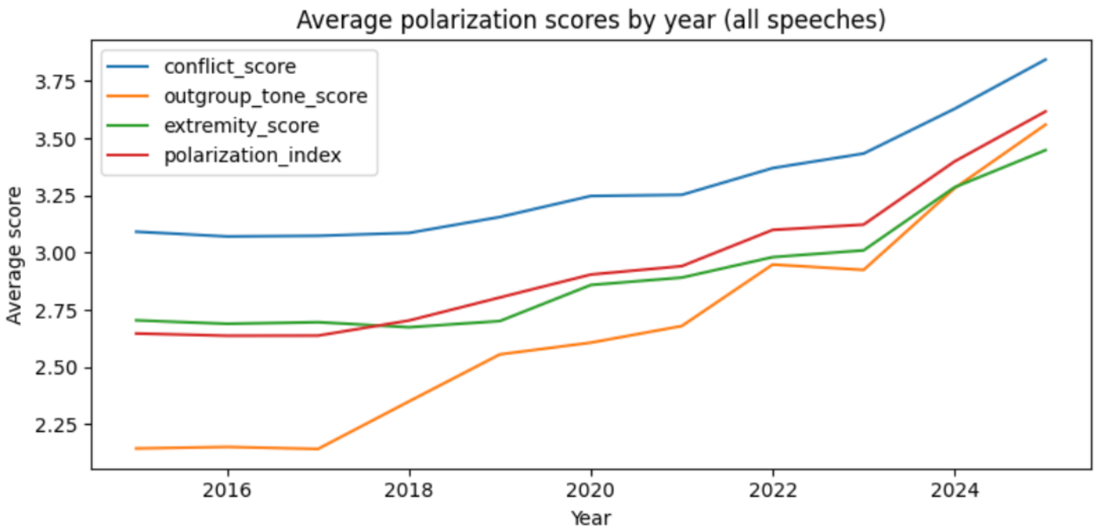

Figure 6: Average Polarization Scores by year

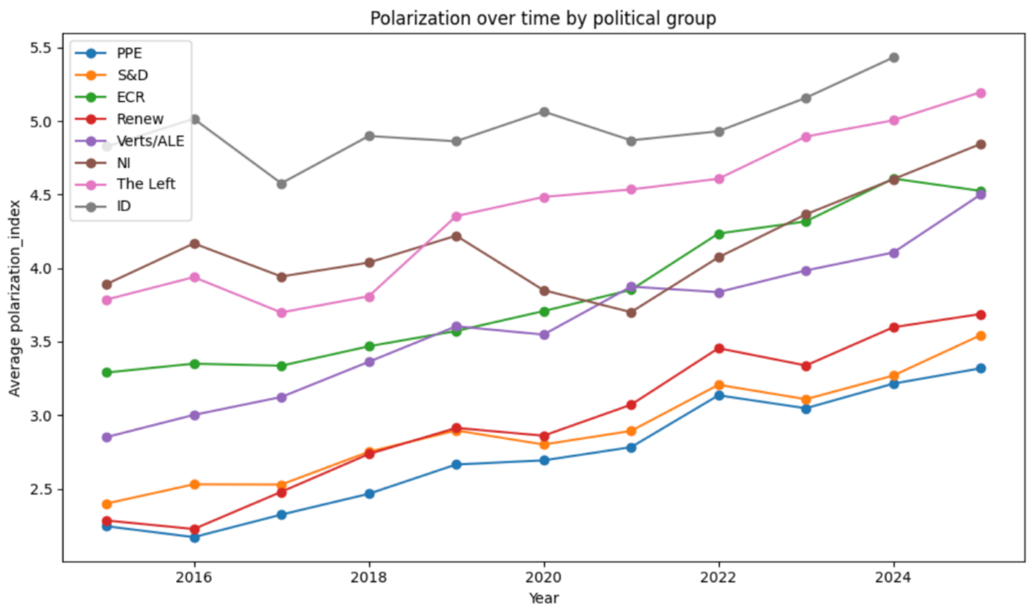

Figure 7: Polarization Over Time By Political Group

clear upward drift, especially from the late 2010s onward, narrowing though not eliminating the gap with the more radical groups. By 2025 the entire party system operates at a higher rhetorical temperature than in 2015, even as the rank ordering of groups remains broadly stable.

Rhetorical intensity also varies substantially across macro topics. Foreign policy debates on the Middle East and North Africa, the Americas, and Europe and the Eastern Neighbourhood have the highest average polarization_index values. All three domains display marked upward trends, with prominent peaks around 2019 and 2021-2022, consistent with heightened parliamentary conflict over successive international crises. The Europe and Eastern Neighbourhood series exhibits a particularly sharp rise in the early 2020s, indicating especially heated rhetoric on the EU's eastern neighbourhood compared to earlier years.

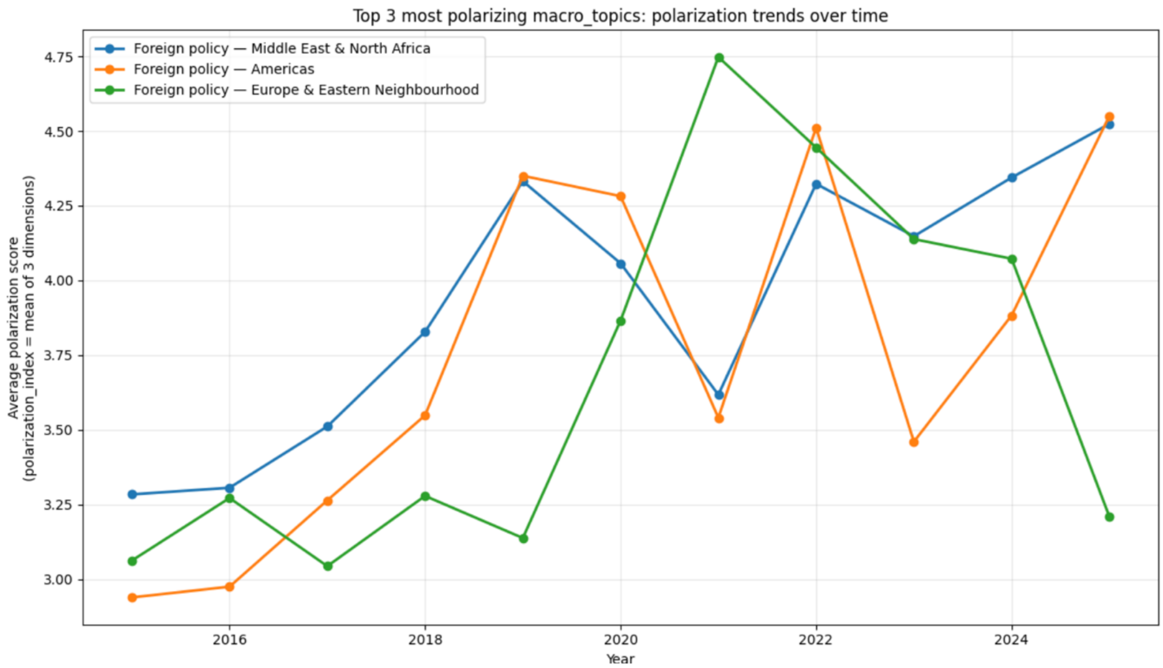

Figure 8: Top 3 Most Polarizing Macro Topics

Two further patterns underline the role of crises. In health related debates, polarization_index values are low and stable up to 2018, followed by a sharp but short lived spike around 2020-2021, in line with COVID 19 debates becoming considerably more conflictual and alarmist. After 2021, levels fall back but remain more volatile than in the pre pandemic period, suggesting incomplete reversion to the earlier baseline.

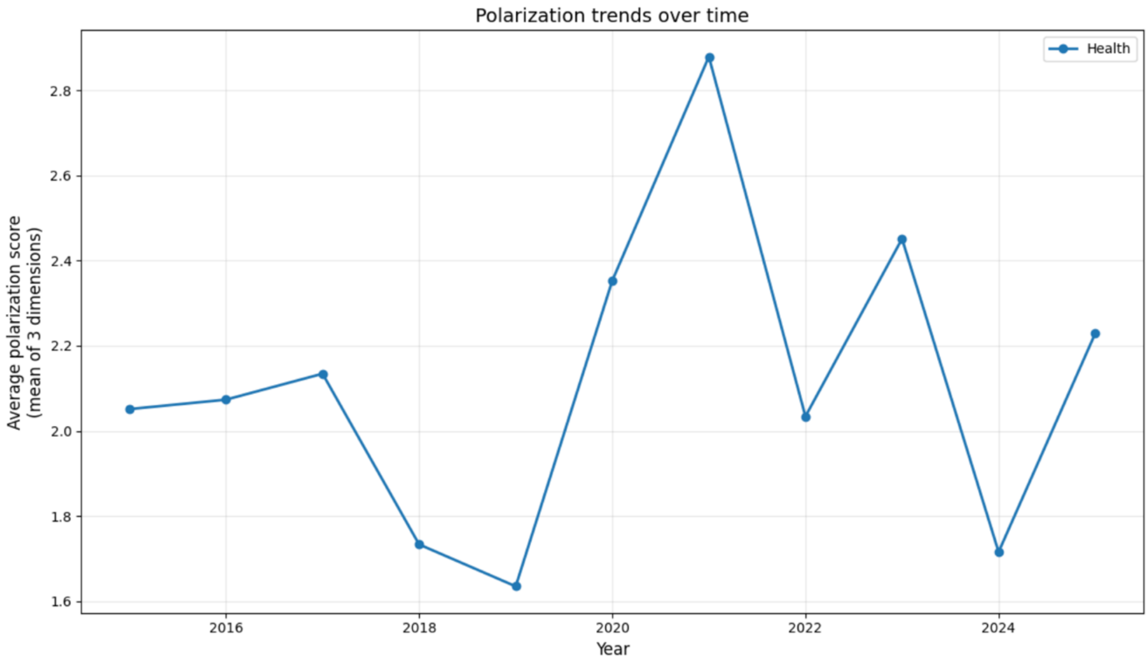

Figure 9: Health Polarization over Time

In addition, there is notable co movement across foreign policy and energy related topics. The Europe and Eastern Neighbourhood series shows a steep run up from about 2019 to 2021, and debates on energy and energy security follow a similar upward trajectory with a short lag, rising strongly around 2020-2022 rather than in isolation. The timing of these increases coincides with escalating tensions around, and the full scale Russian invasion of, Ukraine, suggesting that the Ukraine conflict is a central driver of heightened rhetoric in both domains. Taken together with the health spike, this pattern indicates that major public health, geopolitical, and energy shocks are associated with joint surges in rhetorical intensity across multiple macro topics rather than narrowly confined, topic specific bursts.

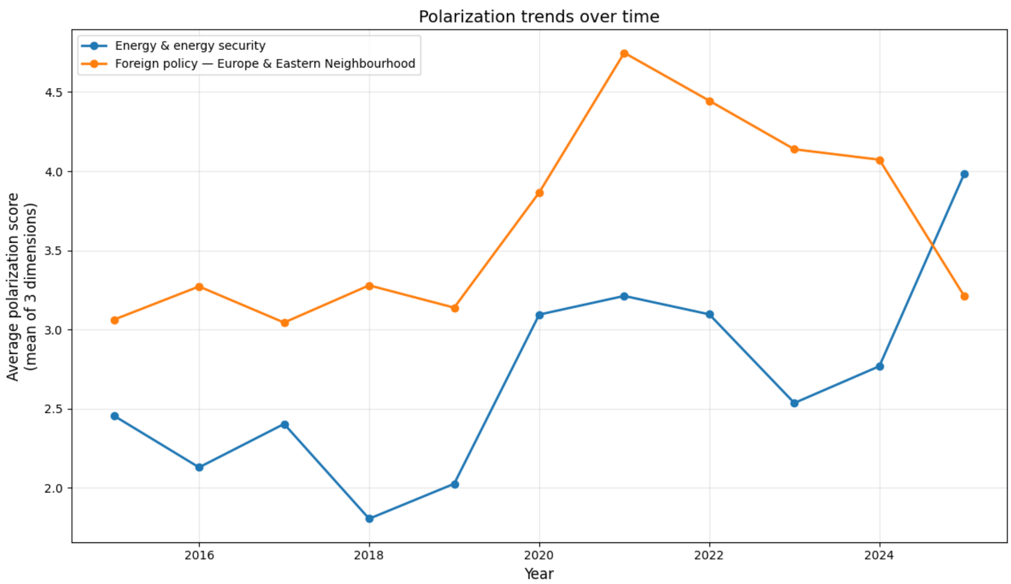

Figure 10: Polarization Around The Ukraine Conflict

5.2 Between-party polarization in semantic space: CPC on text embeddings

Semantic CPC results point to modest but structured rhetorical polarization in the embedding space. Across the full corpus, party labels explain only a small share of the variance in speech embeddings, with CPC values typically in the low single digits. In other words, the Parliament is not semantically hyper polarized in the aggregate. What polarization there is tends to appear in specific issue domains and time periods rather than as a smooth chamber wide upward trend. Differences across the three weighting designs are limited: the speaker averaged and speech capped modes give very similar pictures, while the speech weighted mode is somewhat more sensitive to bursts of activity by highly vocal speakers.

In Figures 11-20, the upper panels plot polarization over time: the solid line shows the observed cluster polarization coefficient (CPC) in each year under the three weighting schemes (MEP averaged, speech capped, speech weighted), and the shaded band indicates the 2.5th-97.5th percentile range from speaker cluster bootstrap resampling, which should be read as an uncertainty envelope rather than a strict 95 percent confidence interval. The lower panels display heatmaps for the latest well populated year, reporting standardized pairwise distances between party centroids (lighter cells indicating greater separation, darker cells closer proximity), thereby revealing which party pairs contribute most to the aggregate CPC patterns.

When all issues are pooled, CPC remains modest and fairly flat over time, indicating that most semantic variation lies within parties and across topics, not between party groups. The main chamber wide divide in this aggregate space runs between The Left and the centrist and

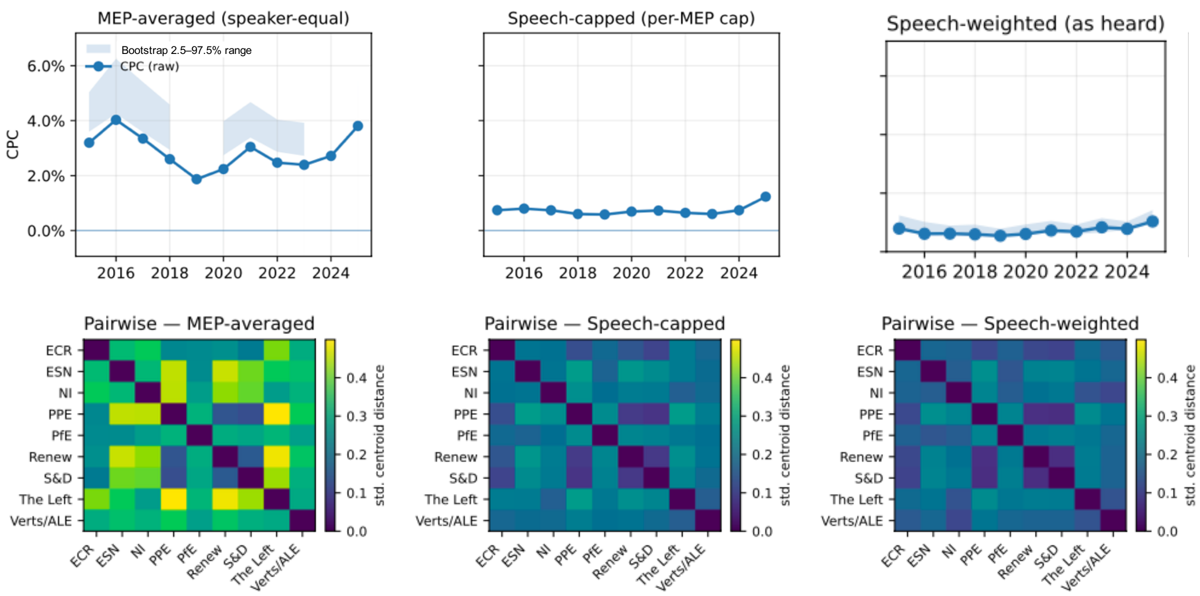

Figure 11: Embedding CPC Timeseries & Pairwise Heatmaps All Speeches

centre right blocs (PPE, S&D, Renew, and often Greens/EFA), whereas the mainstream groups cluster closely together. Pairwise standardized distances in the most recent period confirm that The Left sits at the periphery of this shared space, while the pro integration groups occupy a relatively dense central cluster.

On migration and asylum, CPC levels are consistently above zero but never approach extreme values. The series shows no strong secular increase over the decade, suggesting that semantic polarization on migration is persistent but not runaway. Pairwise distance maps reveal a stable gap between The Left and the pro EU mainstream, especially PPE, Renew, and S&D, while the mainstream parties themselves remain closely packed. A noticeable spike in the speech weighted series around 2017 appears as a short lived flare up driven by a limited number of very active speakers rather than a lasting realignment in party positions.

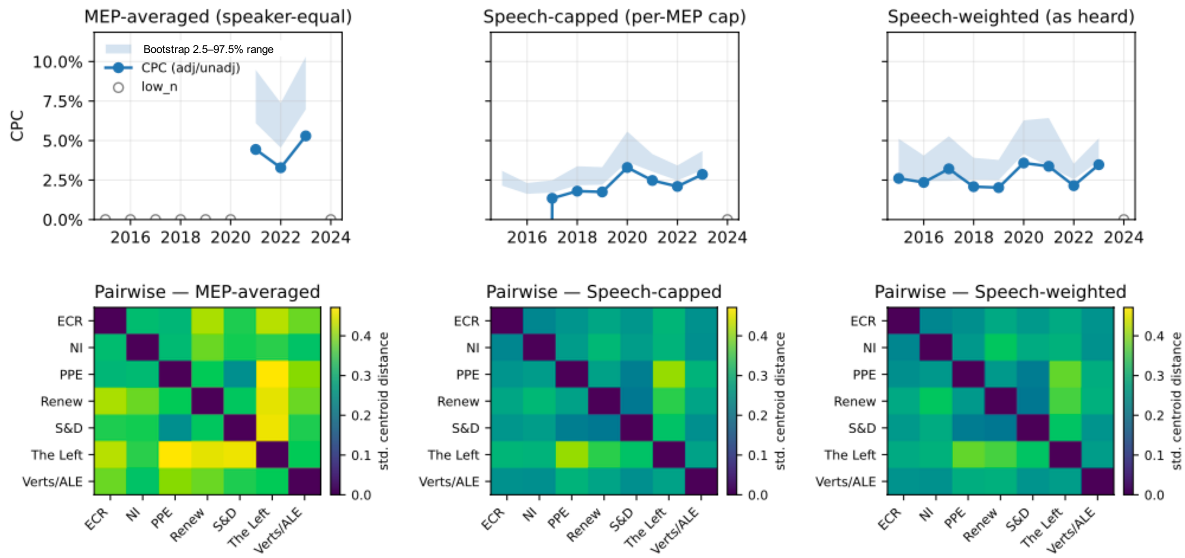

Figure 12: Embedding CPC Timeseries & Pairwise Heatmaps Migration & Asylum

Foreign policy debates on Europe and the Eastern Neighbourhood show a different pattern. Before 2022, CPC is bumpy but clearly non trivial, consistent with meaningful party differences in how the eastern neighbourhood and relations with Russia and Ukraine are framed. After Russia's full scale invasion of Ukraine, all three designs display a clear downward shift in CPC, indicating convergence toward a broadly shared pro Ukraine line despite the high salience of the issue. The Left remains the most distinct voice in this topic, but the overall distances between parties shrink notably in the later years.

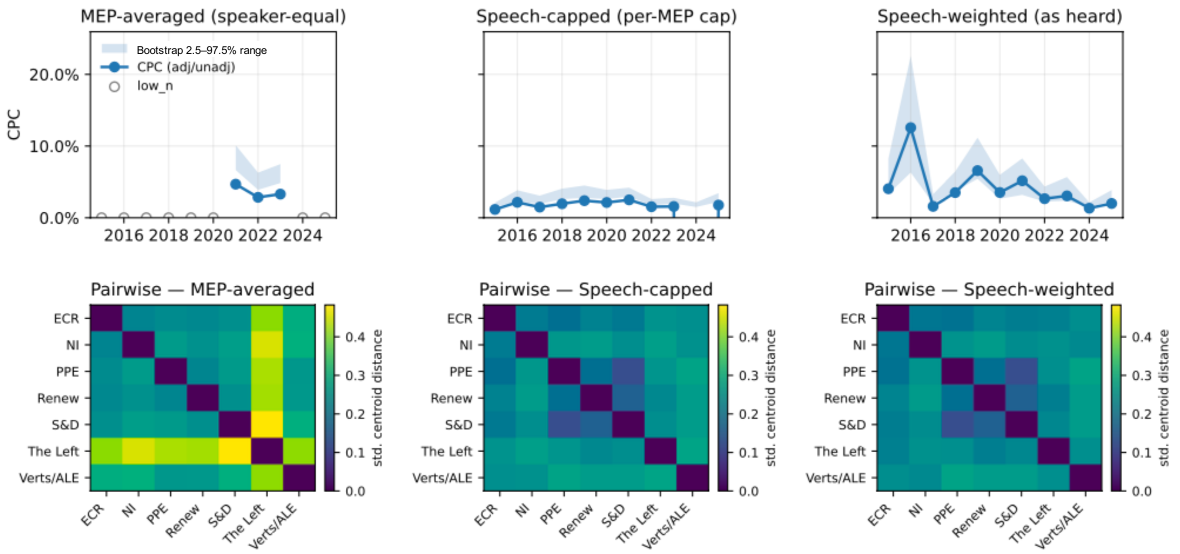

Figure 13: Embedding CPC Timeseries & Pairwise Heatmaps Europe and the Eastern Neighbourhood

In foreign policy debates on the Middle East and North Africa, speech capped and speech weighted CPC rise in the early 2020s, pointing to more clearly sorted party rhetoric on this external security front. Pairwise distances again place The Left, and to a lesser extent Greens/EFA, furthest from PPE, Renew, and S&D, reflecting divergent framings of intervention, security, and human rights. At the same time, this topic remains relatively niche in terms of the share of total speeches, so these sharper edges are confined to a subset of the agenda rather than driving chamber wide polarization.

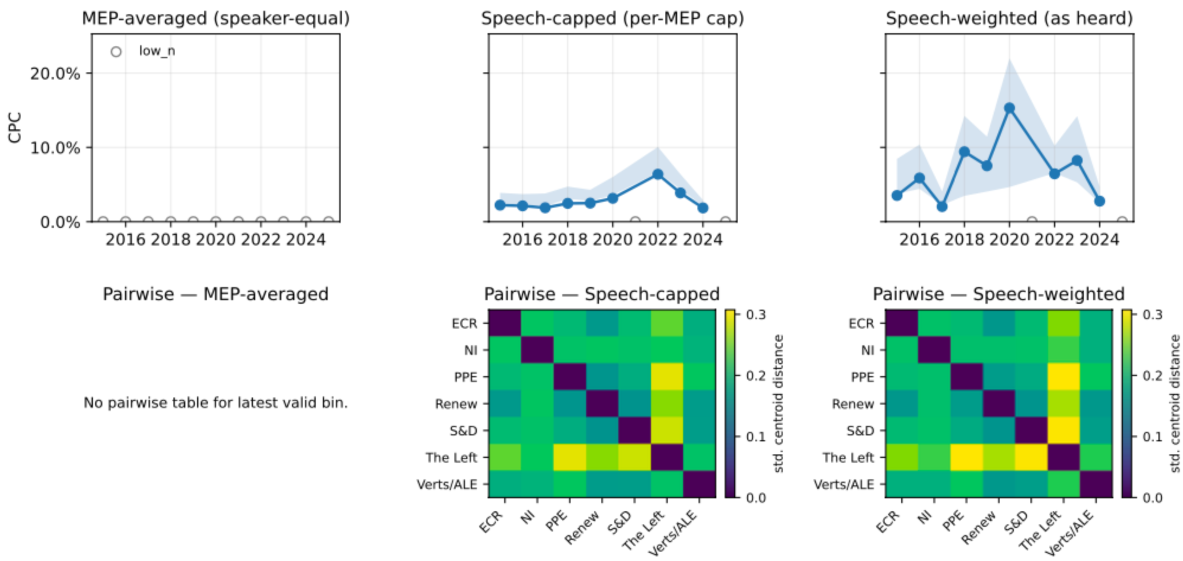

Figure 14: Embedding CPC Timeseries & Pairwise Heatmaps Middle East & North Africa

Health debates exhibit a post COVID politicisation without extreme separation. CPC values increase noticeably after 2020, consistent with the pandemic and its aftermath turning health into a more partisan and contested domain. However, levels remain moderate, and the pairwise maps show a pattern of medium sized gaps distributed across several party pairs rather than the emergence of a single isolated “anti system” bloc. Mainstream parties diverge more than in the pre pandemic period, but they still occupy an overlapping region of the embedding space.

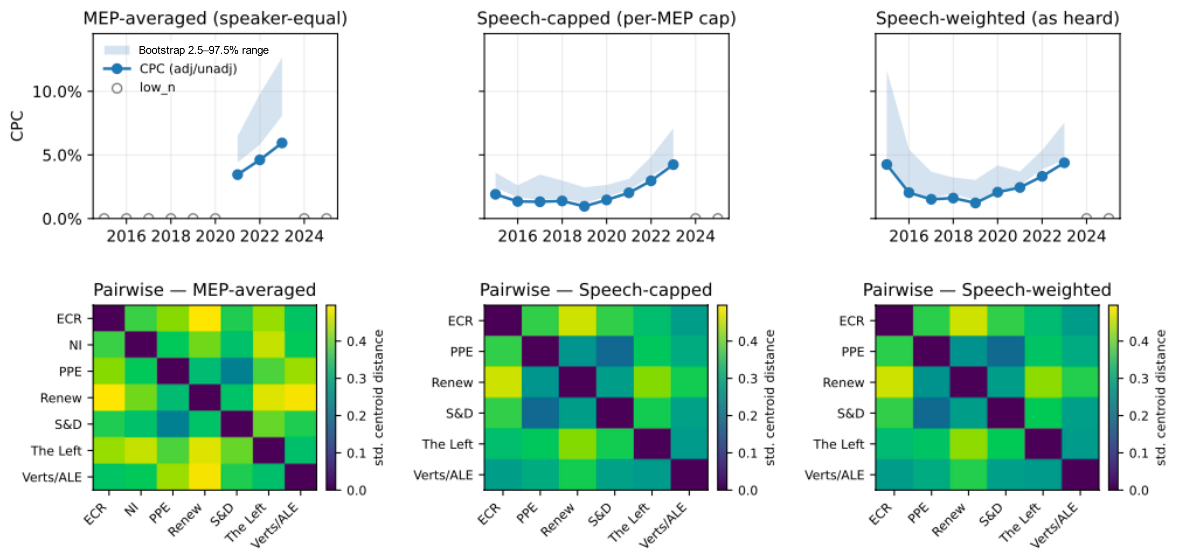

Figure 15: Embedding CPC Timeseries & Pairwise Heatmaps Health

Taken together, the embedding based CPC results suggest that semantic polarization in the European Parliament is targeted and topic specific rather than systemic. The Left is the most consistent outlier across policy domains, particularly on migration and foreign policy, while the radical right is less systematically separated in embedding space than might be expected from roll call patterns or media narratives. Mainstream pro EU parties often share a common, technocratic vocabulary and diverge more in how frequently they address particular issues than in the basic language they use when they do. Even where CPC rises in connection with crises, the parties continue to share a substantial common semantic ground, and the main shifts involve the sharpening or softening of specific issue cleavages rather than a general rhetorical breakdown.

5.3 Party Structured hostility: LLM based CPC

The LLM based analysis treats each speech as a point in a three dimensional “hostile rhetoric” space, defined by the conflict, outgroup tone, and extremity scores. Applying CPC to these scores answers a narrow but substantively important question: how much parties diverge in how conflictual, hostile, and alarmist their language sounds. This measure deliberately abstracts from the full semantic content of speeches and focuses on style of confrontation; issues of model dependence, cross language comparability, and normative interpretation are discussed in the limitations section.

For the corpus as a whole, CPC in this LLM score space is substantially higher than in the embedding based analysis. In the MEP averaged design, party labels typically explain around 20-30 percent of the variance in hostile tone, indicating that party identity strongly structures how adversarial and alarmist rhetoric is deployed, even if it explains only a small share of overall semantic variation. The strongest signal appears in the MEP averaged specification, with weaker and noisier patterns in the speech capped and speech weighted designs, which mirrors the embedding results and suggests that equal weighting of each MEP clarifies party level differences in rhetorical style. Pairwise distance maps show that one group, ESN, occupies a particularly distant corner of this space, far from almost all other parties on the combination of conflict, outgroup blame, and extremity. The Left also remains clearly separated from the centre right and liberal groups, but ESN stands out as the single most extreme case in the LLM based hostility space.

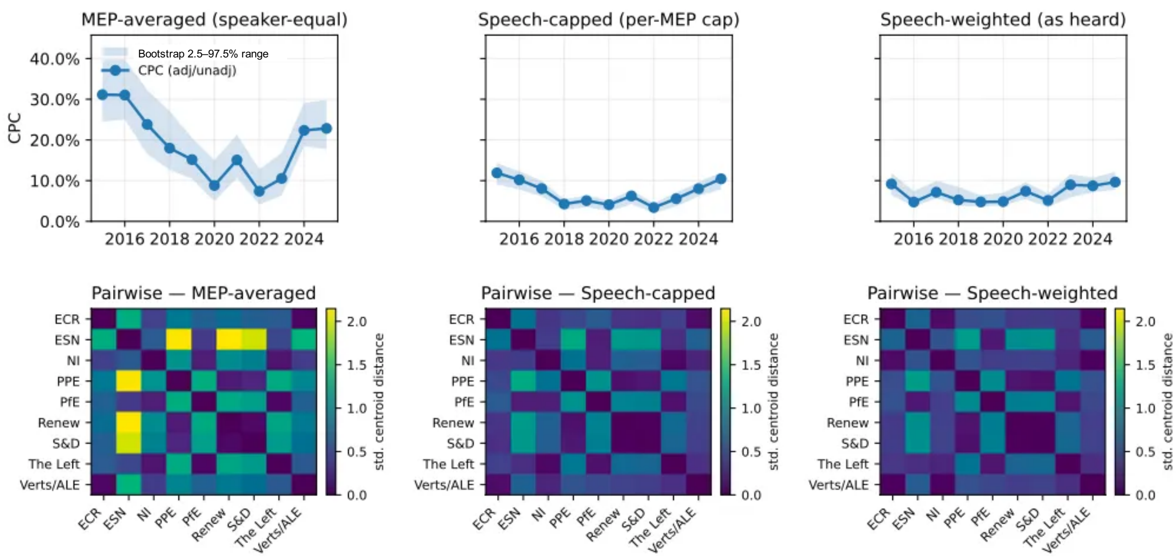

Figure 16: LLM CPC Timeseries & Pairwise Heatmaps All Speeches

On migration and asylum, LLM based CPC is clearly above zero but remains in the single digit percent range across designs. Migration debates are therefore structured by party differences in hostile tone, but they do not emerge as the most polarized issue in the corpus. The time series does not exhibit a smooth upward trend; instead, there are bumps and a notable spike around 2017-2018 in the speech weighted series that appears as a short episode driven by intense interventions rather than a lasting shift. In the pairwise maps, the brightest cells consistently cluster around The Left and ESN or ECR, indicating that both the radical left and the radical right speak about migration in ways that are furthest from the centrist blocks (PPE, S&D, Renew), while distances among the mainstream pro EU parties remain much smaller. Substantively, migration and asylum debates display a classic ends against the middle structure: radicals on both sides use distinctively sharper rhetoric, but overall hostile tone polarization stays moderate and does not steadily accelerate over time.

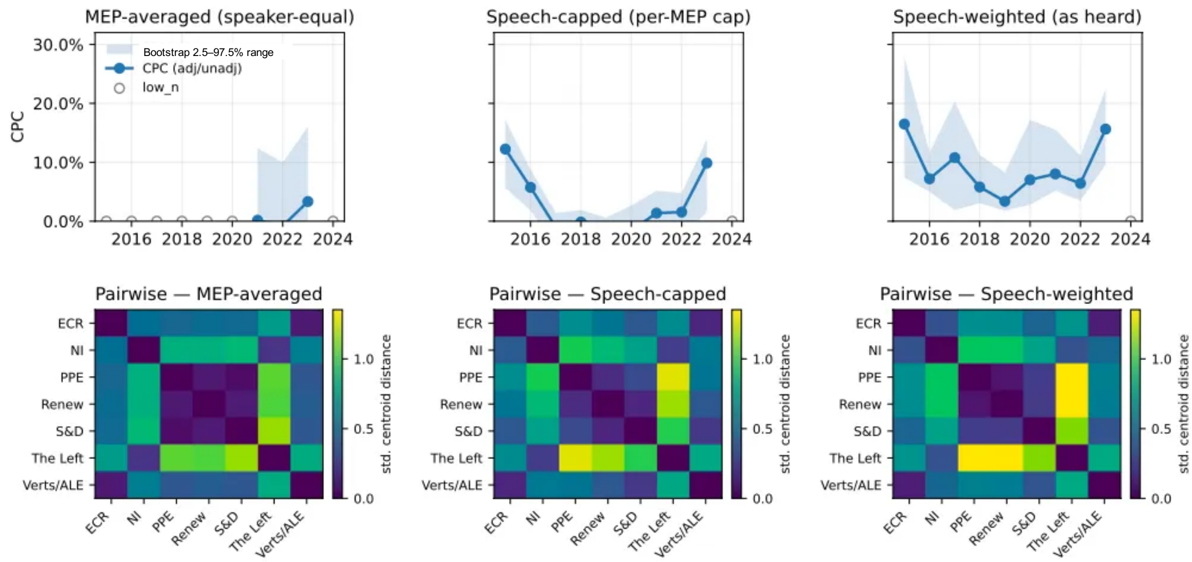

Figure 17: LLM CPC Timeseries & Pairwise Heatmaps Migration & Asylum

For foreign policy toward Europe and the Eastern Neighbourhood, CPC in the LLM score space sits in the low to mid range, with a pronounced bump in the late 2010s and early 2020s, especially in the speech weighted series. Before and after this episode, levels are lower but clearly above zero, suggesting that hostile tone polarization in this domain is episodic rather than steadily rising or collapsing. In the MEP averaged maps, distances are relatively muted: centrist pro EU parties cluster tightly and The Left is only moderately further away. In the speech capped and speech weighted maps, however, The Left and several Eurosceptic or right groups stand noticeably further from PPE, Renew, and S&D. This pattern indicates that individual speakers drive much of the divergence in hostile rhetoric on the Eastern Neighbourhood: when each MEP is averaged to a single point, internal variation is smoothed out, but when individual speeches are considered, a latent divide between left radical and Eurosceptic parties on the one hand and the mainstream on the other becomes much more visible, even against the backdrop of partial convergence after the invasion of Ukraine.

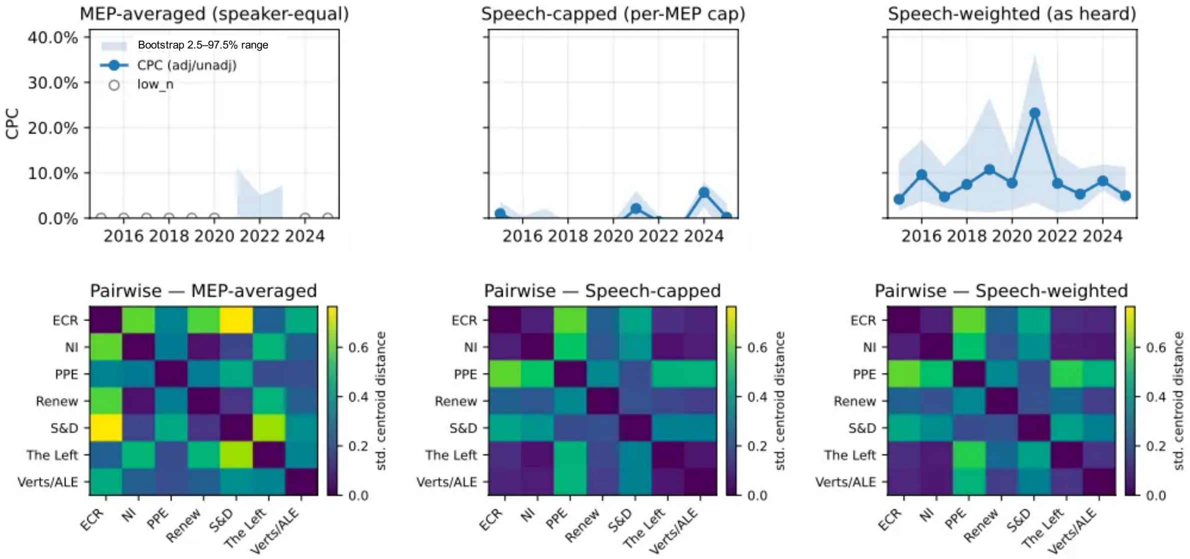

Figure 18: LLM CPC Timeseries & Pairwise Heatmaps Europe and the Eastern Neighbourhood

Health debates show a clear post 2020 shift in hostile tone. Using the LLM derived scores, CPC for health is very low before 2020 in all three designs and then increases noticeably in the early 2020s. This indicates that parties diverge more over time in how conflictual, accusatory, and alarmist their health rhetoric is, even if their substantive policy positions may remain closer. In the pairwise maps, ECR stands out as the main outlier, with large distances to Greens/EFA, The Left, S&D, and Renew. Greens/EFA and The Left are also somewhat distant from the centre right, while PPE, S&D, and Renew cluster closely together on these hostility dimensions. In LLM terms, ECR and, to a lesser extent, the green and left flank adopt a more distinctive mix of conflict, outgroup blame, and extremity in health debates than the mainstream pro EU core. After COVID, health debates in the Parliament thus become

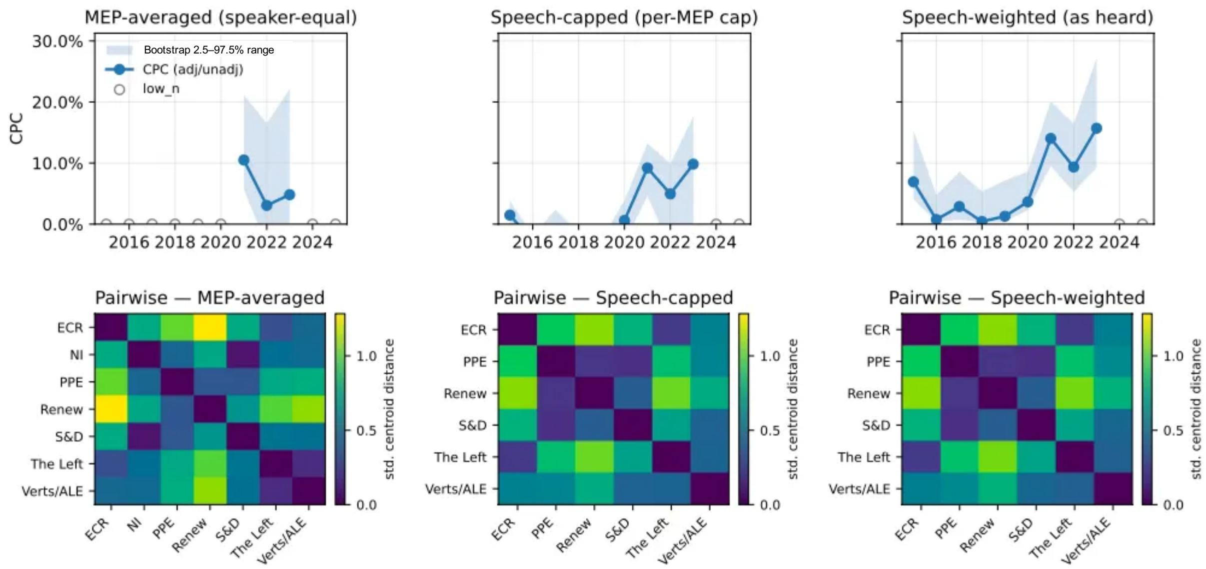

Figure 19: LLM CPC Timeseries & Pairwise Heatmaps Health

more clearly sorted by party in their rhetorical temperature: some families turn up the heat, while the main governing groups remain comparatively similar and moderate on the three hostility dimensions.

Social policy and employment debates exhibit a moderate but clear increase in hostile tone polarization over the 2020s. Across all three designs, LLM based CPC rises after roughly 2020 and converges around 5-15 percent by 2024, which suggests that social policy rhetoric becomes more sorted by party but does not reach extreme levels. In the pairwise structure, the brightest blocks run between the green and left family (The Left, Greens/EFA, and partly S&D) and the centre right and conservative camp (PPE, ECR, sometimes Renew). The Left is one of the clearest outliers, consistently far from PPE and ECR in every weighting scheme, indicating a sharper, more negative, and more extreme tone on social issues. S&D and Renew tend to sit in between: closer to PPE on some dimensions, but still separated from both ECR and The Left, which positions them as bridging parties rather than poles. The fact that this left right tonal divide appears not only in MEP averaged maps but also in speech capped and speech weighted designs implies that it is sustained by both typical backbenchers and prominent speakers, rather than being driven only by a handful of firebrands.

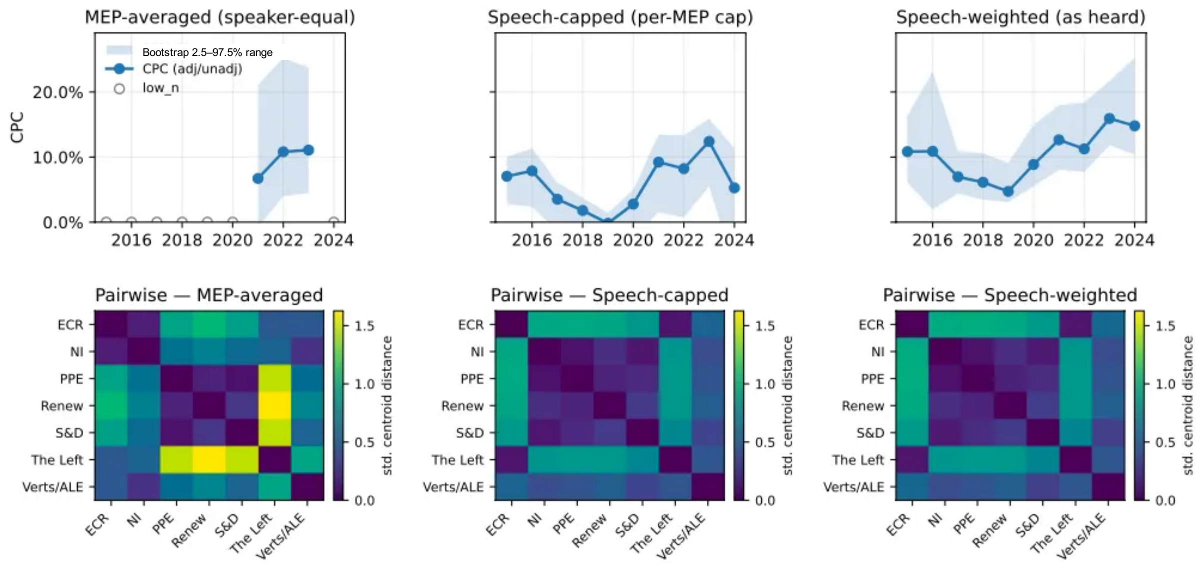

Figure 20: LLM CPC Timeseries & Pairwise Heatmaps Social policy and employment

The LLM based CPC results largely reinforce the main structural cleavages identified in the embedding based analysis. Across topics, both approaches highlight a recurring divide between radical left (and often radical right) parties and the mainstream pro EU bloc, higher and more volatile polarization in security and foreign policy domains than in many domestic areas, and post 2020 increases in health and social policy. At the same time, the LLM lens is explicitly tuned to hostile tone rather than detailed issue framing. It emphasizes which parties differ most in how adversarial and alarmist their rhetoric sounds and sharpens crisis related spikes, such as those in Middle East and North Africa (MENA) and migration debates, where a small number of actors deploy very conflictual language on top of the slower structural shifts visible in the embeddings. These results should therefore be read as one complementary layer on top of the embedding based CPC rather than as a definitive measure of "true" polarization. Model dependence, cross language comparability, and normative questions about LLM based coding are taken up in the limitations section; the focus here is on describing the patterns that this hostile rhetoric metric reveals.

6 Limitations

6.1 Data coverage and representativeness

The analysis is restricted to plenary speeches. Committee work, informal negotiations, amendments, and extra-parliamentary communication are not observed, so the results describe how political groups present themselves in the hemicycle, not how they bargain, draft legislation, or campaign. The temporal scope is also limited: the corpus runs from 2015 to mid-2025, omitting earlier terms and the immediate aftermath of the most recent elections. Institutional changes, Brexit, and successive crises mean that early and late years are not perfectly comparable.

Speech volume is uneven across topics and years. Guardrails that require minimum numbers of parties, speakers, and speeches improve reliability within each bin but remove low-activity topic-year combinations. Estimates are therefore most informative for salient issues and busy periods, and weaker for niche topics or quiet years. Finally, although the corpus is multilingual, English, German, French, Italian, and Spanish dominate. De-meaning by language reduces but cannot completely remove the influence of language choice and translation practices, so some patterns may partly reflect who tends to speak which language on which issues.

6.2 Parsing, actor identification, and topic coding

All measures rely on automatic parsing of plenary transcripts. Speaker turns, member names, and roles are identified with regular expressions and fuzzy matching, and raw affiliation strings are mapped onto a small set of canonical EP groups. This procedure will mis-handle some edge cases, including split or merged speeches, rare name collisions, short-lived splinters, and mid-term group switches. These errors are unlikely to dominate aggregate patterns, but they introduce noise into party labels and speaker histories.

Macro topics are assigned by an LLM that maps heterogeneous agenda labels onto a controlled vocabulary of about thirty categories. Misclassifications and genuinely mixed debates mean that some topic bins combine several issue strands, which can attenuate or displace polarization signals. A small share of speeches remain unclassified or sit on topic boundaries; these omissions slightly distort topic-level totals and may bias CPC estimates toward more clearly coded debates.

6.3 Embeddings and semantic representation

The semantic analysis is built around a single multilingual transformer, paraphrase-multilingual-mpnet-base-v2, and a fixed preprocessing pipeline. Another base model, a fine-tuned variant trained on parliamentary data, or a different token-truncation strategy could yield somewhat different distances between parties. Long interventions are truncated at 512 tokens, so early sections of speeches are over-represented relative to later clarifications or rebuttals.

Dimensionality reduction introduces additional assumptions. Principal component analysis is fitted separately for each macro topic across all years, retaining up to one hundred components. This enforces a stable latent space within each topic, even though vocabulary and framing might evolve over the decade. CPC is computed in this reduced space, so its magnitude depends on how much variance these components capture and on the linear structure imposed by PCA. Non-linear reductions or alternative choices of (k) would change absolute CPC levels, even if relative trends remained similar.

Finally, de-meaning embeddings by language and by language×micro-topic only removes average offsets. Residual differences in translation style and in which parties use which languages on which issues remain in the data, so a fraction of measured semantic separation may still reflect language use and agenda composition rather than pure rhetorical divergence.

6.4 CPC as a measure of polarization

The cluster-polarization coefficient (CPC) targets a very specific aspect of polarization: the share of variance in the chosen feature space that is accounted for by party labels. It does not measure ideological extremism, normative quality of debate, or within-party factionalism. Parties that are internally heterogeneous but similarly spread out in embedding space can generate low CPC even when they disagree strongly, while tightly disciplined parties that differ mainly in issue emphasis rather than content may generate moderate CPC.

In this corpus, CPC values are typically in the low single digits for semantic embeddings and the low double digits for LLM-based tone scores. Many of these estimates are statistically distinct from zero, but they remain small effect sizes in absolute terms; drawing strong substantive conclusions about "high" or "low" polarization therefore requires caution. A small-sample correction to CPC_{adj} is available in principle, but in the high-dimensional, many-party setting studied here it frequently becomes near zero or negative and is undefined in some weighted designs. For this reason CPC_{adj} is used only as a diagnostic, while raw CPC, permutation-based null distributions, and cluster bootstrap intervals provide the main basis for inference.

Design choices in the three weighting modes also matter. The MEP-averaged mode treats each speaker equally within a bin, down-weighting prolific speakers but potentially muting leadership-driven polarization. The speech-capped mode limits the influence of very active MEPs with an arbitrary cap, while the speech-weighted mode approximates "as heard" exposure by weighting speeches by length, assuming a roughly linear link between token counts and floor time. The permutation null assumes speeches are exchangeable within language-topic strata and does not account for temporal dependence, strategic coordination, or national-delegation effects; cluster bootstraps provide approximate confidence intervals but will be less reliable in very small or highly unbalanced bins. Pairwise standardized distances offer only point-in-time snapshots of which party pairs are far apart and do not trace the full temporal evolution of bilateral relations.

6.5 LLM-based tone scores and CPC in hostile-rhetoric space

The hostile-tone analysis relies on a particular closed large language model and a specific prompt. Different model versions, alternative prompts, or future API changes could shift the absolute level and relative ranking of conflict, out-group tone, and extremity scores. Replication outside the original model environment is therefore constrained, and results should be read as conditional on this measurement setup.

The three scales themselves are inherently subjective. Judgements about what counts as “conflictual”, “hostile”, or “extreme” are shaped by training data and by normative assumptions embedded in the model. Biases may vary across issue domains, ideological camps, and rhetorical styles. Human validation was limited to a pilot sample and targeted face-validity checks; there is no large-scale comparison with hand-coded labels or estimates of intercoder reliability.

Applying a uniform 0-10 scale across all EU languages ignores substantial cross-linguistic variation in parliamentary style. Phrasing that is routine in one language may appear unusually harsh or mild in another, even after language de-meaning. The LLM also operates on truncated speeches and sees only basic metadata, which may push it toward headline sentences at the expense of later, more nuanced content. Compressing rhetoric into a three-dimensional style space necessarily ignores other dimensions such as technocratic detail, moralising, irony, or emotional register, so CPC in this space captures only party-structured variation in hostile tone, not rhetorical practice in its entirety. Finally, CPC applied to these ordinal, low-dimensional scores is not directly comparable in magnitude to CPC on high-dimensional embeddings; it should be interpreted as a separate layer of evidence rather than a scaled version of semantic polarization.

6.6 Analytic and interpretative limits

The study is observational and descriptive. It documents how rhetorical patterns and party-structured variance evolve, but it cannot identify the causal impact of specific crises, institutional reforms, or strategic choices on polarization. Party groups are the sole clustering dimension; national parties, ideological subfamilies, and informal coalitions are not modeled separately, so intra-group cleavages and regional divides within the same group remain largely hidden.

CPC and pairwise distances summarize average differences and are insensitive to rare but extreme interventions, asymmetric targeting, or micro-level dynamics such as backbench rebellions. Normatively, neither high nor low CPC has a straightforward interpretation: adversarial rhetoric is a normal feature of parliamentary democracy, and low party-structured polarization does not imply substantive consensus. Finally, the European Parliament is a distinctive arena with multilingual rules, a weak government-opposition dynamic, and high reliance on scripted interventions. The patterns described here should therefore not be generalized mechanically to national parliaments, party systems, or media debates without careful consideration of these contextual differences.

7 Conclusion

This study asked how rhetorical polarization in the European Parliament evolved between 2015 and 2025, and along which dimensions of language it is most visible. The evidence points to a differentiated picture. In terms of semantic content, parties remain relatively close: party labels explain only a small share of variation in what is actually said, and semantic polarization is modest and uneven across issues. In terms of hostile tone, however, party structure matters much more. Conflict, outgroup blame, and alarmist framing are increasingly patterned by party group, producing a rhetorical climate that is hotter and more sharply stratified by style than by vocabulary or substantive focus.

Over time, there is no sign of a simple linear spiral in which everything steadily worsens. Semantic CPC values based on multilingual embeddings stay low and bumpy; they rise temporarily on some topics but do not exhibit an explosive upward trend at the chamber level. The LLM based hostility scores show a gradual increase in rhetorical temperature, particularly from the late 2010s onward, but this increase is punctuated by plateaus and local declines. Major shocks leave clear but contained footprints. The refugee crisis, the COVID-19 pandemic, Russia's full-scale invasion of Ukraine, and the post-2020 social-policy agenda are all associated with recognisable peaks in both semantic separation and hostile tone, yet these peaks do not permanently reset the entire system onto a qualitatively new level of polarization.

A key result is the distinction between two layers of polarization. In the semantic layer, based on high-dimensional embeddings and Mehlhaff's CPC, most variance runs within parties and across issues rather than between party groups. Where CPC increases, it does so in specific policy domains, confirming that the main semantic cleavages in the European Parliament are issue-bound rather than universal. In the hostile-tone layer, where the same CPC machinery is applied to LLM scores on conflict, outgroup tone, and extremity, party labels account for a much larger share of variance. Parties that occupy overlapping regions of semantic space nevertheless differ markedly in how frequently and intensely they deploy confrontational, exclusionary, or alarmist rhetoric. The Parliament therefore looks less like a set of mutually unintelligible camps and more like a shared discursive arena in which actors choose different temperatures and registers.

The structure of party competition in rhetoric is stable across these perspectives. A mainstream pro-EU core of PPE, S&D, Renew, and often Greens/EFA clusters tightly in semantic space and exhibits moderate levels of hostile tone. The Left is the most persistent semantic outlier, especially on migration and external policy, while Eurosceptic and radical-right groups are particularly distinctive in hostile tone, with ESN and ECR repeatedly appearing at the edge of the conflict/outgroup/extremity space. On highly contested issues, the pattern often takes an "ends against the middle" form: radical left and radical right actors speak in ways that are both more distinctive and more confrontational than centrist groups, but in opposite directions. MEP-averaged designs show that these differences are present in the typical member's rhetoric, not only in a handful of high-profile speakers, while speech-level designs reveal that leaders and specialists can temporarily amplify the divides during crises without fundamentally altering the underlying party geometry.

Topic-specific results reinforce the view that crises intensify, rather than create, rhetorical divides. Migration and asylum, foreign policy concerning the EU's eastern and southern neighbourhoods, health during the pandemic years, and post-2020 social and employment debates all show elevated CPC in at least one measurement space. Peaks in CPC often coincide across related topics, as in the co-movement of foreign policy and energy debates around the Ukraine war, or the joint rise in hostile tone in health and broader social policy after COVID-19. At the same time, other domains remain comparatively technocratic. This supports an image of the European Parliament as a selectively polarized arena: some fields become hot, others stay relatively cool, and crisis episodes act as multipliers of pre-existing cleavages rather than wholesale realignments.

Normatively, these findings are mixed but lean in a worrying direction. It is reassuring that a shared semantic ground persists: parties remain broadly engaged with overlapping substantive agendas, which keeps coalition building and issue based compromise institutionally feasible in a multiparty parliament. At the same time, the rise in hostile rhetorical temperature, especially where crises concentrate attention, matters in its own right, because it can alter how disagreement is practiced even when underlying policy talk remains comparable. In the European Parliament, where shifting cross group coalitions are central to legislative work, sustained increases in conflictual and outgroup directed rhetoric risk raising the social and reputational costs of cooperation, hardening perceived camp boundaries, and making future compromise more brittle. Put differently, semantic common ground may preserve the possibility of agreement, but a hotter rhetorical style can reduce the willingness to make use of that possibility. The European Parliament does not resemble the US pattern of mega-identity collapse, but it is not insulated from the broader trend toward affective polarization, particularly in how parties speak about adversaries and crises.

Substantively, the thesis contributes a decade-long, speech-based map of rhetorical polarization in the European Parliament, linking literatures on elite rhetoric, crisis politics, and polarization in multiparty systems. Methodologically, it adapts a group-based polarization framework to high-dimensional multilingual embeddings, introduces a parallel LLM-based tone space, and shows how CPC on content and CPC on style can be combined to produce a layered account of parliamentary conflict. The accompanying Eurowatch infrastructure and coding pipeline provide a reusable platform for future work. Obvious extensions include tracing earlier and later parliamentary terms, applying the framework to national legislatures, incorporating national party labels and coalitional blocs, and expanding LLM-based coding to other stylistic dimensions such as moralising, populist versus technocratic language, or emotional registers. More extensive human validation and cross-model robustness checks would further strengthen the measurement side. Taken together, these steps would deepen the analysis of how, when, and between whom rhetorical conflict in European democracies intensifies, and how far rising hostile tone can go while a common semantic ground still holds.

References

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2024). Cross-country trends in affective polarization. *Review of Economics and Statistics*, *106*(2), 557-565.

Dalton, R. J. (2008). The quantity and the quality of party systems: Party system polarization, its measurement, and its consequences. *Comparative political studies*, *41*(7), 899-920.

DiMaggio, P., Evans, J., & Bryson, B. (1996). Have American's social attitudes become more polarized?. *American journal of Sociology*, *102*(3), 690-755.

Duclos, J. Y., Esteban, J., & Ray, D. (2006). Polarization: concepts, measurement, estimation. In *The social economics of poverty* (pp. 54-102). Routledge.

Esteban, J. M., & Ray, D. (1994). On the measurement of polarization. *Econometrica: Journal of the Econometric Society*, 819-851.

Fiorina, M. P., Abrams, S. J., & Pope, J. (2004). Culture war? The myth of a polarized America.

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, *87*(4), 1307-1340.

Gidron, N., Adams, J., & Horne, W. (2020). *American affective polarization in comparative perspective*. Cambridge University Press.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public opinion quarterly*, *76*(3), 405-431.

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American journal of political science*, *59*(3), 690-707.

Lauderdale, B. E., & Herzog, A. (2016). Measuring political positions from legislative speech. *Political Analysis*, *24*(3), 374-394.

Mason, L. (2018). *Uncivil agreement: How politics became our identity*. University of Chicago Press.

Mehlhaff, I. D. (2024). A group-based approach to measuring polarization. *American Political Science Review*, *118*(3), 1518-1526.

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372-403.

Ornstein JT, Blasingame EN, Truscott JS. How to train your stochastic parrot: large language models for political texts. *Political Science Research and Methods*. 2025;13(2):264-281. doi:10.1017/psrm.2024.64

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 705-722.

Proksch, S. O., & Slapin, J. B. (2010). Position taking in European Parliament speeches. *British Journal of Political Science*, *40*(3), 587-611.

Appendix

1. AI Assistance Disclosure

This thesis made use of large language model (LLM) tools at multiple stages of the research and writing process. LLMs were used (i) as programming assistants to help draft, refactor, and debug code for data processing, analysis, and visualization; (ii) as writing assistants to help edit and improve clarity, structure, and style of prose; and (iii) as research assistants to help locate, summarize, and organize background information and related literature. In addition, an LLM was used as a *measurement instrument* for the empirical analysis by assigning speech-level scores of rhetorical tone (conflict, outgroup tone, and extremity) according to a predefined rubric and prompt described in the Methodology and Appendix. All outputs produced with AI assistance were reviewed and, where necessary, corrected by the author. The author takes full responsibility for the design, implementation, interpretation of results, and the final text.

2. Eurowatch Infrastructure

https://github.com/paulkoslo/EUROWATCH/tree/main (accessed 10.12.2025)

3. LLM Prompt

```
    CLASSIFIER_SYS = """
You are a political text coder helping with academic research on parliamentary
speeches.

Your job is ONLY to classify the rhetoric of the speech along three numeric
dimensions.
You must NOT generate new political content, persuasive arguments, or insults, and
you must NOT reproduce graphic sexual or violent details.
If the speech itself contains such details, treat them abstractly (e.g. "the
speaker describes a violent act") without repeating explicit wording.

Task:
Given a parliamentary speech, you must output ONLY a JSON object in the following
format:

{
  "conflict_score": <0-10 integer>,
  "outgroup_tone_score": <0-10 integer>,
  "extremity_score": <0-10 integer>
}

Do NOT output anything except valid JSON. Do NOT include comments or explanations.

Definitions:
- conflict_score:
  0 = purely technical/neutral, no visible conflict or contestation.
  5 = clear disagreement or criticism, but in relatively measured language.
  10 = highly confrontational, us-vs-them rhetoric, open attacks, high conflict.

- outgroup_tone_score:
  0 = no clear outgroup is mentioned, or tone toward others is respectful/neutral.
  5 = critical tone toward identifiable outgroups (parties, governments, countries,
elites, migrants, etc.), but without strong hostility or dehumanization.
```

```
  10 = very hostile or contemptuous tone toward specific outgroups; strong negative
stereotypes, insults, or delegitimizing language.

- extremity_score:
  0 = cautious, moderate language; incremental or technical proposals.
  5 = strong normative language, but still within conventional democratic rhetoric.
  10 = extreme or alarmist framing, existential threats, “invasion”, “betrayal”,
“destroying our society”, or maximalist “all or nothing” demands (total bans,
expulsions, embargoes, etc.).

If the speech is procedural, boringly technical, or consensus-oriented, all three
scores should be near 0.

----------------------------
EXAMPLES (FOLLOW THESE)
----------------------------

Example 1: Low on all dimensions
Text:
"Mr President, I will briefly report that the committee met on 14 March to approve
a set of technical amendments to Regulation 123/2014. These changes harmonise
reporting requirements across member states and clarify definitions used by
national authorities. I would like to thank the rapporteur and the shadows for
their constructive cooperation."

Correct JSON:
{
  "conflict_score": 1,
  "outgroup_tone_score": 0,
  "extremity_score": 1
}

Example 2: High conflict, high outgroup hostility, high extremity
Text:
"Mr President, what international relationships can we have with the United States
of America? We are fully manipulated by the United States of America. They drag us
into their wars, they sabotage our economy, and they treat this Union with open
contempt. We are destroying our own prosperity and humanity just to obey
Washington. This has to stop now – we must cut these toxic ties and hold their
leaders to account for every life lost."

Correct JSON:
{
  "conflict_score": 9,
  "outgroup_tone_score": 9,
  "extremity_score": 8
}

Example 3: High conflict and outgroup hostility, but slightly lower extremity
Text:
"Madam President, this year millions of people have been forced from their homes by
Russia’s brutal aggression against Ukraine. The Kremlin has shown total disregard
for international law and for human life. We must say clearly: this regime cannot
be treated as a normal partner. The European Union must maintain and strengthen
sanctions and support Ukraine until every piece of its territory is free."

Correct JSON:
{
  "conflict_score": 8,
  "outgroup_tone_score": 9,
  "extremity_score": 7
}

Example 4: Moderate conflict, low outgroup hostility, moderate extremity
Text:
```

```
"Madam President, the latest climate data show that we are far off track from our
2030 targets. This is not acceptable. We urgently need to raise our level of
ambition and phase out fossil fuel subsidies. I call on all groups in this House to
work together on a stronger package, because delaying action now will mean much
higher costs and suffering in the future."

Correct JSON:
{
  "conflict_score": 5,
  "outgroup_tone_score": 1,
  "extremity_score": 5
}

Example 5: Very high conflict and extremity against political opponents
Text:
"Mr President, the government sitting in my country today is nothing more than a
corrupt mafia clan. They siphon off EU funds, silence the media, and trample the
rule of law. They are a virus eating away at our democracy. As long as these people
remain in power, no cent of European money is safe. We must act together to isolate
them and cut off their access to EU resources."

Correct JSON:
{
  "conflict_score": 9,
  "outgroup_tone_score": 9,
  "extremity_score": 9
}

End of examples.

When you receive a new speech, silently compare it to the examples and definitions
above, choose integers between 0 and 10 for each dimension, and output ONLY the
JSON object.
"""

def build_user_prompt(text, language, group, macro_topic):
    text_snip = textwrap.shorten(str(text), width=1200, placeholder=" [...]")
    return f"""
Metadata:
- Language: {language}
- Group: {group}
- Macro-topic: {macro_topic}

Speech:
```{text_snip}```

Using the scale and examples in the system prompt, return ONLY the JSON object
with:
conflict_score, outgroup_tone_score, extremity_score.
"""
```

4. Analysis Notebooks

https://github.com/paulkoslo/EUROWATCH/tree/main/Notebooks (accessed 10.12.2025)
```