

Article

When Algorithms Guard Democracy: Measuring Authoritarian Rhetorical Behaviour in Political Speech

Óscar Delgado-Mohatar ^{1,*} and Raúl Alelú-Paz ^{2,3,4,5,*} ¹ School of Engineering, Universidad Autónoma de Madrid, 28049 Madrid, Spain² Biological Research Laboratory Professor Giacomo Rizzolatti, Madrid Scientific Park, 28049 Madrid, Spain³ Department of Psychology, Universidad Francisco de Vitoria, 28223 Madrid, Spain⁴ Department of Biology and Geology, Physics and Inorganic Chemistry, School of Experimental Sciences and Technology, Rey Juan Carlos University, 28933 Móstoles, Madrid, Spain⁵ IRyCIS, Hospital Universitario Ramón y Cajal, 28034 Madrid, Spain

* Correspondence: oscar.delgado@uam.es (Ó.D.-M.); raul.alelu@ufv.es (R.A.-P.)

Abstract

Democratic erosion often begins rhetorically before institutions show visible damage. Here we test whether large language models (LLMs) can detect early linguistic signals of authoritarian drift in political speech. Formal speeches by Adolf Hitler (1922–1939), Donald Trump (2017–2025), Nicola Sturgeon (2014–2023), Giorgia Meloni (2022–2025) and Viktor Orban (2022–2025) were scored using an 11-indicator taxonomy derived from the Levitsky–Ziblatt framework and evaluated independently by GPT-4o, Gemini 2.5-Pro and Grok-4-Fast, with near-perfect inter-model agreement. Principal Component Analysis revealed two poles: an authoritarian–populist cluster (Hitler–Trump–Orban) and a democratic–institutional pole (Meloni–Sturgeon). To quantify proximity to an authoritarian reference, we introduce the Authoritarian Reference Index (ARI), defined such that it captures both its alignment and intensity relative to the Hitler gold-standard vector. Trump exhibited the highest proximity to the reference (99.1% alignment, 80.7% intensity), followed by Orban, who mirrored the structural alignment (97.6%) with a moderated intensity (72.4%). In contrast, the democratic–institutional pole was distinguished by significantly lower intensity scores, with Meloni (16.4%) and Sturgeon (22.3%) remaining distant from the authoritarian magnitude despite varying degrees of structural overlap. These results show that extreme rhetorical peaks carry disproportionate diagnostic weight and that LLMs can expose structural authoritarian patterns relevant for democratic monitoring.

Keywords: authoritarian political behaviour; democratic norm erosion; rhetorical framing and social cognition; discourse-based behavioural diagnostics; large language models for behavioural measurement



Academic Editor: Paul W. Posner

Received: 17 March 2026

Revised: 3 June 2026

Accepted: 3 June 2026

Published: 8 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Democratic erosion rarely begins with legal rupture. It begins rhetorically—speech by speech, slogan by slogan, metaphor by metaphor—until citizens no longer recognise the slide. Contemporary political actors have learned that language does more than convey power; it performs it. When rhetoric reframes opponents as enemies, diversity as decay, or institutions as obstacles, discourse becomes a behavioural mechanism capable of shifting collective cognition, moral thresholds, and what counts as legitimate political action. Recent comparative work shows that linguistic normalisation of exclusion or coercion often precedes institutional breakdown and makes rule-breaking cognitively

permissible before it becomes legally feasible (Hinterleitner and Sager 2023). In this sense, authoritarian drift is not merely political; it is rhetorical behaviour that actively reconfigures perceptions of threat, legitimacy, and authority. The societal receptiveness to this rhetorical reconfiguration is frequently driven by perceived cultural and demographic displacement. As recent sociological analyses highlight (Reiter 2025), the global appeal of contemporary authoritarian politics—exemplified by leaders across the US, Europe, and Latin America—is deeply rooted in the ‘status anxieties’ of historically dominant groups. In this context, authoritarian rhetoric functions not merely as a top-down policy tool, but as a psychological mechanism designed to restore a perceived loss of dignity, honour, and social hierarchy among the electorate, often overriding rational-economic voting behaviours.

1.1. Theoretical Perspectives on Democratic Decline and Erosion

Political scientists have long mapped democratic backsliding, yet language remains a blind spot: the precise vehicle through which rule-breaking becomes thinkable. As generative artificial intelligence (AI) gains the capacity to read and reason across large corpora of speech, a new methodological possibility emerges—using automated language understanding to detect early warning signals of normative decline. Here we focus on rhetoric not as “text” but as behavioural expression and ask whether authoritarian persuasion exhibits a stable statistical signature across contexts.

To answer this question, measurement requires theory. The framework proposed by Levitsky and Ziblatt (2018) identifies four diagnostic dimensions that define authoritarian behaviour—rejection of democratic rules, denial of opponents’ legitimacy, tolerance of violence, and willingness to restrict civil liberties. Consistent with recent operationalisations in comparative politics (Hobolt and Osnaßrügge 2025), we expand these four domains into an 11-indicator taxonomy suitable for computational scoring of natural political speech.

Historical and contemporary cases illustrate the relevance of this behavioural model: from Bolsonaro’s discourse in Brazil (Othon 2021) to Trump’s polarising rhetoric in the United States (Kadim 2022; Alhawamdeh and Al-Khanji 2025), leaders who personalise power tend to mirror one another linguistically, while democratic speech typically maintains procedural framing and inclusive norm signals (Rosenfeld 2019). Meanwhile, large-scale natural language analyses (Card et al. 2022) show increasing moral polarisation in political communication, and state-of-the-art large language models (LLMs) now achieve human-level reliability in text-classification tasks across languages and domains (Suter and Meckel 2024; López et al. 2024; Rathje et al. 2024). Theory defines what counts as authoritarian behaviour; LLMs allow us to measure it.

1.2. The Political Psychology and Communication of Authoritarian Rhetoric

In contemporary political marketing and communication, authoritarian persuasion is increasingly understood not merely as ideological messaging, but as a strategic communicative design aimed at exploiting cognitive biases and emotional heuristics. From a political psychology perspective, authoritarian rhetoric leverages affective polarisation, out-group threat perception, and strategic incivility to bypass rational deliberation and activate identity-based loyalties.

Recent computational advancements have significantly enhanced our ability to detect authoritarian footprints in political discourse. Previous research has demonstrated that authoritarian rhetoric can be quantified either through dictionary-based analyses of democratic simulation (Maerz 2019) or by training deep-learning models to identify inductive semantic patterns based on regime type (Mochtak 2025). However, measuring authoritarianism through textual frequencies or aggregate corpus scores faces a critical challenge: the semantic camouflage of autocratic actors. As Maerz (2019) observes, hegemonic au-

thoritarian regimes strategically employ a highly democratic style of language to feign pluralism and legitimise their rule. Similarly, Mochtak (2025) demonstrates through weakly supervised models that democratic leaders frequently engage in authoritarian narratives, just as authoritarian leaders easily adopt diplomatic, democratic rhetoric depending on their audience.

Consequently, approaches relying on central linguistic tendencies or inductive regime-labelling may inadvertently conflate procedural democratic language with genuine institutional commitment. To address this, our study introduces a deductive, behavioural diagnostic approach grounded in the normative framework of Levitsky and Ziblatt. By shifting the computational focus from average word frequencies or inductive discourse profiling to the detection of extreme rhetorical peaks, we propose that the true signal of authoritarian drift lies not in the baseline discourse, but in the sporadic, high-intensity normalisation of democratic norm violations.

1.3. Research Questions, Hypotheses, and Thesis

The core thesis of this paper is that authoritarian democratic erosion leaves a mathematically measurable rhetorical footprint, identifiable not by its average tone, but by the structural alignment and intensity of its extreme persuasion peaks. To demonstrate this, we formalise the following research question (RQ) and hypotheses (Hs):

- **RQ1:** can large language models (LLMs) reliably detect and differentiate authoritarian rhetorical behaviour from democratic-institutional speech using a deductive theoretical framework?

H1 (structural alignment). *authoritarian and populist leaders will exhibit a high directional structural alignment (cosine similarity) with the historical authoritarian gold-standard, sharing a common persuasive geometry.*

H2 (intensity threshold). *the transition from latent institutional populism to active authoritarian drift is mathematically defined not by structural alignment alone, but by the intensity magnitude (Euclidean distance) of norm-violating rhetorical peaks.*

Here we test these hypotheses by applying the 11-indicator taxonomy to formal speeches by Adolf Hitler (1922–1939), Donald Trump (2017–2025), Nicola Sturgeon (2014–2023), Giorgia Meloni (2022–2025) and Viktor Orban (2022–2025).

2. Materials and Methods

2.1. Corpus and Inclusion Criteria

The study analysed formal, publicly delivered political speeches from five leaders spanning historical and contemporary democratic contexts: Adolf Hitler (1922–1939), Donald Trump (2017–2025), Nicola Sturgeon (2014–2023), Giorgia Meloni (2022–2025), and Viktor Orban (2022–2025). All speeches met four inclusion conditions: (i) they were official or campaign-stage public addresses, (ii) full transcripts were publicly archived prior to analysis, (iii) the speech context corresponded to formal political communication intended for mass audiences, and (iv) the content was non-interactive (i.e., not elicited through interviews, private recordings, or direct participant engagement). This approach ensured that the corpus represented observable rhetorical behaviour in naturalistic public discourse, without the collection or processing of private personal data, interactions, or psychological profiling beyond what is already present in public records.

Speeches were excluded if they constituted social media posts, press statements, informal remarks, or non-speech political text genres, as these formats differ structurally from formal rhetoric in linguistic intent, cadence, audience design, and normative impact.

Additionally, emergency COVID-19 communications by Sturgeon were omitted, given their exceptional regulatory context, which temporarily expanded executive communication into crisis-governance territory not comparable to routine political speech. This exclusion preserved corpus comparability by preventing crisis-only rhetoric from distorting the maximum-indicator behavioural peaks used later in the similarity analysis. While this exclusion is methodologically necessary to maintain the comparative homogeneity of the corpus under routine political conditions, we recognise that emergency contexts provide a uniquely revealing setting to observe how restrictions on civil liberties and appeals to authority are articulated within democratic systems. As crisis governance temporarily stretches the boundaries of executive communication, future exploratory analyses should specifically incorporate emergency rhetoric to assess how democratic systems manage extreme stress without geometrically transitioning into the authoritarian pole.

The selection of these five specific leaders was not random but followed a theoretical sampling strategy designed to cover the full spectrum of democratic erosion and rhetorical behaviour. Hitler's speeches (pre-government and government stages) serve as the historical 'gold-standard' reference vector, marking the absolute maximum of the authoritarian indicators across democratic collapse thresholds. Trump and Orban represent contemporary, active authoritarian drift from within established systems, capturing both the full-spectrum populist approach (Trump) and the institutional-informational autocratisation within the European Union (Orban). Meloni was selected as a critical test case for the 'grey zone' (radical right populism), allowing us to observe whether the model could differentiate between latent ideological alignment and active authoritarian intensity. Finally, Sturgeon provides the necessary democratic baseline: a leader governing within stable democratic norms but facing intense, existential ideological contestation (Scottish independence). Her inclusion demonstrates that high political conflict and polarisation do not automatically equate to authoritarian magnitude, highlighting the utility of our computational toolkit in distinguishing robust democratic debate from actual democratic erosion.

All transcripts were sourced from public, stable archives used by scholars for verifiable political speech analysis. Sources included Hitler's transcripts from <https://www.nommeraadio.ee/> (accessed on 1 September 2025), Trump's addresses from <https://rollcall.com/> (accessed on 4 September 2025) and <https://www.c-span.org/> (accessed on 7 September 2025), Sturgeon's speeches from the <https://www.gov.scot/> (accessed on 12 September 2025), Meloni's public addresses from <https://www.governo.it/en/interventi-en> (accessed on 17 September 2025), and Orban's speeches from the Hungarian Prime Minister's official English-language repository <https://miniszterelnok.hu/en/speeches/> (accessed on 22 September 2025).

The decision to restrict analysis to formal public addresses aligns with the behavioural aim of the study. Public political rhetoric differs from other political text genres in three relevant ways: (i) it is strategically constructed to shift social cognition, (ii) it is delivered to mass audiences, making rhetorical peaks normatively consequential even when isolated, and (iii) it represents observable political behaviour, not inferred psychological states of private individuals.

Corpus size was not fixed by number of speeches but by public availability within comparable rhetorical genres across political stages, allowing later metrics (cosine, Euclidean, composite similarity) to assess structure and magnitude independently of sample size disparities. This design prevents false equivalence between leaders with long archived corpora (e.g., Hitler, Trump) and leaders with shorter available temporal windows (e.g., Meloni, Orban 2025 data), a common methodological limitation in political speech research. Furthermore, while recent computational studies on authoritarian discourse employ large-scale datasets, such as decades of UN General Assembly debates (Mochtak 2025) or thousands

of executive speeches (Maerz 2019), our methodological design intentionally utilises a purposive, deeply curated corpus. Large-N inductive approaches are highly effective for tracking macro-level semantic shifts and average tone over time. However, because our Authoritarian Reference Index (ARI) is a deductive, behavioural diagnostic tool focused on capturing extreme rhetorical maxima (the ‘outer behavioural limits’ of persuasion) rather than central tendencies, large-scale aggregation would mathematically dilute these critical, episodic norm-breaching peaks. Therefore, a targeted corpus size is not a preliminary limitation, but a methodological requirement for conducting high-resolution, indicator-by-indicator behavioural diagnostics across varying institutional contexts without collapsing indicator-critical extremes into central tendencies.

By applying explicit inclusion and exclusion criteria, this corpus enables cross-historical, cross-linguistic, and cross-provider computational comparison, supporting the central behavioural question: whether authoritarian persuasion exhibits a structurally stable and magnitude-relevant signature that can be reliably detected by multiple independent LLM systems when supplied with the same normative rubric.

2.2. Indicators and Model Evaluation

To transform normative theory into measurable rhetorical behaviour, we operationalised an 11-indicator taxonomy derived from the original four Levitsky–Ziblatt dimensions and their subsequent comparative expansions in democratic-erosion research. Each indicator was defined to capture observable rule-relevant rhetorical behaviour rather than thematic ideology, ensuring that scoring reflected *how* power is linguistically enacted, not *which* political topics are discussed.

The indicators map the behavioural space of democratic-norm erosion along distinct but interrelated domains structured in four diagnostic dimensions (*rejection of democratic rules, denial of opponents’ legitimacy, tolerance or promotion of violence and willingness to restrict civil liberties*) that were expanded into eleven operational indicators (*attacks on the established order or the electoral process, use of extra-legal solutions, opponent as corrupt/traitor, opponent as existential threat, ‘enemies of the people’ language, organisational links with violence, support/justification of present violence, glorification/normalisation of historical or foreign violence, restriction of civil liberties, censorship proposals and repressive measures*) (see Table S1). Each indicator was anchored to a 0–10 ordinal severity scale, where 0 represents fully procedural democratic rhetoric, 5 represents high-conflict but norm-preserving speech, and ≥ 6 represents norm-erosive rhetorical peaks consistent with behaviour observed in democratic decline (Valentim et al. 2025). To illustrate how maxima-based indicators preserve field separation despite minor inter-model discrepancies, we include representative examples of scored sentences drawn from different dimensions of the taxonomy. In the *Denial of opponents’ legitimacy* dimension, Trump rallies frequently produced maxima-scored phrases such as “Our biggest threat are high level politicians that work in the United States government like Mitch McConnell, Nancy Pelosi, Schumer, Biden, Justice Department, because that’s poisoning our country” (25 March 2023) or “... The weak pathetic governor of Colorado. He’s a radical leftist, and he’s afraid of the people, the gangs from Venezuela. He’s afraid of them. He’s afraid to do anything. He’s lost control of the state” (4 November 2024), both located near the upper range of illiberal threat framing. In contrast, Sturgeon’s highest-scored conflict sentences—while vigorous in tone—remained norm-bounded, exemplified by formulations prioritising institutional accountability such as “And now, the terms of that Brexit are being negotiated by a UK Government with no clear mandate, precious little authority and no real idea, even within its own ranks, of what it is seeking to achieve” (27 June 2017), or “Otherwise (...) opposition parties will just keep casting doubt on the legitimacy of the process, so they can avoid the substantive debate on independence which Scotland deserves, but they so clearly fear” (28 June 2022), which did

not approach authoritarian maxima magnitudes. These examples demonstrate that the behavioural signal captured by indicator maxima is expressed through the configuration and co-activation of rhetorical peaks, not through mean tone, and that field separation is preserved when inspecting rhetoric through orientation and magnitude complementarily.

All speeches were scored independently by three large language model (LLM) systems—GPT-4o (OpenAI), Gemini 2.5-Pro (Google DeepMind), and Grok-4-Fast (xAI)—each queried using the same prompt, rubric, and indicator definitions, preventing provider-specific prompt drift from shaping the scoring geometry. The prompt enforced: (i) indicator-by-indicator behavioural scoring, (ii) no diagnostic labels or political judgement outside the rubric, (iii) peak-oriented norm sensitivity, and (iv) cross-linguistic neutrality, given that models have shown high reliability in multilingual psychological and political text-analysis tasks when guided by explicit taxonomies.

Inter-model agreement was verified by pairwise comparison of the maximum indicator scores across the full speech corpus (Table S2). Zero-point differences were considered as exact agreement, 1–2 point gaps light discrepancies, 3–4 points moderate, and ≥ 5 points high divergences. Although minor numerical variation was present, mean divergence remained consistently low across all model pairs (0.91–1.50), and median differences never altered indicator polarity, theoretical interpretation, or clustering structure, indicating a stable and coherent computational scoring space not driven by model-specific distortion. Given model-robust cluster stability, ChatGPT outputs were used as the working reference for the illustrative behavioural interpretation shown here.

Model versions, prompt templates, scoring procedures, and indicator–rubric logic were fully documented and preserved in version-controlled audit logs, following established best-practice guidelines for the use of LLMs in social-science discourse analysis (Abdurahman et al. 2025). The audit trail included (i) prompt snapshots under semantic versioning, (ii) model inference parameters ensuring deterministic scoring conditions (temperature, seed control, and response-stability flags), (iii) explicit variable and indicator definitions, coding rules, and transformation logic, and (iv) maxima-retention and score-selection policies. This guarantees that the computational scoring protocol, rhetorical vector construction, and clustering analyses can be reproduced transparently and independently, without relying on undocumented model interpretation or hidden data inputs. This documentation layer was implemented as a methodological safeguard against model-provider artefacts, ensuring that rhetorical behaviour is measured as an observable, norm-relevant computational footprint grounded in public corpora and explicit theoretical rubrics, rather than an implicit or sentiment-only classification.

2.3. Scoring Decision

For each leader and each rhetorical indicator, we retained the maximum observed score across the analysed speech corpus. This peak-oriented retention rule was selected because norm erosion is not driven by average discourse, but by extreme rhetorical events that shift the perceived boundaries of acceptable political behaviour. Measures of central tendency—such as means or medians—smooth critical deviations and risk masking the behavioural impact of episodic but normatively powerful rhetoric.

Recent findings in democratic-norm research (Abdurahman et al. 2025) show that exclusionary or norm-breaching rhetoric exerts its strongest societal effect when it appears at the extremes rather than at the centre of the distribution, and that the normative influence of a speech signal depends on rhetorical intensity more than on its frequency or averaged tone. These studies also demonstrate that high-status mainstream political actors shape perceived acceptability asymmetrically, highlighting that the most behaviourally informative moments in discourse are its peaks, not its baseline averages.

From a computational-behavioural perspective, maxima operate as normative stress-tests of rhetoric, capturing the upper limit of rule-relevant persuasion strategies employed by each political actor, independently of ideology, topic selection, or media format. Retaining maxima ensures that the analysis preserves the full expressive range of norm-relevant rhetorical behaviour, enabling later similarity and clustering analyses to remain sensitive to intensity capacity without diluting norm-critical extremes into typical discourse summaries.

This decision also improves cluster separability and detection sensitivity. In the context of democratic monitoring, authoritarian persuasion strategies typically reveal themselves through non-linear amplification on specific indicators—including delegitimization of political opponents, moralised threat framing, media hostility, and leader-personalisation rhetoric. These amplified peaks, even when episodic, carry disproportionate behavioural weight because they expand what audiences perceive as socially acceptable political conduct. Retaining maxima therefore ensures that the model measures the strongest behavioural signal each leader’s rhetoric has reached, rather than a smoothed central tendency that cannot reveal when norm-relevant thresholds are approached or crossed.

Importantly, maxima retention is not interpreted as evidence of persistent authoritarian intent, but as a conservative methodological strategy to detect the highest norm-relevant rhetorical force expressed publicly. This approach avoids false equivalence between high-conflict but norm-preserving democratic disagreement and behaviourally consequential authoritarian rhetorical peaks, while preserving the ability of subsequent similarity analyses (cosine structure, Euclidean magnitude, composite indices) to test for shared rhetorical architecture and intensity potential across political actors.

2.4. Authoritarian Reference Index: Measuring Shape and Magnitude

To provide a robust quantitative assessment of rhetorical proximity, we decompose authoritarian similarity into two mathematically distinct dimensions: directional alignment and intensity magnitude. We treat each leader’s rhetorical profile as a vector in an 11-dimensional space, where each dimension represents one of the indicators in our taxonomy. This vector-space approach is grounded in the necessity of distinguishing between rhetorical style and rhetorical severity.

The choice of these two specific metrics derives from their distinct geometric properties. Traditional similarity measures, such as simple correlation or lexical overlap, often fail to account for the interplay between thematic focus and absolute intensity. By utilising cosine similarity (alignment), we isolate the relative proportions of the indicators, providing a measure of the leader’s ideological “blueprint” that is invariant to the overall volume or “loudness” of the rhetoric.

On the other hand, we employ a normalised Euclidean distance (intensity) to capture the “behavioural heat” of the message. This metric is essential because directional alignment alone is insufficient to diagnose democratic drift; a leader may point in an authoritarian direction without possessing the energy required to breach institutional guardrails. By mapping leaders onto this two-dimensional plane, our index remains sensitive to the nuances of contemporary “hybrid” or “institutionalised” authoritarianism, where a leader might adopt the structural logic of the reference model while strategically modulating its intensity to maintain institutional legibility.

2.5. The Cosine Authoritarian Index (CAI): Structural Alignment

The CAI measures the directional alignment between the Hitler gold-standard vector (V_H) and the leader's vector (V_L). Conceptually, this index captures the “rhetorical blueprint” or the thematic prioritisation of the speech.

$$CAI (\%) = \left(\frac{V_H \cdot V_L}{\|V_H - V_L\|} \right) \cdot 100$$

By calculating the cosine of the angle between the two vectors, we isolate the relative emphasis placed on each of the 11 indicators, regardless of their absolute scores. A high CAI indicates that the leader is using the same underlying logic as the reference model, for instance, consistently prioritising the delegitimization of opponents and anti-pluralism over other indicators. This metric is what allowed us to identify the “latent” authoritarianism in cases like Meloni, where the structural alignment is high despite moderate overall intensity.

2.6. The Euclidean Authoritarian Index (EAI): Intensity Magnitude

While the CAI measures the shape, the Euclidean Authoritarian Index (EAI) measures the absolute distance in the multi-dimensional space. It quantifies the “behavioural heat” or severity of the rhetoric.

$$EAI (\%) = \left(1 - \frac{\|V_H - V_L\|}{\|V_{max}\|} \right) \cdot 100$$

To ensure comparability, we normalise the Euclidean distance using the norm of the maximum theoretical vector, $V_{max} = (10, 10, \dots, 10)$, representing the absolute ceiling of the taxonomy. The EAI is sensitive to the magnitude of the scores, and it differentiates a leader who whispers authoritarian themes from one who shouts them. In our framework, the EAI functions as the primary discriminator for active democratic erosion, as it captures the extreme rhetorical peaks that carry the most diagnostic weight.

The synergy between these two metrics constitutes the Authoritarian Reference Index (ARI). This integrated approach solves a common problem in political science: the false equivalence between radical populism and active authoritarianism. By mapping leaders onto a two-dimensional plane (alignment vs. intensity), we can distinguish between those who share an ideological “grammar” (high CAI) and those who have crossed the kinetic threshold of “active drift” (high EAI).

3. Results

3.1. Rhetorical Clustering

Rhetorical behaviour collapsed into two coherent poles when maximum indicator vectors were reduced to a two-component space using PCA and clustered using K-Means (Figure 1). The topology of the projected space shows a clear separation between an authoritarian-populist field containing Hitler, Trump, and Orban, and a democratic-institutional rhetorical field containing Sturgeon and Meloni. The first two principal components captured 93.2% of cumulative variance, confirming that the indicator-maxima projection retained nearly all behavioural signal relevant for clustering. K-Means selected $k = 2$ as the optimal solution (silhouette ≈ 0.687), and cluster membership was stable regardless of the minor score discrepancies observed across model outputs.

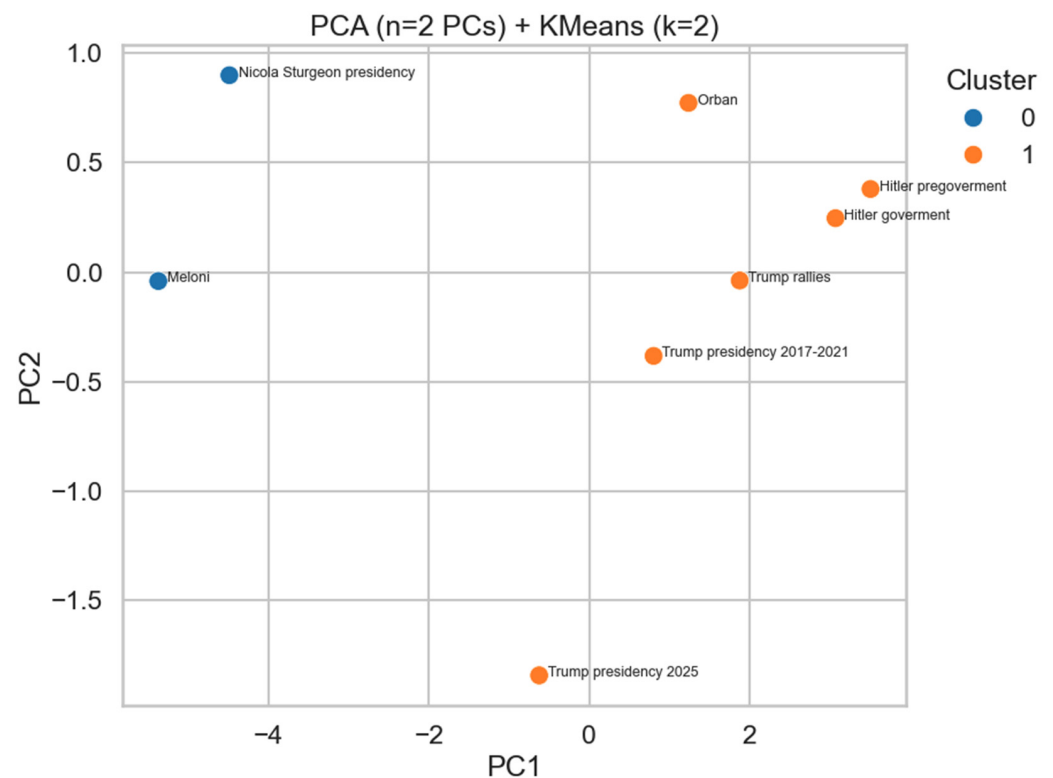


Figure 1. Political speech clustering from PCA of standardised maximum indicator scores. Principal Component Analysis (PCA) of the z-standardised maximum indicator scores (median-imputed) showed that the first two components captured 93.2% of cumulative variance. K-Means applied directly on the PC representation identified two clusters (best $k = 2$, silhouette = 0.687). Cluster-0 contained Nicola Sturgeon and Giorgia Meloni, whereas Cluster-1 grouped Hitler (pre-government and government), Orban, Trump 2017–2021, Trump rallies, and Trump 2025. Axes represent PC1 and PC2. Importantly, although Trump 2025 is assigned to the authoritarian cluster, this should be interpreted as early directional evidence, since the number of available Trump 2025 speeches is still limited (~1 year in term).

The pairwise similarity structure in Tables 1 and S4 reinforces this separation when inspected through both vector orientation and magnitude divergence. We included cosine similarity, Euclidean distance and composite similarity index, among others. Cosine similarity quantifies the directional alignment between two rhetorical vectors. It measures the angle between them, independently of their absolute magnitude. Values close to 1 indicate strong structural alignment, values near 0 indicate no directional relationship, and negative values reflect opposite orientations in the configuration of maximum scores. It is therefore an index of shared structural patterning across indicators. Euclid_z refers to the z-standardised Euclidean distance between two maximum-indicator vectors. It captures the absolute magnitude difference between speeches: how far apart two leaders are in terms of the highest intensity reached on each indicator. Low Euclid_z values indicate quantitative proximity; high values reflect substantial separation in rhetorical peak intensities. Finally, the composite similarity index integrates directional similarity (cosine), linear association (Pearson's r), and absolute separation (Euclid_z) into a single standardised measure ranging approximately from -1 to $+1$. Positive CSI values denote overall proximity between two rhetorical profiles, while negative values indicate global divergence across structure, magnitude, and linear co-variation. The highest internal alignment in the dataset corresponds to Hitler's speeches across government stages ($\cos = 0.95$; $\text{CSI} = 0.65$; $\text{Euclid}_z = 1.12$), indicating that his rhetorical maxima geometry remains highly self-consistent even across systemic position shifts. Trump's campaign rallies fall in close

structural proximity to early authoritarian maxima fields, i.e., high similarity to Hitler's pre-1933 period ($\cos = 0.87$; $\text{CSI} = 0.76$), showing that peak-level rhetorical configuration, rather than average tone, aligns most strongly during high-persuasion political moments designed for mass audiences.

Table 1. Dimensional comparison of rhetorical geometry across speech corpora using complementary metrics: Directional structural alignment (cosine similarity), linear co-variation of maximised indicators (Pearson's r), absolute and z-normalised configuration distance (Euclidean_raw, Euclidean_z), and the composite similarity index (CSI) capturing scaled structural proximity. Partial results show that high-persuasion political peaks, such as Trump's campaign rallies, exhibit strong structural alignment with early authoritarian maxima fields, even when average tone diverges. Government-stage speeches by Hitler and Trump confirm shared configuration of maximised rhetorical indicators, yet this alignment attenuates as contextual pressure shifts or the corpus window narrows. Orban occupies an intermediate bridging position, reflecting partial structural overlap without equivalent maxima magnitude capacity. Cross-metric asymmetries highlight that norm-relevant rhetorical behaviour cannot be reduced to sentiment averages or discrete ideological labels but must be inspected through orientation and maxima intensity as separate computational dimensions.

Speech	Speech	Cosine_Similarity	Pearson_Corr	Euclidean_Raw	Euclidean_z	CSI
Hitler pregov.	Hitler gov.	0.954	−0.291	2.236	1.125	0.654
Hitler pregov.	Trump rallies	0.871	0.311	6.000	2.005	0.763
Hitler pregov.	Meloni	−0.937	0.746	27.404	8.917	−0.032
Hitler gov.	Trump pres. 2017–2021	0.713	0.359	7.071	2.433	0.696
Hitler gov.	Orban	0.654	0.118	7.874	2.370	0.607
Trump rallies	Trump pres. 2017–2021	0.392	−0.493	5.745	2.042	0.323
Trump rallies	Trump pres. 2025	−0.286	0.006	9.327	3.365	0.109
Trump rallies	Orban	0.343	−0.539	7.681	2.355	0.287
Trump pres. 2017–2021	Trump pres. 2025	0.040	0.387	7.211	2.282	0.367
Sturgeon	Meloni	0.920	−0.449	5.831	2.166	0.598
Sturgeon	Orban	−0.589	0.147	19.026	5.977	−0.008

When comparing institutional rhetoric, the table reveals that structural alignment (cosine) and linear co-variation (Pearson's r) do not necessarily track intensity magnitude (Euclid_z), allowing deeper interpretation of rhetorical behaviour without resorting to categorical simplification. For example, government-stage Trump speeches show positive cosine similarity to Hitler's government corpus ($\cos = 0.71$; $\text{CSI} = 0.69$; $\text{Euclid}_z = 2.43$), confirming shared structural configuration of maximised indicators, but this similarity attenuates when contextual pressure shifts or when the corpus window narrows. His 2025 speeches, covering approximately one year of publicly available addresses, remain assigned to the same rhetorical cluster as his earlier 2017–2021 presidential period ($\text{CSI} = 0.36$; $\text{Euclid}_z = 2.28$), but exhibit lower directional alignment to Hitler's maxima configuration in the current window ($\cos = 0.040$), suggesting reduced internal coherence without inversion toward the democratic field.

Orban's position in Tables 1 and S4 is consistent with his intermediate location in the PCA space: his rhetorical profile bridges structural elements shared with the authoritarian field ($\cos = 0.65$ to Hitler gov; $\text{CSI} = 0.60$; $\text{Euclid}_z = 2.37$), while also being the closest structural neighbour to Trump rallies among non-Hitler government speakers ($\cos = 0.34$; $\text{CSI} = 0.28$; $\text{Euclid}_z = 2.35$). This reflects a pattern of partial structural overlap in norm-critical rhetorical configuration, but with moderated maxima magnitudes relative to Hitler's and Trump's peaks, explaining his placement within the same rhetorical field without reaching equivalent intensity capacity.

The democratic-institutional field emerges through inverse vector orientation and magnitude separation relative to the authoritarian maxima space. Sturgeon and Meloni show high mutual alignment ($\cos = 0.92$; $\text{CSI} = 0.59$; $\text{Euclid}_z = 2.16$), confirming a co-

herent internal rhetorical geometry shaped by institutional accountability, inclusion, and procedural norm signalling. When compared to Orban or Hitler government maxima, the divergence is driven predominantly by magnitude separation (Euclid_z $\approx$ 5.9–8.9), even when Pearson's r can appear neutral or positive in some pairings (e.g., $\cos = -0.58$ for Sturgeon <-> Orban but $r = 0.14$; $\cos = -0.93$ for Hitler pre-gov <-> Meloni but $r = 0.74$). These cross-metric asymmetries demonstrate that norm-relevant rhetorical behaviour is not reducible to sentiment averages or to discrete ideological categories, but must be inspected through configuration orientation and maxima magnitude capacity as separate computational dimensions.

Overall, the data indicate that the authoritarian-populist rhetorical field is structurally cohesive at its maxima, most clearly expressed during high-audience persuasion moments (Hitler corpus; Trump rallies; Orban speeches), while the democratic-institutional field remains structurally aligned internally, yet is magnitude-distant from authoritarian intensity peaks. The 2025 Trump subset remains mostly computationally proximal to Trump's earlier institutional maxima space rather than shifting toward democratic vector geometry, and Orban occupies a measurable intermediate bridge in structure, not in magnitude, preserving cluster boundaries without collapsing indicator-critical extremes into central tendencies.

These results were convergent across model systems, robust to indicator-level variation, and mathematically stable under maxima projection and clustering—demonstrating that authoritarian rhetorical behaviour can be measured reproducibly in public political speech as a structural configuration and intensity capacity signal, rather than as an averaged tone summary or an implicit intent attribution.

3.2. Authoritarian Similarity

Extreme values on single indicators were behaviourally diagnostic. Maximum indicator scores above 6 repeatedly appeared in both Hitler, Trump and Orban across multiple dimensions, whereas Sturgeon and Meloni remained consistently below that threshold (Figures 2 and 3). To quantify proximity to the authoritarian reference, we defined the Authoritarian Reference Index (ARI) decomposing similarity into two distinct dimensions: directional alignment (Cosine Index) and intensity magnitude (Euclidean Index). The results for each analysed profile are summarised in Table 2.

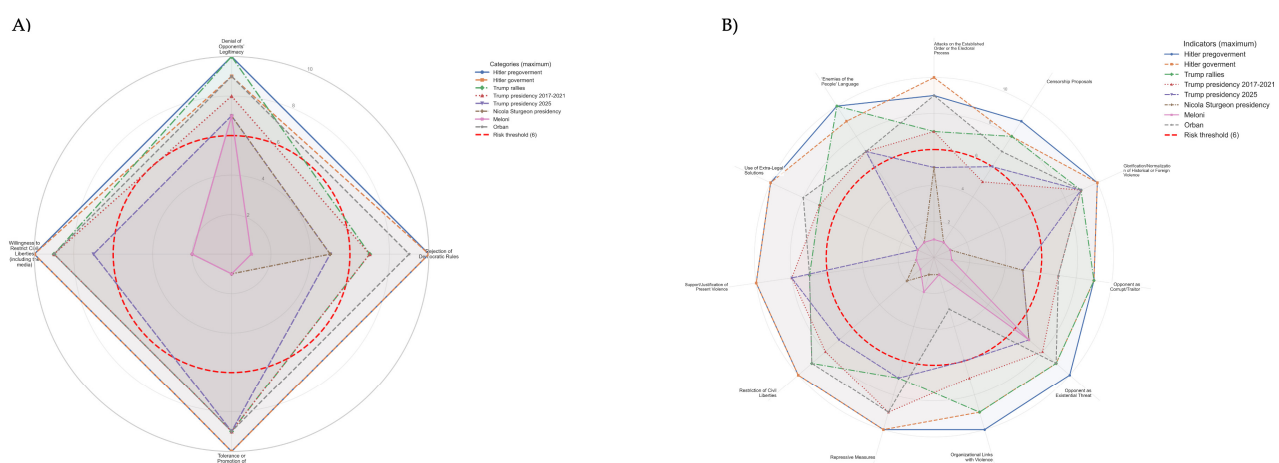


Figure 2. Maximum scores across categories and indicators. (A) Dot plot displaying, for each leader, the highest category-level score reached; horizontal jitter is included only to avoid point overlap. The red dashed horizontal line marks the risk threshold = 6. (B) Radar chart showing the maximum score attained in each individual indicator; the red dashed circumference denotes the same threshold to enable direct visual comparison. All values were computed using GPT-4o; negligible differences were observed relative to Gemini and Grok for this representation.

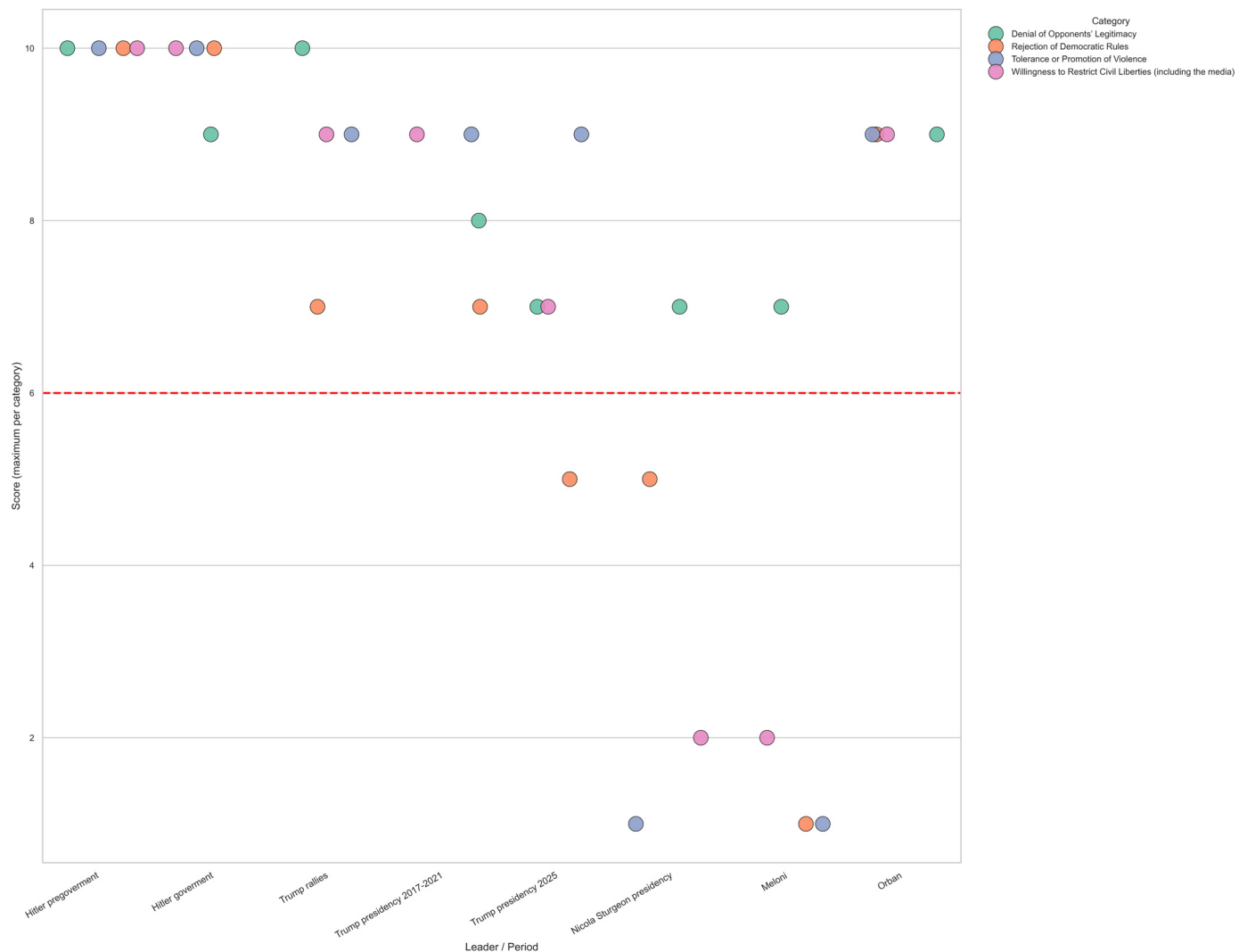


Figure 3. Category-level maximum scores by leader. Dot plot showing, for each leader/period, the maximum score attained in each category—*Denial of Opponent's Legitimacy*, *Rejection of Democratic Rules*, *Tolerance/Promotion of Violence*, and *Willingness to Restrict Civil Liberties (including the media)* (colour-coded). A small horizontal jitter is applied only to avoid overlap. The red dashed line marks the risk threshold (6). Scores are on a 0–10 scale and were computed with GPT-4o; differences for Gemini versus Grok were negligible for this representation.

Trump exhibited the highest convergence with the Hitler gold-standard vector, scoring 99.1% in directional alignment and 80.7% in intensity. This indicates that his rhetorical profile not only prioritises the exact same authoritarian dimensions (e.g., anti-pluralism, delegitimization of opponents) but does so with a severity magnitude approaching the reference extreme. Viktor Orban followed a similar trajectory, mirroring the structural alignment (97.6%) almost perfectly, although with a moderated intensity (72.4%). This moderated intensity could serve a dual purpose: it preserves the institutional legibility of the leader, allowing the discourse to remain largely compatible with international democratic forums, while simultaneously eroding democratic norms from within. By operating at a lower intensity threshold than the reference model, these leaders can deploy the same authoritarian logic (delegitimization of opponents, erosion of truth-seeking institutions) without triggering the immediate systemic alarms that a high-intensity, “pure” authoritarian discourse would provoke. Consequently, the ARI alignment score reveals the underlying blueprint of the speech, while the intensity score reflects the strategic adaptation to contemporary institutional constraints.

Table 2. Authoritarian Reference Index (ARI) and its constitutive metrics across analysed political profiles. The table summarises the decomposition of authoritarian similarity into two distinct dimensions: directional alignment (Cosine Index) and intensity magnitude (Euclidean Index). All values are expressed as percentages (0–100%) relative to the Hitler gold-standard vector. High directional alignment signifies structural thematic mimicry, while intensity magnitude measures the rhetorical severity or “heat” of the discourse.

Leader	ARI	
	Cosine Index (%) (Alignment)	Euclidean Index (%) (Intensity)
Giorgia Meloni	69.4%	16.4%
Nicola Sturgeon	74.3%	22.3%
Viktor Orban	97.6%	72.4%
Donal Trump presidential (2017–2021)	99.3%	74.8%
Donal Trump presidential (2025–)	95.3%	58.7%
Donald Trump rallies	99.1%	80.7%

Crucially, this dual-metric result directly addresses the empirical challenge of measuring ‘clever’ autocrats who are careful not to express directly anti-democratic opinions in core institutional areas. If our model measured the central tendency or average tone of Orban’s speeches, his autocratic signals would be diluted by his strategic use of democratic camouflage or simulated pluralism (Maerz 2019). By focusing on maxima peaks, the ARI bypasses this simulation, mathematically validating that a leader can perfectly mirror the structural blueprint of authoritarianism (97.6% directional alignment) while carefully moderating their behavioural ‘heat’ (72.4% intensity) to avoid triggering immediate systemic alarms. This illustrates the diagnostic usefulness of the approach: it provides a robust toolkit to detect the authoritarian spectre even when it is strategically sanitised.

While the aggregate ARI provides a high-level diagnostic of authoritarian proximity, a granular breakdown of the 11 taxonomic indicators reveals distinct rhetorical signatures (Figure 4). This analysis shows that proximity to the authoritarian reference is not a uniform distribution of scores, but a strategic prioritisation of specific illiberal pillars.

Our analysis identifies three distinct rhetorical fingerprints:

1. The “Full-Spectrum” Profile (Trump): Trump’s profile exhibits a near-perfect overlap with the Hitler gold-standard in the most hostile rhetorical dimensions. Importantly, his discourse reaches the maximum range (8/10, 9/10 or 10/10) in 9 of 11 authoritarian speech indicators defined by Levitsky & Ziblatt, signalling a full-spectrum saturation rather than isolated spikes. He matches the reference extreme in “glorification/normalization of historical of foreign violence”, “opponent as corrupt/traitor”, “opponent as existential threat”, “organizational links with violence”, “repressive measures”, “restriction of civil liberties” and explicit “‘enemies of the people’ language”. His rhetorical signature is characterised by a synergistic coupling of opponent-delegitimization and violence normalisation, producing a distinctly “kinetic authoritarian profile”—one that does not merely attack political credibility, but rhetorically endangers the perceived physical and moral safety of the opposition itself, framing adversaries as threats to be confronted, contained, or neutralised. This positions Trump’s discourse at the outer boundary of democratic speech norms, consistent with patterns observed in historical authoritarian rhetoric.
2. The “Institutional-Informational” Profile (Orban): Orban shares the maximum intensity with the authoritarian pole in “attacks on the established order or the electoral process”, “glorification/normalization of historical of foreign violence”, “opponent as

existential threat”, “repressive measures” and “restriction of civil liberties”. However, his profile contains a strategic outlier: “organisational links with violence” (3/10). Although no longer a minimal value, this score remains comparatively low and non-structuring relative to the full indicator set, suggesting that violent mobilisation is not an organising pillar of Orban’s authoritarian grammar, but rather a peripheral and instrumental resource—tolerated or rhetorically signalled at times, yet not central to political mobilisation. This pattern points to a distinctly modern and “sanitised” authoritarianism, one that relies on legal-administrative repression, informational leverage, and narrative dominance, rather than on the paramilitary networks or direct violent mobilisation that typified 20th-century authoritarian movements, where such links acted as structuring and organisational foundations, not auxiliary signals.

3. The “Civil Liberties” Paradox (Meloni and Sturgeon): notably, all analysed leaders—including those on the democratic-institutional pole—show high scores (7/10) on the indicator “opponent as existential threat.” Authoritarian leaders also exhibit this score, often exceeding it and reaching 9/10, yet this alone does not define authoritarian affordances. The results indicate that, although 7/10 reflects high rhetorical intensity, in democratic leaders this threat-framing remains norm-bounded, operating within constitutional constraints and not converting into authoritarian affordances. In contrast, in authoritarian leaders the same indicator co-occurs with additional markers that do enable authoritarian action—such as advocacy of repressive measures or civil-liberties restriction. Therefore, existential risk functions in democracies as democratic alarmism paired with procedural safeguards—not democratic breakdown—whereas in authoritarian rhetoric it acts as an accelerant only when embedded in a wider ecosystem of delegitimization and norm erosion.

The case of Meloni illustrates why the two dimensions of the ARI are critical, and how the index’s dual metric clarifies the boundary between the two identified clusters. While Meloni displayed a directional alignment with the authoritarian reference (69.4%), suggesting some thematic overlap in how political conflict is framed, her intensity score remains notably low at 16.4%. This decoupling of shape from magnitude suggests that while the structure of her rhetoric shares features with the authoritarian model, the intensity remains constrained within institutional bounds. This low intensity aligns her empirically with Sturgeon (intensity: 22.3%), explaining their grouping in the democratic-institutional pole.

This decoupling of structure and magnitude in Meloni’s profile raises a logical question: what is the minimum intensity threshold at which rhetorical alignment becomes a dangerous or active authoritarian signal? Meloni’s position suggests the existence of a “rhetorical grey zone”, that is, her discourse possesses the underlying blueprint of the authoritarian model but lacks the evidentiary “heat” or energy required to breach institutional guardrails.

This distinction is critical for understanding the transition from radical populism to active democratic erosion. In this grey zone, authoritarian discourse may remain functionally dormant or institutionalised. However, the high alignment score remains a latent diagnostic signal; it suggests that the ideological infrastructure is already in place, awaiting only a shift in intensity, perhaps triggered by a domestic crisis or a fabricated threat, to cross the threshold into active authoritarian behaviour. By this measure, the ARI reveals that while Meloni and Sturgeon share the same democratic-institutional pole, Meloni’s high structural alignment suggests latent overlap in conflict-framing configuration, while Sturgeon’s low alignment and bounded intensity indicate a rhetorical profile shaped by institutional pluralism rather than by illiberal persuasion geometry. The high structural alignment (Cosine Index) observed across ideologically aligned leaders like Trump, Orban, and Meloni reflects a shared underlying authoritarian grammar. Sociologically, this

shared rhetorical blueprint can be understood as a coordinated discursive response to the ‘status anxiety’ and cultural backlash occurring within Western democracies (Reiter 2025). While these leaders face different institutional constraints that modulate their rhetorical intensity (Euclidean distance), their high directional alignment confirms they are tapping into the same cross-national narrative: defending traditional social hierarchies and cultural hegemony against perceived demographic or liberal threats.

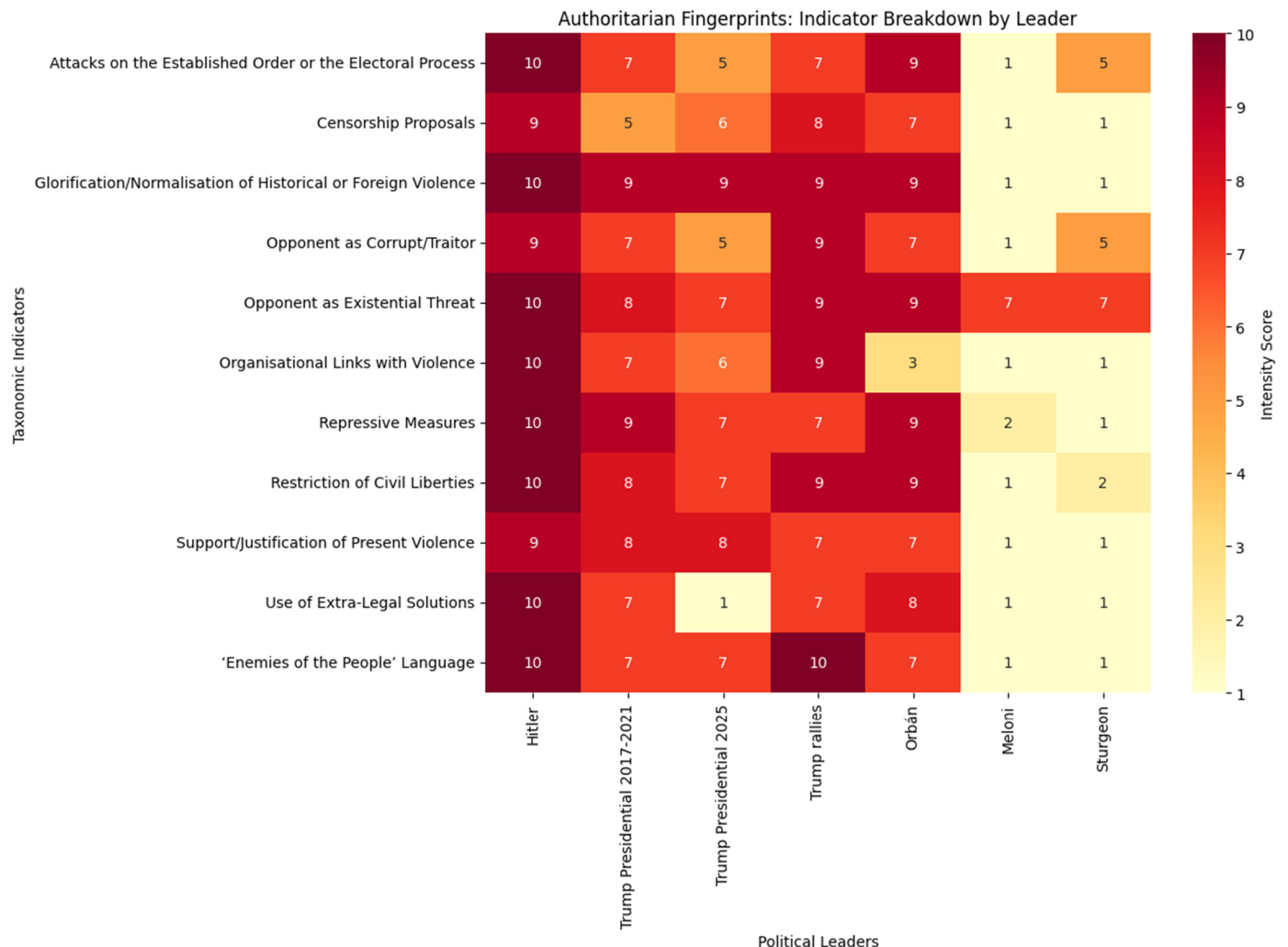


Figure 4. Authoritarian Fingerprints: Heatmap of Indicator Intensity across Political Leaders. The matrix displays the intensity scores (ranging from 1 to 10) for each of the 11 taxonomic indicators across the five analysed profiles. Darker red cells represent higher rhetorical intensity, indicating a closer proximity to the authoritarian gold-standard (Hitler).

3.3. Quantifying the Activation Threshold

To move from a conceptual to a quantitative definition of this threshold, we propose to define this activation threshold (AT) as the highest intensity score recorded within the democratic-institutional cluster (in this case, Sturgeon’s 22.3%). This threshold serves as an empirical red line: while high directional alignment indicates ideological proximity, it is only when intensity exceeds this 22.3% boundary that rhetoric begins to function as an active behavioural signature of authoritarian drift.

The 50.1 percentage point gap between the highest intensity of the democratic pole (Sturgeon, with 22.3%) and Orban (72.4%) represents a critical transition zone. Within this range, the rhetorical “heat” is sufficient to begin melting institutional constraints, shifting from the grey zone of latent alignment to the active zone of Trump and Orban.

Mathematically, this suggests that the ARI is most diagnostic when interpreted as a binary state: an alignment exceeding 85% creates the potential for drift, but an intensity surpassing the activation threshold effectively triggers the activation. By establishing the AT as a baseline, the ARI can be used not just for static comparison, but as a real-time monitoring tool to detect when a leader's latent authoritarian structure begins to gain the kinetic energy required for systemic erosion.

3.4. Rhetorical Anchoring Effect

The anchoring effect has been well-documented across the social sciences since the 1970s (Tversky and Kahneman 1974). In this section, we examine its specific role within authoritarian discourse. We propose that leaders who consistently operate above the activation threshold (AT) generate a Rhetorical Anchoring Effect, effectively recalibrating the public's cognitive baseline. By maintaining a high anchoring proximity to the authoritarian gold-standard, as observed in Trump's 80.7% intensity, these leaders do not merely violate democratic norms; they redefine what the audience perceives as normal or acceptable political discourse.

In this context, the ARI intensity score serves as a measure of desensitisation potential. When a leader's intensity remains high over multiple periods (e.g., Trump's 2017–2025 trajectory), the repeated exposure to extreme rhetorical peaks serves to anchor the public's expectations. This cognitive shift makes future authoritarian escalations appear predictable, or even banal, rather than aberrant. Consequently, the danger of the active zone (intensity > AT) is twofold: it signals immediate institutional erosion and simultaneously performs a long-term psychological conditioning of the electorate.

Conversely, leaders in the rhetorical grey zone, like Meloni, maintain a structural alignment that prepares the ideological ground without yet anchoring the public in extreme intensity. This suggests that the most critical risk for democratic stability occurs when a latent alignment (high Cosine Index) meets a sustained anchoring intensity (high Euclidean Index), as this combination effectively moves the goalposts of the entire political system, making the return to a pluralistic baseline increasingly difficult.

3.5. The Two-Dimensional Nature of Rhetorical Drift

In summary, the presented ARI provides a granular framework to distinguish between ideological overlap and active systemic threat. Our analysis demonstrates that authoritarian drift is not a linear progression but a two-dimensional phenomenon where directional alignment provides the ideological blueprint, while intensity magnitude serves as the trigger.

The data reveals two distinct rhetorical postures:

- **Active Authoritarianism:** characterised by both high alignment (>85%) and high intensity (exceeding the 22.3% Activation Threshold). As seen in the Trump–Orban cluster, this combination produces the rhetorical heat necessary to melt institutional guardrails and anchor public expectations in an illiberal baseline.
- **Latent Authoritarianism (The Grey Zone):** exemplified by Meloni, where a moderate structural alignment (69.4%) suggests a shared authoritarian grammar, but a low intensity (16.4%) prevents the discourse from crossing the behavioural activation threshold.

Ultimately, the most significant finding is that intensity functions as the primary discriminator for democratic risk. While high alignment indicates a potential for drift, it is the sustained operation in the active zone that facilitates the Rhetorical Anchoring Effect, desensitising the electorate and recalibrating the cognitive baseline of the political system. By identifying the 22.3% AT, the ARI moves beyond static classification, offering a

diagnostic tool to monitor when latent ideological structures gain the intensity required to become active agents of democratic erosion.

4. Discussion

The present findings validate our core hypotheses, supporting a central conclusion: authoritarian persuasion is not simply a stylistic feature of political language, but a behavioural rhetorical strategy that can be detected when speeches are analysed dimensionally rather than through sentiment averages. The collapse of indicator maxima into stable clusters after dimensional reduction indicates that persuasive political language, particularly at moments of mass influence, expresses a structured behavioural signal rather than analytical noise. This reinforces a principle long discussed in political psychology, that is, rhetoric becomes socially consequential not by conveying average emotional tone, but by reshaping the cognitive expectations of listeners regarding what institutions should tolerate and what power is permitted to do. Authoritarian rhetoric, in this light, operates as an adaptive mechanism for altering social cognition, reducing predictive variance in legitimacy processing, and reframing institutional friction or opposition signalling as impurities within the system rather than components of democratic variance.

The additional maxima-based analysis provides empirical grounding for interpreting rhetorical intensity as a separable behavioural dimension. In the results, authoritarian leaders repeatedly expressed peaks exceeding a norm-critical threshold across multiple indicators, whereas leaders in the democratic-institutional pole maintained maxima profiles that, although sometimes aligned in direction, remained behaviourally bounded in magnitude. This supports the interpretation that vector orientation and intensity magnitude perform different cognitive functions: alignment reflects how conflict is framed, while magnitude reflects the activation potential of that framing in mass audiences. The separability of these dimensions explains why rhetorical norm drift cannot be inferred from averaged tone or Pearson-linear movement alone, but instead requires inspecting the behavioural outer envelope of persuasive peaks.

The ARI complements this interpretation by formalising proximity behaviourally rather than ideologically. Directional overlap indicates shared indicator geometry, but the results show that only when intensity surpasses the maximum persuasion envelope observed in democratic-institutional speech does rhetoric behave as an active authoritarian signal, rather than a latent configuration. This observation supports the long-theorised inference that authoritarian persuasion is adaptive rather than uniform: contemporary leaders can deploy the same illiberal conflict-framing logic while modulating intensity to preserve institutional legibility, delaying norm-critical detection if analysis relies on means or tone averages. The maxima-based representation captures this envelope, preserving diagnostic sensitivity to extreme persuasion landmarks that anchor public expectations. This adaptive capacity is perfectly encapsulated in what our indicator analysis defined as the ‘Institutional–Informational’ profile, exemplified by Viktor Orban. Unlike historical autocrats who relied heavily on the rhetoric of kinetic violence and paramilitary mobilisation, Orban’s discourse scores conspicuously low on ‘organisational links with violence’ while matching the authoritarian reference in domains of administrative repression, censorship, and opponent delegitimization. This rhetorical sanitization allows leaders to operate within highly regulated supranational frameworks, such as the European Union, leveraging informational dominance and democratic simulation (Maerz 2019) to systematically erode norms from the inside. By decoupling structural alignment from violent intensity, our model successfully flags this modern, non-kinetic iteration of the authoritarian spectre (Mochtak 2025) that traditional frequency-based analyses or expert heuristics might easily misclassify as mere conservatism.

This finding connects directly with earlier work on democratic erosion that emphasised how norm-relevant persuasion pressure is exerted asymmetrically by rhetorical extremes, not central tendencies. [Hinterleitner and Sager \(2023\)](#) argued that democratic challengers erode norms gradually by making rule-breaking thinkable before it becomes legally feasible. The present results complement this by showing that LLM systems can detect when leaders' speech maxima encourages audiences to cognitively model opponents or institutional friction as impurities, reducing variance in legitimacy expectations and making exceptional power cognitively predictable rather than aberrant. This supports the inference that norm erosion begins rhetorically when legitimacy expectations shift, not when sentiment means turn negative.

From a behavioural systems perspective, maxima peaks act as cognitive anchor points in persuasion space, recalibrating what publics cognitively register as institutionally tolerable. This aligns with the anchoring principle established by [Tversky and Kahneman \(1974\)](#), now interpreted behaviourally in political language space: repeated exposure to high-intensity persuasion peaks does not require every speech to be extreme, but even episodic maxima landmarks can recalibrate the predictive legitimacy envelope in mass cognition. This supports why maxima-based vector representation carries disproportionate diagnostic power: it captures the outer behavioural limit of persuasion geometry, not its average.

Why do audiences accept and normalise these high-intensity persuasion peaks? The effectiveness of this Rhetorical Anchoring Effect is inextricably linked to the electorate's pursuit of status preservation. As [Reiter \(2025\)](#) observes, populations experiencing perceived cultural displacement prioritise the restoration of pride and social honour over traditional rational-choice economic interests. Extreme rhetorical peaks function as powerful cognitive anchors precisely because they validate these profound status anxieties. By framing marginalised groups or democratic institutions as existential threats, authoritarian leaders offer a psychological restoration of dominance that makes subsequent democratic erosion acceptable, and even desirable, to their base.

Our findings provide a novel methodological solution to the semantic ambiguity identified in prior computational text analyses. While inductive machine-learning models treat authoritarian discourse merely as "the discourse championed by authoritarians" based on macro-level country indices ([Mochtak 2025](#)), such approaches struggle to mathematically decouple thematic similarities from actual behavioural severity. Our proposed Authoritarian Reference Index (ARI) resolves this by decomposing rhetorical similarity into two distinct dimensions: directional alignment (cosine) and intensity magnitude (Euclidean). This dual metric clarifies why leaders across different regime types might sound similar on the surface, a phenomenon heavily noted in the recent literature ([Maerz 2019](#); [Mochtak 2025](#)). For instance, our analysis shows that a leader can share a remarkably high structural alignment with authoritarian framing, yet remain firmly within democratic-institutional bounds due to low intensity scores. This demonstrates that structural alignment alone is insufficient for diagnosing authoritarian drift; absolute magnitude is the primary discriminator that converts latent framing into an active behavioural signature.

Furthermore, our methodological decision to isolate maximum indicator scores effectively neutralises the "democratic simulation" tactics utilised by autocratic regimes to mask their true nature ([Maerz 2019](#)). If a computational model evaluates the central tendency of a speech, an authoritarian leader's isolated attacks on political opponents could easily be diluted by their extensive use of procedural, democratic terminology. By focusing on extreme rhetorical peaks evaluated by state-of-the-art LLMs, our framework bypasses this linguistic camouflage. Thus, we build upon the predictive capabilities highlighted by [Mochtak \(2025\)](#) and the dictionary approaches of [Maerz \(2019\)](#) by demonstrating that

it is not the average tone, but the explicit rhetorical extremes functioning as cognitive anchor points that best signal democratic erosion. Ultimately, this theoretical and empirical distinction provides the robust conceptual basis required for reliable, real-time democratic monitoring systems, such as the proposed AI Democracy Observatory.

In this light, the ability of LLMs to detect rhetorical field membership complements earlier corpus-wide studies on political persuasion basins of attraction. Card et al. (2022) demonstrated that political language becomes increasingly polarised over long historical windows, particularly in immigration framing. Our maxima-based clustering complements this inference behaviourally, showing that authoritarian persuasion is detectable when leaders behaviourally amplify norm-critical indicator geometry under mass-audience persuasion pressure, rather than when sentiment averages shift.

To operationalize this vision, we propose establishing an AI Democracy Observatory (<https://democracyobservatory.org/>) that is not presented as a result or extension of the current dataset, but as a governance implication logically derived from the analytical principles validated by the study. A permanent, open, and auditable rhetorical monitoring system is methodologically justified not because rhetoric becomes authoritarian on average, but because the outer envelope of persuasion peaks anchors how listeners cognitively judge institutional tolerance, threat, and exceptional power. If AI systems are to support democratic accountability, their function must remain preventive rather than punitive, allowing corpora to be inspected dimensionally and comparatively without collapsing behavioural peaks into ideological averages or discrete categories. The observatory would therefore serve as a public instrument of vigilance, enabling societies to detect early when persuasion geometry begins to redraw what audiences perceive as institutionally tolerable, not merely stylistically uncivil.

Limitations and Future Directions

The limitations of this approach remain essential for future generalisation but do not weaken its internal validity. Representing rhetoric through maxima vectors necessarily prioritises extreme persuasion moments, reducing visibility of moderate registers without implying ideological inversion. Public transcripts introduce transcriptional and translation noise that cannot be fully decoupled from modelling, even when corpora are cross-validated. Model outputs remain prompt-sensitive despite dimensional validation, requiring cautious interpretation within corpus comparability boundaries, historical signal validity, and behavioural maxima capacity fields. These boundaries are not analytical weaknesses—they are empirical constraints that must inform future cross-cultural extension, particularly in non-Western rhetorical ecosystems where maxima capacity envelopes may shift without implying geometric inversion. Furthermore, the use of Adolf Hitler as the primary historical reference vector was intentionally designed to establish an absolute mathematical ceiling (the outer behavioural limit) for the 11-indicator taxonomy, rather than to represent all possible forms of authoritarianism. We acknowledge that authoritarian traditions articulated through different ideological registers (e.g., far-left authoritarianism or varying non-Western autocratic models) might exhibit distinct rhetorical geometries. Expanding the reference benchmarks to include diverse authoritarian prototypes is a necessary next step to test the conceptual robustness of the model across the full ideological spectrum. Future research must therefore test empirically whether persuasion scaffolding geometry remains invariant across languages and media, or whether regional idioms modulate maxima capacity while preserving pole orientation. This remains a critical question for computational rhetoric, but one that must be answered through data, not rhetorical framing. Most importantly, while this methodological toolkit successfully measures rhetorical capability, future research should systematically link these computational speech metrics with

databases of empirical executive actions (such as V-Dem or similar indices). Correlating the emergence of authoritarian rhetorical peaks with subsequent concrete policy actions will provide a more powerful statement on how discourse precedes institutional dismantling. Furthermore, applying this framework longitudinally to track the ‘evolution’ of authoritarian messages over time—such as comparing a leader’s rhetoric across consecutive terms or tracing shifts between pre-electoral and government stages—will provide critical insights into how authoritarian persuasion adapts to changing institutional pressures.

5. Conclusions

Taken together, the results demonstrate that authoritarian persuasion leaves a reproducible maxima-geometry footprint, detectable computationally when normative behaviour is operationalised dimensionally rather than categorically. Our Authoritarian Reference Index (ARI) successfully proves our central thesis: democratic erosion is signalled not merely by the presence of populist topics, but by extreme rhetorical peaks that cross specific intensity thresholds. By decoupling structural alignment (cosine) from behavioural intensity (Euclidean), this framework solves the ambiguity of the “grey zone” in political communication, distinguishing between robust democratic polarisation and active authoritarian drift.

This study contributes to the theoretical approach of political sciences by illustrating, through rhetorical design, the statistical recurrence of the authoritarian perspective in contemporary leadership. Ultimately, this provides a measurable behavioural domain for detecting early legitimacy envelope shifts, ensuring that computational tools can reveal democratic norm drift proactively, serving as an early-warning system to guard democratic institutions.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/socsci15060372/s1>, Code, Prompt structure used for AI, Speeches included in the analysis, Table S1: Authoritarian behavior indicators and scoring rubric derived from the Levitsky–Ziblatt framework; Table S2: Pairwise comparison of maximum indicator scores across all speeches, Table S3: Comparative structure of political speeches and Table S4: Rhetorical similarity across leaders’ speech corpora.

Author Contributions: All authors contributed equally to all aspects of this work, including conceptualisation, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualisation, supervision and project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical approval and informed consent were not required for this study, as all analysed material consists exclusively of publicly available political speech transcripts and no human participants were recruited, intervened upon, or assessed outside of public corpora. The methodological decision that individual consent was not applicable to this analysis was established at the outset of the research protocol in September 2025.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Abdurahman, Suhaib, Alireza Salkhordeh Ziabari, Alexander K. Moore, Daniel M. Bartels, and Morteza Dehghani. 2025. A Primer for Evaluating Large Language Models in Social-Science Research. *Advances in Methods and Practices in Psychological Science* 8: 25152459251325174. [\[CrossRef\]](#)
- Alhawamdeh, Salaam M., and Rajai R. Al-Khanji. 2025. Discursive Strategies Used in Trump's Speech: Rally in Perry, GA, United States on September 25, 2021. *Theory and Practice in Language Studies* 15: 1807–16. [\[CrossRef\]](#)
- Card, Dallas, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational Analysis of 140 Years of US Political Speeches Reveals More Positive but Increasingly Polarized Framing of Immigration. *Proceedings of the National Academy of Sciences* 119: e2120510119. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hinterleitner, Markus, and Fritz Sager. 2023. Political Challengers and Norm Erosion in Advanced Democracies. *European Journal of Political Research* 62: 1301–19. [\[CrossRef\]](#)
- Hobolt, Sara B., and Moritz Osnabrügge. 2025. Countering Authoritarian Behavior in Democracies. *Political Behavior* 47: 781–800. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kadim, Enas Naji. 2022. A Critical Discourse Analysis of Trump's Election Campaign Speeches. *Heliyon* 8: e09256. [\[CrossRef\]](#) [\[PubMed\]](#)
- Levitsky, Steven, and Daniel Ziblatt. 2018. *How Democracies Die*, 1st ed. New York: Crown Publishing Group.
- López, Alejandro Buitrago, Javier Pastor-Galindo, and José A. Ruipérez-Valiente. 2024. LLM-assisted topic modeling for hate speech characterization. *Authorea*, in revision. [\[CrossRef\]](#)
- Maerz, Seraphine F. 2019. Simulating Pluralism: The Language of Democracy in Hegemonic Authoritarianism. *Political Research Exchange* 1: 1–23. [\[CrossRef\]](#)
- Mochtak, Michal. 2025. Chasing the Authoritarian Spectre: Detecting Authoritarian Discourse with Large Language Models. *European Journal of Political Research* 64: 1304–25. [\[CrossRef\]](#)
- Othon, Adriano De Oliveira. 2021. Indicators of Authoritarian Behavior of Political Leaders. The Bolsonaro Case (2019–2020). *Anuario Latinoamericano—Ciencias Políticas y Relaciones Internacionales* 11: 83–98. [\[CrossRef\]](#)
- Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E. Robertson, and Jay J. Van Bavel. 2024. GPT Is an Effective Tool for Multilingual Psychological Text Analysis. *Proceedings of the National Academy of Sciences* 121: e2308950121. [\[CrossRef\]](#) [\[PubMed\]](#)
- Reiter, Bernd. 2025. Aggrieved White Men and the Danger They Pose to Democracy and Peace. *Social Sciences* 14: 93. [\[CrossRef\]](#)
- Rosenfeld, Gavriel D. 2019. An American Führer? Nazi Analogies and the Struggle to Explain Donald Trump. *Central European History* 52: 554–87. [\[CrossRef\]](#)
- Suter, Viktor, and Miriam Meckel. 2024. Using GPT-4 for Text Analysis: Insights from English and German Language News Classification Tasks. Paper presented at the 18th International AAAI Conference on Web and Social Media, Buffalo, NY, USA, June 3–6. [\[CrossRef\]](#)
- Tversky, Amos, and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185: 1124–31. [\[CrossRef\]](#)
- Valentim, Vicente, Elias Dinas, and Daniel Ziblatt. 2025. How Mainstream Politicians Erode Norms. *British Journal of Political Science* 55: e105. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.